

machine learning

semantiek

retrieval

kennisorganisatie

classificeren

PDF

Maak het vindbaar

op schijven, sites
en SharePoint

ontologie

hyperlinks

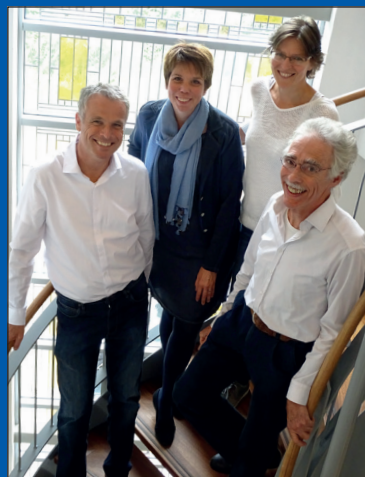
beeldanalyse

XML

thesaurus

taxonomie

Het beheer van grote hoeveelheden documenten vormt een steeds grotere uitdaging voor organisaties. Of het nu gaat om bestaande of nieuwe informatiesystemen, lokaal of in de cloud, informatiemanagers stellen zich de vraag welke mogelijkheden er zijn om de toegankelijkheid en vindbaarheid van de informatie zo effectief mogelijk te realiseren. Een goede vindbaarheid bespaart immers vele uren arbeidstijd en voorkomt incomplete dossiers. Traditionele instrumenten als taxonomieën, thesauri en autorisatielijsten bewijzen daarbij nog dagelijks hun waarde en de technische ontwikkelingen hebben de mogelijkheden uiteraard verruimd: automatische indexering en klassering, ontologieën en hyperlinking zijn waardevolle aanvullingen.



In dit boek behandelen we belangrijke methoden en technieken om informatie (documenten) van een organisatie vindbaar te maken. De theorie van de toegankelijkheidsleer wordt vanaf de basis behandeld en aan de hand van vele voorbeelden komen technieken en instrumenten als taxonomie, thesaurus, ontologie, zoekmachine en classificatie aan de orde, inclusief stappenplannen om hier zelf mee aan de slag te gaan.

Omdat SharePoint een veelgebruikt platform is voor het beheren en delen van documenten, besteden we een apart hoofdstuk aan de wijze waarop documenten binnen SharePoint zo goed mogelijk vindbaar kunnen worden gemaakt.

De hoofdstukken worden afgewisseld met kaderteksten waarin specifiek wordt ingegaan op gerelateerde onderwerpen als XML, machine learning en cardsorting.

Iedereen die in de praktijk betrokken is bij de implementatie van een informatiesysteem of in opleiding is tot informatieprofessional kan putten uit de uitgebreide beschrijvingen en handvatten die dit werk biedt. Omdat studenten een deel van de doelgroep vormen, is dit werk als open textbook onder een Creative Commons licentie gratis te downloaden.

UDOC.,





Marjolein van der Linden (1969) is vanaf 1999 docente aan de Hogeschool van Amsterdam, opleiding Media Informatie en Communicatie en sinds 2014 ook docente bij de opleiding Communicatie van de HvA. Zij studeerde informatiewetenschap aan de UvA en Psychosynthese aan het Instituut voor Psychosynthese in Hilversum. Daarnaast is Marjolein freelance docente bij GO Opleidingen.



Joyce van Aalten (1978) is sinds 2010 zelfstandig adviseur bij Invenier en docent bij GO Opleidingen. Ze adviseert organisaties in de toepassingsmogelijkheden en het gebruik van metadata, taxonomieën en search in het algemeen en binnen SharePoint in het bijzonder. Hiervoor werkte ze als adviseur en deskresearcher voor diverse organisaties, waaronder DNV-CIBIT en Berenschot.



Eric Sieverts (1947) is van huis uit natuurkundige. Van 1980 tot 2012 hield hij zich bij de Universiteitsbibliotheek Utrecht bezig met zoeksystemen en met Innovatie & Ontwikkeling. Van 1991 tot 2012 was hij docent bij de Hogeschool van Amsterdam, bij de Opleiding Media, Informatie en Communicatie. Op dit moment is Eric nog docent bij onder meer VOGIN en GO Opleidingen, redacteur van vakblad IP en organisator van de jaarlijkse VOGIN-IP-lezing.



Peter Becker (1967) studeerde Culturele Informatiewetenschappen en werkte 10 jaar als muziekbibliothecaris. Vervolgens was hij een aantal jaren als opleidingsmanager verbonden aan GO Opleidingen. Sinds 2007 is Peter docent informatiebeheer aan de Haagse Hogeschool en bij GO Opleidingen.

Maak het vindbaar

op schijven, sites en Sharepoint

Joyce van Aalten, Peter Becker
Marjolein van der Linden, Eric Sieverts

Open Textbook

Open textbook



Tenzij anders vermeld is alles in dit werk gelicenseerd onder een Creative Commons Naamsvermelding 4.0-licentie. Wanneer je gebruik wilt maken van dit werk, hanteer dan de volgende methode van naamsvermelding: Joyce van Aalten, Peter Becker, Marjolein van der Linden, Eric Sieverts: Maak het vindbaar op schijven, sites en SharePoint (2017), CC-BY-SA 4.0 gelicenseerd.

De volledige licentie-tekst is te lezen op: <https://creativecommons.org/licenses/by/4.0/>.

Tekstredactie: Christa van de Graaf

ISBN 9789492388001

NUR 123

D

U

O

I

N

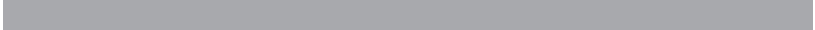
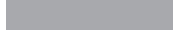
I

Inleiding	8
Hoofdstuk 1 Toegankelijk maken van informatie	10
1.1 Informatie en documenten	10
1.2 Opslaan om gevonden te worden	13
1.3 Toegankelijkheid	14
1.4 Metadata	16
1.5 Informatietalen	19
1.6 Bouwer, indexeerder en zoeker	21
1.7 Samengestelde onderwerpen	22
1.8 Samenvattend	24
Hoofdstuk 2 Informatie zoeken	26
2.1 Informatiecollecties en hun zoekhulpmiddelen	26
2.2 Aard van de zoekhulpmiddelen	27
2.3 Zoekvragen	29
2.4 Zoektechnieken	30
2.4.1 Zoeken met woorden	30
2.4.2 Booleaans zoeken	31
2.4.3 Best-match en relevantieordening	34
2.4.4 Nabijheidsoperatoren	36
2.4.5 Truncatie en woordstammen	37
2.5 Velden, filteren en facetzoeken	38
2.5.1 Zoeken in auteursveld	39
2.5.2 Zoeken op titelwoorden	39
2.5.3 Zoeken op trefwoordvelden	40
2.5.4 Andere velden om op te zoeken of filteren	41
2.6 Zoekresultaten	43
2.6.1 Kwaliteit van het resultaat	43
2.6.2 Opbrengst, vangst en precisie	44
2.7 Samenvattend	47
Kader SEO en SEA	50
Hoofdstuk 3 Ordening	52
3.1 Ordening en hiërarchie	52
3.2 Betekenis van het begrip taxonomie	54
3.3 Werkwijze en regels	58
3.3.1 Stap 1: Doel bepalen	59
3.3.2 Stap 2: Begrippen verzamelen: analyse van de content en de processen	60
3.3.3 Stap 3: Facetten en groepen	61
3.3.4 Stap 4: Collocatievolgorde	65
3.4 Samenvattend	66
Kader Cardsorting	68

Hoofdstuk 4 Van keuzelijst tot thesaurus	72
4.1 Woordenlijsten	72
4.2 Doel van een thesaurus	75
4.3 Wat is een thesaurus?	76
4.4 Concepten en termen	78
4.5 Proces thesaurusbouw	79
4.6 Samenvattend	93
Kader ISO 25964	95
Kader Checklist opstellen randvoorwaarden thesaurus	97
Hoofdstuk 5 Eigennamen in thesauri en autorisatielijsten	98
5.1 Toegankelijk maken van (eigen)namen	99
5.2 Verwijzingen	100
5.3 Algemene richtlijnen voor het opnemen van namen	100
5.4 Vaker voorkomende namen	102
5.5 Geografische namen	102
5.6 Namen van tijdvakken, tijdperken en gebeurtenissen	104
5.7 Persoonsnamen	105
Hoofdstuk 6 Digitale en fysieke ordening	108
6.1 Digitale ordening	109
6.1.1 Doel, begrippen, homogene groepen en collocatievolgorde	110
6.1.2 Stap 5: Combinatievolgorde	111
6.1.3 Stap 6: Bepaal de volgorde van de mappen (schemavolgorde)	114
6.2 Fysieke plaatsing van documenten	116
6.2.1 Stap 1 t/m 4: Doel bepalen, verzamelen van begrippen, homogene groepen en collocatievolgorde	117
6.2.2 Stap 5 en 6: combinatievolgorde en schemavolgorde	118
6.2.3 Stap 7 en 8: werk de classificatie uit in een schema en ken notatie toe	121
6.2.4 Schema bouwen en notaties	126
6.2.5 Stap 9 en 10: Register samenstellen en handleiding schrijven	126
6.3 Samenvattend	127
Hoofdstuk 7 Toegekend indexerend	129
7.1 Inhoudsanalyse	130
7.2 Van inhoudsanalyse naar informatietaal	133
7.2.1 Indexeerdiepte	134
7.2.2 Globaalindexering en diepte-indexering	135
7.3 Regels bij het toekennen van termen	137
7.4 Tagging en folksonomies	140
7.5 Samenvattend	145

Kader Beeldmateriaal	147
Kader Wordpress	153
Hoofdstuk 8 SharePoint	154
8.1 Site-architectuur	155
8.1.1 Site-architectuur: sites binnen een siteverzameling	155
8.1.2 Site-architectuur: site-onderdelen per site	156
8.2 Metadata-architectuur	157
8.2.1 Metadata: eigenschappen in kolommen	157
8.2.2 Beheerde metadata	158
8.2.3 Ondernemingstrefwoorden	160
8.2.4 Inhoudstypen	160
8.2.5 Toekennen van gecontroleerde termen	161
8.2.6 Zoeken met termensets	161
8.2.7 Mappenstructuur	163
8.3 Pagina-architectuur	164
8.4 Samenvattend	164
Hoofdstuk 9 Vormen van kennisrepresentatie	166
9.1 Semantische netwerken	167
9.2 Ontologieën	170
9.3 Vergelijking van kennisrepresentaties	175
9.4 RDF en tripels	176
9.5 Graphs	180
9.6 Toepassingen	182
9.6.1 SKOS	182
9.6.2 Linked data	185
9.6.3 Knowledge Cards, Knowledge Graph en Knowledge Vault bij Google	188
9.6.4 Schema.org	190
9.7 Samenvattend	192
Kader XML	197
Hoofdstuk 10 Automatisch metadateren en klasseren	201
10.1 Toepassingen: automatisch ontlenen en automatisch verrijken	201
10.1.1 Automatisch ontlenen	202
10.1.2 Automatisch verrijken	205
10.2 Methoden: rule based en getraind met voorbeelden	205
10.2.1 Methode 1: Rule based	206
10.2.2 Methode 2: getraind met voorbeelden	206
10.3 Technieken: statistisch en linguïstisch	209
10.4 Kwaliteit van het resultaat	210
10.5 Praktijkvoorbeelden van automatische verrijking	214
10.6 Samenvattend	218

Kader Automatische beeldherkenning	222
Kader Machine learning	230
 Hoofdstuk 11 Verbetering van zoeken met natuurlijke taal	 233
11.1 Problemen met eenvoudige freetext retrieval	234
11.2 Woordstammen, fuzzy zoeken, spellingscorrecties en splitsen van samengestelde woorden	235
11.3 Gebruik van semantische kennis	237
11.4 Technieken voor relevantie-ordering	240
11.5 Clusteren van zoekresultaten	241
11.6 Gegenereerde suggesties voor verfijnen van zoekresultaat	242
11.7 Terugkoppeling door gebruikers	245
11.8 Semantisch zoeken	246
11.9 Samenvattend	247
 Kader Vectormodel	 250
Kader TF*IDF	253
 Hoofdstuk 12 Hypertext, hyperlinks en relaties	 255
12.1 Aard van relaties	255
12.2 Hypertext	258
12.3 Structuur in hypertext	262
12.4 Gebruiksproblemen van hypertext	267
12.5 Hyperlinks maken toegankelijk	272
12.6 Hyperfictie	272
12.7 Samenvattend	273
 Hoofdstuk 13 Praktijkvoorbeelden	 275
13.1 AAT	275
13.2 MeSH thesaurus (Medical Subject Headings)	282
13.3 CAB-thesaurus	285
13.4 Thesaurus Zorg en Welzijn en het gebruik binnen kanker.nl	287
13.5 Thesaurus politiekunde	293
13.6 DDC, UDC en SISO	295
13.7 Iconclass	297
13.8 Automatisch klasseren bij het ministerie van Binnenlandse Zaken en Koninkrijksrelaties	302
13.9 Schema.org	305
 Definities	 313
Begrippenlijst	324



Inleiding

De dagelijkse groei van beschikbare informatie biedt niet alleen ongeken- de mogelijkheden, het stelt ons ook voor de uitdaging om het enorme aanbod beheersbaar en vindbaar te houden. Op het web bedienen we ons dagelijks van zoekmachines om snel bij de gewenste informatie te komen. We accepteren dat het zoekresultaat vaak niet compleet is omdat we in de meeste gevallen meer dan voldoende informatie vinden. Bij informatie die binnen de context van een organisatie is opgeslagen voldoet dit lang niet altijd, omdat vaak geen cruciale informatie gemist mag worden. Bovendien kunnen niet alle technieken van het web in een gesloten omgeving worden toegepast. Wie een informatiesysteem als een intranet of documentmanagementsysteem inricht, zal zich dan ook moeten verdiepen in de toepassing van andere methoden om de informatie toegankelijk te houden.

In dit boek behandelen we alle belangrijke methoden en technieken om de informatie, en met name documenten, in de context van een organisatie vindbaar te maken. We doen dit door de theorie van de toegankelijkheidsleer vanaf de basis te behandelen. Daarbij is gekozen voor een praktische insteek. In de eerste hoofdstukken komen aan de hand van verschillende voorbeelden de gebruikelijke instrumenten als taxonomie, thesaurus en classificatie aan de orde, inclusief stappenplannen om hiermee aan de slag te gaan. Omdat de gekozen oplossingen moeten aansluiten bij de wijze waarop gebruikers informatie zoeken, wordt hier in hoofdstuk 2 specifiek aandacht aan besteed. SharePoint is een veelgebruikt platform voor het beheren en delen van documenten. In hoofdstuk 8 gaan we daarom in op de wijze waarop de besproken instrumenten binnen SharePoint kunnen worden ingezet om documenten zo goed mogelijk vindbaar te maken. Nadat we in hoofdstuk 9 modernere vormen van kennisrepresentatie hebben behandeld, bespreken we vanaf hoofdstuk 10 de geavanceerde technische mogelijkheden als automatische classificatie, zoekmachines en de oplossingen van hyperlinking.

Naast de vele voorbeelden die tussen de theorie worden aangeboden, beschrijven we in hoofdstuk 13 concrete casussen waarbij alle besproken technieken in de praktijk worden toegepast. Dit boek kan worden beschouwd als de opvolger van het in 2010 verschenen werk “Organiseer je informatie”. Waar dat relevant is, zijn enkele passages uit dat werk overgenomen. Vergeleken met “Organiseer je informatie” zijn niet alleen veel onderwerpen geactualiseerd, uitgebreid en toegevoegd (zoals Sharepoint), maar is er ook voor gekozen om een aantal onderwerpen minder gedetailleerd te behandelen (zoals thesaurusbouw). In deze

gevallen wordt verwezen naar andere bronnen, waaronder de eerder genoemde voorganger van dit boek, waar de specifieke details kunnen worden bestudeerd.

Hoofdstuk 1

Toegankelijk maken van informatie

We zijn er helemaal aan gewend; snel informatie zoeken in een enorm aanbod. Op het web gebruiken we daar in de meeste gevallen een zoekmachine voor die in een fractie van een seconde resultaten geeft. Als de resultaten niet direct bevredigend zijn, worden we ongeduldig; we vinden het al vervelend als we naar de volgende pagina met resultaten moeten klikken of als we een nieuwe zoekactie moeten uitvoeren. Ook in meer gesloten omgevingen als een intranet of documentsysteem zoals SharePoint, willen we niet te veel tijd besteden aan het vinden van de juiste informatie of een bepaald document. De techniek helpt ons daarbij. Zoekmachines kunnen zoekacties vele malen sneller uitvoeren dan de mens dankzij automatisch aangelegde indexen, geavanceerde software, snelle processoren en oneindige opslagcapaciteit. Als gevolg hiervan zijn databanken, websites, digitale tijdschriften, boeken en e-books, maar ook interne kennissystemen, intranetten en systemen voor documentmanagement met slechts enkele muisklikken te doorzoeken. Over het algemeen zijn er altijd wel zoekresultaten, maar als daar geen bruikbare informatie tussen zit weten we niet of deze niet bestaat, niet toegankelijk is, of moeilijker te vinden dan we dachten. Ook weten we niet altijd of de gevonden resultaten compleet zijn. Bovendien kunnen de resultaten weinig relevant blijken. Om deze redenen worden verschillende instrumenten toegepast ter ondersteuning van of als alternatief voor een zoekmachine. Uit verschillende onderzoeken van onder andere IDC (Feldman & Sherman, 2004), AIIM en Gartner blijkt namelijk dat kenniswerkers tot wel 35% van hun tijd besteden aan het zoeken naar de juiste informatie. Het loont dus de moeite om verschillende instrumenten in te zetten om informatie beter vindbaar te maken. Het vindbaar maken van informatie werd traditioneel aangeduid met het begrip ontsluiting, maar dat woord wordt in de praktijk niet veel meer gebruikt. Ook in dit boek zullen we dat niet doen en spreken we over “vindbaar maken” en “toegankelijk maken”. Voordat we de belangrijkste methoden en technieken voor het toegankelijk maken in theorie en praktijk gaan behandelen, bespreken we in dit hoofdstuk een aantal basisbegrippen uit de informatiewetenschap.

ontsluiting

1.1 Informatie en documenten

Door het enorme aanbod en de geavanceerde devices zijn we eraan gewend geraakt om bijna ieder moment van de dag digitale informatie in

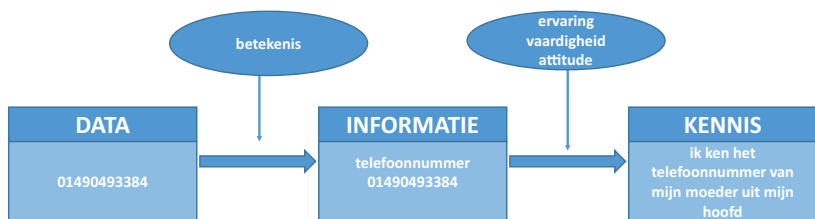
verschillende vormen te raadplegen zoals nieuwsberichten, verkoopcijfers, filmpjes, boeken, wikipedia's enzovoort. Met recht spreken we dan ook van een informatiemaatschappij. Het begrip informatie is echter niet zo makkelijk uit te leggen omdat je er op verschillende manieren naar kunt kijken. In de ICT ligt de nadruk op het feit dat we informatie blijkbaar snel over grote afstanden kunnen verplaatsen. James Gleick (2011) beschrijft in zijn boek *Informatie* hoe de vele technische ontwikkelingen op het gebied van natuurkunde en wiskunde (en vaak een combinatie daarvan) in enkele decennia hebben geleid tot de huidige mogelijkheden. De term informatie wordt in zijn huidige betekenis dan ook pas gebruikt vanaf 1948 als onder andere Claude Shannon, één van de grondleggers van de transistor, op zoek is naar een term voor de grote hoeveelheid boodschappen die op afstand kunnen worden verstuurd. In minder technische domeinen is sprake van andere benaderingen. Zo wordt informatie in de filosofie beschouwd als een wezenlijk onderdeel van ons bestaan. In de literatuur is het aantal definities dan ook oneindig, al naar gelang de nadruk. Wij leggen het begrip informatie uit in relatie tot enkele andere termen die eveneens een rol spelen in het vakgebied van de informatieprofessional.

Al vele jaren wordt in de informatiewetenschap de relatie tussen data, informatie, kennis en wijsheid uitgelegd aan de hand van het zogenoemde DIK(W) model. Hoewel dit model niet onomstreden is en er verschillende interpretaties over de details bestaan, is er wel een gemeenschappelijk idee dat we hier toelichten. Het model gaat uit van een opbouw in betekenis en waarde:

- | | |
|-------------------|---|
| data | <ul style="list-style-type: none"> • Data komt van het Latijnse werkwoord “dare”, dat geven betekent. Data betekent letterlijk: het gegevene, of gegevens. Het betreffen dan ook de kale gegevens, ook wel aangeduid met feiten: enen en nullen, cijfers of letters waar een gebruiker zonder toelichting niets mee kan, bijvoorbeeld 86534. Je ziet dat het cijfers zijn, maar het zegt verder niets. |
| informatie | <ul style="list-style-type: none"> • Zodra data betekenis krijgen, wordt het informatie. Als bij de cijfers 86534 vermeld wordt dat het om het netto maandsalaris van een voetballer bij een Spaanse club gaat, krijgt het betekenis en is er sprake van informatie voor de ontvanger. Deze betekenis wordt geboden door de context die bij de data vermeld is: netto, maandsalaris, euro's, Spanje en voetballer. Daarmee geeft informatie ook antwoorden op vragen. In het eerdergenoemde boek van James Gleick beschrijft hij hoe de beroemde cryptograaf Alan Turing na een gesprek met Claude Shannon uitroept: “Shannon wil een Brein niet alleen met gegevens volstoppen, maar ook met culturele dingen”. Shannon dacht dus meer in de toepassing van informatie- en kennissystemen waar het Turing nog vooral ging om zo efficiënt mogelijk opslaan en versturen van data. |
| kennis | <ul style="list-style-type: none"> • Of informatie ook tot kennis leidt, hangt sterk af van de ontvanger. Deze kan de informatie begrijpen en misschien direct weer vergeten, maar hij kan er ook actief mee aan de slag gaan. In het genoemde |

salarisvoorbeeld kan iemand de informatie gebruiken in een discussie over de vercommercialisering van het voetbal of bij een onderzoek naar inkomensverschillen in Spanje. Ontvangers kunnen niet alleen verschillende doelen hebben om de informatie te gebruiken, ze hebben ook verschillende soorten voorkennis en vaardigheden om de informatie te verwerken en van nieuwe context te voorzien.

Weggeman (1997) drukt de stap van informatie naar kennis dan ook uit in een soort formule: Informatie maal Ervaring, Vaardigheid en Attitude (I.EVA). Wie ontvangen informatie actief verwerkt, bouwt kennis op. Deskundigen op het gebied van kennismangement en kenniscreatie maken vaak onderverdelingen in soorten kennis. De bekendste onderverdeling is die van Nonaka (1994) die een onderscheid maakt tussen impliciete en expliciete kennis. Impliciete kennis is de persoonsgebonden kennis die iemand bezit. Dit kan zowel theoretische kennis (feiten, theorieën) betreffen als het bijna automatisch handelen omdat je dat ooit zo geleerd hebt, zoals fietsen of autorijden. Zodra de kennis wordt vastgelegd en overgedragen, spreken we van expliciete kennis. In de context van het DIKW model zouden we eigenlijk moeten spreken van informatie zodra de kennis door de zender is vastgelegd omdat het maar van de ontvanger afhangt of het werkelijk tot kennis leidt.



Afb. 1.1 - De stappen van data via kennis naar informatie.

wijsheid

- Wijsheid tenslotte is een abstracter begrip dat onder andere ingaat op morele aspecten. Het wordt dan ook wel aangeduid als “doing the right things”. Je ziet dat de stappen in het DIKW model niet alleen opschuiven in betekenis, maar ook van machinegebonden context naar de menselijke, psychologische context gaan. We zullen wijsheid dan ook niet verder uitdiepen omdat dat buiten het kader van dit boek valt.

content

Veel van de informatie die we gebruiken is opgenomen in documenten. Over de betekenis van het begrip document is veel discussie en er bestaan dan ook verschillende definities (zie bijvoorbeeld Buckland, 1991). In dit boek hanteren we een ruime definitie: vastgelegde en reproduceerbare tekst, geluid of beeld in analoge of digitale vorm beschouwen we als een document. Een webpagina valt dus ook onder deze definitie, evenals een filmpje op Youtube. Tenslotte bespreken we hier het begrip content. Letterlijk is content de inhoud. Bij een informatiesysteem kan dat van alles

contentsilo

zijn: documenten, tabellen, losse data. In de wereld van de marketing legt men de nadruk op het feit dat de inhoud een doel dient. Niet voor niets worden systemen voor het bijhouden van websites en intranetten ook wel contentmanagementsystemen genoemd. Deze platforms worden immers vaak beheerd door de afdeling marketing en communicatie. In het kader van dit boek beperken we ons niet tot alleen een webomgeving en maken we ook geen onderscheid tussen inhoud die wel of geen doel dient. We hanteren dan ook de volgende definitie van content: datgene wat als data of informatie in een informatiesysteem is geplaatst. Het begrip contentsilo is een populaire term voor een omgeving waar je content bewaart. Daarbij kun je denken aan een informatiesysteem als een intranet, website of documentmanagementsysteem. Ook onderdelen van zo'n systeem kunnen met het begrip contentsilo worden aangeduid. Wie constateert dat de netwerkschijf een onduidelijke mappenstructuur heeft, kan verzuchten dat hij zich een weg door allerlei contentsilo's (in dit geval dus de mappen) moet ploeteren om het juiste document te vinden.

1.2 Opslaan om gevonden te worden

Dit boek gaat over het vindbaar maken van informatie. Om dit zo goed mogelijk te doen, is het belangrijk om ons te verplaatsen in het gedrag van de zoeker en de methoden die hij hanteert. In het volgende hoofdstuk wordt daarom aandacht besteed aan het zoekproces en de overwegingen die daarbij een rol spelen. Daarop vooruitlopend kan worden vastgesteld dat met name in een bedrijfsomgeving het resultaat van een zoekactie vaak compleet en efficiënt moet zijn. Dat betekent dat er geen documenten mogen ontbreken (compleetheid) en dat er niet te veel irrelevante documenten moeten worden getoond als resultaat van een zoekactie (efficiency). Bij het gebruik van een zoekmachine die documenten fulltext doorzoekt, gaat dat nog weleens mis, met name als er wordt gezocht op onderwerp. Dit kan bijvoorbeeld gebeuren als een relevant document niet de term bevat waarop werd gezocht. Het zoekresultaat is dan incompleet. Ook kan een zoekmachine documenten tonen die misschien wel de zoekterm bevatten, maar verder niet over dat onderwerp gaan, hetgeen dan ruis oplevert. Wie bijvoorbeeld alle verslagen van een projectgroep zoekt, kan in de zoekmachine de zoekterm 'verslag' intikken. Daarmee bestaat de kans dat er documenten gemist worden waar niet het woord 'verslag' wordt gehanteerd, maar bijvoorbeeld 'notulen' of 'besluitenlijst'. Aan de andere kant zal niet ieder document waar het woord 'verslag' in voorkomt ook werkelijk een verslag zijn. Dit alles is het gevolg van het feit dat de zoekmachine letterlijk zoekt in alle woorden die in een documentcollectie aanwezig zijn. De zoekmachine verzint geen termen om de inhoud van documenten beter te definiëren en interpreteert het resultaat niet of nauwelijks. Om dit gebrek te ondervangen kunnen termen aan documenten worden toegekend waarna de zoeker met behulp van deze termen gericht

te werk kan gaan. Degene die het document plaatst, formuleert dan betekenisconcepten in de vorm van termen om het document toegankelijk en vindbaar te maken voor de zoekers. Hoewel bij tekstdocumenten de tekst wordt bestudeerd om de betekenisconcepten te formuleren, kan het gebeuren dat de toegekende term uiteindelijk niet voorkomt in het document. Bij beeldmateriaal kunnen de interpretaties sterk uiteenlopen, zeker als er geen regels zijn gesteld aan de wijze waarop de betekenisconcepten moeten worden vertaald in termen zoals het geval is bij de meeste deelsites voor foto's. Een foto van de Eiffeltoren krijgt dan bijvoorbeeld niet alleen de term 'Eiffeltoren' en in veel gevallen (maar hoogstwaarschijnlijk niet alle) 'Parijs', maar individuele personen zullen ook andere termen toekennen zoals bijvoorbeeld 'vakantie', 'zomer' of 'oom Kees'. Bij tekstdocumenten doen zich vergelijkbare problemen voor. Wie een boek met informatie over locomotieven en trams van een term moet voorzien kan kiezen uit woorden als: 'treinen', 'spoorwegen', 'locomotieven', 'trams', 'railvervoer', enzovoort. Als er dan ook nog termen bij komen die betrekking hebben op de beschreven landen of het tijdperk, ontstaat er al snel willekeur. Bovendien is de keuze van degene die de term heeft toegekend niet per definitie dezelfde als die van de zoekers. In dit boek bespreken we methoden en afspraken om dit soort problemen te ondervangen.

1.3 Toegankelijkheid

ordering

Wie grote hoeveelheden content beheert, moet dus een methode vinden om de afzonderlijke documenten te onderscheiden en vindbaar te houden. Bibliotheken en archieven doen dat traditioneel op twee manieren: via ordening en door middel van een catalogus. Met ordening bedoelen we dat collecties door middel van rubrieken en subrubrieken worden georganiseerd. Een eenvoudige vorm van ordening van papieren collecties is op formaat of alfabetisch op naam van de auteur. Een ordening op onderwerp vraagt vaak om het gebruik van een codering om de onderwerpen van elkaar te scheiden en onder te verdelen. Op basis van de gekozen ordening kan de gebruiker gericht zoeken in de kast omdat bijvoorbeeld de boeken van één auteur of die over een bepaald onderwerp bij elkaar staan. We noemen dat de directe toegankelijkheid. Een andere mogelijkheid is om met behulp van een catalogus of inventaris vanuit verschillende zoektermen zoals 'auteur', 'onderwerp' of 'titel' te worden doorverwezen naar de plaats van het document, de zogenoemde indirecte toegankelijkheid. De gebruiker heeft dan een tussenstap (catalogus of inventaris) nodig om de vindplaats te achterhalen.

directe toegankelijkheid

indirecte toegankelijkheid

De technieken voor zowel de directe (ordering) als indirecte toegankelijkheid van papieren collecties kunnen we grotendeels ook toepassen in digitale omgevingen. Wat betreft de directe toegankelijkheid via ordening kun je bijvoorbeeld denken aan de menustructuur van

een website of de mappenstructuur op een (netwerk)schijf. Voor de indirecte toegankelijkheid gebruiken we in een digitale omgeving een zoekmachine. Daarbij kunnen we een aantal regels rondom het toekennen van trefwoorden voor het bouwen van een kaartcatalogus goed gebruiken als we de resultaten van een zoekmachine willen verbeteren. Je ziet dus dat directe toegankelijkheid en indirecte toegankelijkheid ook in een digitale omgeving kunnen worden ingericht. Toen internet opkwam, lag de nadruk in eerste instantie op directe toegankelijkheid via ordening. Yahoo, en later de Nederlandse Startpagina, ontwierpen rubrieken om zoveel mogelijk websites in categorieën onder te brengen en aan te bieden. Dat kwam tegemoet aan de menselijke behoefte om overzicht te creëren. We kunnen wel zeggen dat Google daar rigoureuus een einde aan heeft gemaakt. Overigens was Google zeker niet de eerste zoekmachine, maar de resultaten op de eerste pagina's waren bij Google vaak wel relevanter dan die bij andere zoekmachines. Zoekmachines zijn dus een vorm van indirecte toegankelijkheid en maken een document op vele manieren vindbaar. Michael Wesch laat in zijn filmpje (Wesch, 2007) feilloos zien dat het denken in rubrieken op internet voorbijgaat aan een belangrijk kenmerk van de digitale omgeving: documenten zijn niet gebonden aan één fysieke plaats. Door middel van een zoekmachine of het gebruik van trefwoorden zijn documenten via heel veel ingangen terug te vinden.

ontlenen

toekennen

isness

aboutness

ofness

Een zoekmachine doet dat automatisch: alle woorden in een tekst worden door de zoekmachine geïndexeerd. Dat wil zeggen dat de zoekmachine alle woorden die hij tegenkomt in een eigen lijst plaatst, met verwijzingen naar de documenten waar dat woord in voorkomt. De zoekmachine bedenkt zelf geen nieuwe woorden, maar haalt ze dus uit de betreffende teksten op de webpagina's. We noemen dat ontlenen. Bij de informatieprofessional die de titel van een boek overneemt om in de catalogus op te nemen is eveneens sprake van ontlenen. Op het moment dat hij echter zelf een trefwoord bedenkt om aan te geven waar het boek over gaat, is sprake van toekennen van een term. We zien dus dat bij indirecte toegankelijkheid sprake kan zijn van zowel ontlenen als toekennen van termen waarop een document gevonden kan worden. De grens tussen toekennen en ontlenen is niet altijd duidelijk te trekken. De woorden in de titel kunnen immers ook de inhoud goed weergeven en als trefwoord worden gebruikt. Het onderscheid tussen ontlenen en toekennen van termen betreft de wijze waarop de termen worden geworven. We kunnen ook een onderscheid in termen maken op basis van het karakter van de termen. We maken daarbij onderscheid tussen isness, aboutness en ofness. Met isness bedoelen we de termen die aangeven van welk type document formeel sprake is (zoals 'boek', 'rapport', 'artikel'). Bij aboutness gaat het om het uitdrukken van het onderwerp (of onderwerpen) waar het document over gaat. In het geval van tekstdocumenten is het onderscheid tussen isness en aboutness voldoende. Bij het beschrijven van afbeeldingen is behoefte aan een derde categorie, de ofness. Deze behoefte komt voort uit de verschillende niveaus waarop je kunt beschrijven wat een afbeelding uitdrukt. Een

afbeelding van een vogel kan als beschrijving krijgen: ‘vliegen’, ‘meeuw’, ‘blauwe lucht’. Deze beschrijvingen vallen onder de ofness. Beschrijvingen van deze afbeelding die vallen onder aboutness zijn bijvoorbeeld ‘vrijheid’ of ‘ruimte’. In een apart kader gaan we verder in op de verschillende niveaus bij het beschrijven van afbeeldingen.

1.4 Metadata

metadata

We hebben het hiervoor gehad over (in)directe toegankelijkheid en het ontlenen of toekennen van termen om de isness en aboutness van een document te beschrijven. De termen die we hiervoor gebruiken noemen we metadata. Letterlijk betekent metadata ‘data over data’: het zijn woorden of codes die gaan over een bepaald document. Metadata kunnen worden gebruikt om een document te beschrijven, maar ook om het beter vindbaar te maken. We illustreren dat aan de hand van een muziekcollectie.

Wie vroeger een cd-collectie in de kast had staan, zal daar over het algemeen geen catalogus van hebben bijgehouden. Toch was het eenvoudig om snel de gewenste titel te vinden in de min of meer geordende collectie. De cd’s konden bijvoorbeeld alfabetisch op naam van de artiest of band zijn geplaatst. Daarmee was het een vorm van directe toegankelijkheid. In een digitaal muziekbestand willen we echter onze muziek op verschillende manieren kunnen terugvinden: op titel, componist, artiest, genre, album, enzovoort. Om dat te kunnen, heeft de gebruiker of de aanbieder gegevens aan het geluidsbestand toegevoegd die het bestand beschrijven en identificeren. Deze gegevens over een item worden dan metadata genoemd. In de jaren ‘70 van de 20e eeuw begonnen ontwikkelaars van databasesystemen de term metadata te gebruiken voor de gestructureerde gegevens die ze in de database vastleggen. Hoewel dat dus al een tijd geleden is, beschouwen we metadata nog altijd als een moderne term voor een principe dat al vele eeuwen bestaat. Archieven en bibliotheken doen al sinds hun ontstaan in de oudheid aan metadatering. De eerste kleitabletten waarop inventarissen werden vastgelegd, de klassieke kaartenbakken en andere, ook digitale, catalogi bevatten allemaal gegevens over objecten of documenten.

Omdat in feite iedere vorm van beschrijvende gegevens als metadata kan worden beschouwd, worden metadata in veel verschillende omgevingen en voor uiteenlopende doeleinden toegepast. Er zijn dan ook talloze vormen van metadata die volgens verschillende invalshoeken kunnen worden ingedeeld.

De volgende indeling is een vrij algemene:

- Administratieve metadata. Dit zijn gegevens zijn van belang bij het beheer van documenten en gegevensobjecten zoals de datum van aanmaak en eigenaar.
- Structurele metadata. Hiermee wordt vastgelegd of gegevensseenheden onderdeel zijn van een groter geheel, zoals een pagina uit een boek, of een relatie hebben met andere bestanden. Structurele metadata spelen bijvoorbeeld een belangrijke rol bij digitaliseringsprojecten omdat ze van afzonderlijke scans duidelijk maken waar deze bij horen en wat hun volgorde is.
- Technische metadata. Deze hebben een relatie met de techniek die de creatie en opslag van het document mogelijk maakt. Voorbeelden hiervan zijn het opslagformaat en de bestandsomvang. Vaak worden deze metadata automatisch door het systeem aangemaakt. Veel camera's geven bijvoorbeeld de gegevens over de sluitertijd en de gebruikte lens bij een foto al automatisch mee.
- Beschrijvende metadata. Met behulp van deze gegevens wordt de inhoud van het object beschreven. Naast het feit dat deze gegevens een informatief karakter hebben, bevorderen ze ook de terugvindbaarheid op inhoud van een document.

Bovenstaande metadata hebben allemaal hun eigen functie. Het hangt dus helemaal af van de aard van de collectie en het verwachte gebruik daarvan welke metadata worden gebruikt en op welke wijze. Ook de keuze voor de eigenschappen van een document die we binnen elke soort metadata willen vastleggen is afhankelijk van de situatie. Het is in de praktijk niet haalbaar en zeker niet efficiënt om alle denkbare eigenschappen van documenten in metadata vast te leggen. Bovenstaande opsomming kan wel helpen bij het inrichten van informatiesystemen. Door de lijst af te lopen voorkom je dat je bepaalde soorten metadata wellicht vergeet. Wie bijvoorbeeld in SharePoint kolommen voor metadatating gaat definiëren, kan deze lijst gebruiken om alle noodzakelijke soorten te dekken. Interessant is daarbij dat een deel van de metadata automatisch worden toegekend op het moment dat een document wordt opgeslagen. Met name administratieve en technische metadata kunnen voor een groot deel door het systeem worden bepaald en vastgelegd.

formele metadata

Informatiespecialisten hanteren over het algemeen een onderscheid in de verschillende soorten metadata dat enigszins van bovenstaande indeling afwijkt. Zij verdelen de metadata in formele metadata, inhoudelijke metadata en administratieve metadata waarbij het begrip metadata nog vaak wordt vervangen door 'gegevens'. Formele gegevens hebben, zoals de naam aangeeft, betrekking op de formele aspecten van een item: titel, auteur, uitgever, aantal pagina's, enzovoort. Veel van deze gegevens worden door de auteur of uitgever van het document vastgesteld voordat een gebruiker het onder ogen krijgt. In de meeste gevallen zijn

de auteursnamen keurig vermeld, maar informatiespecialisten hebben bij bepaalde typen documenten als rapporten, brochures en folders nog weleens problemen om alle gegevens te achterhalen.

Voor de gestructureerde vermelding van deze formele gegevens in een informatiesysteem als een bibliotheekcatalogus of andere database bestaan diverse internationale afspraken (standaarden). Doel van deze afspraken is om de metagegevens over publicaties wereldwijd te kunnen uitwisselen en uniform via het web beschikbaar te maken. Voorbeelden van dergelijke afspraken zijn:

- MARC21 (Machine Readable Cataloging), een standaard voor te gebruiken metadata-velden voor bibliografische beschrijvingen (MARC, 1999).
- AACR2 (Anglo-American Cataloguing Rules), standaardregels voor de wijze van vullen van de velden van bibliografische beschrijvingen (AACR2, 2005).
- RDA (Resource Description and Access), opvolger van AACR2, gebaseerd op het relationele FRBR-model (Functional Requirements for Bibliographic Records) (RDA, 2008; FRBR, 1997).
- Dublin Core, een standaard voor te gebruiken metadata-velden, oorspronkelijk ontwikkeld voor het eenvoudig beschrijven van webpagina's; nu in een veel breder toepassingsgebied in gebruik (DCMI, 1995). Dublin Core heeft vele afgeleiden, zoals de [Overheid.nl](http://overheid.nl) Web Metadata Standaard (OWMS) waar alle informatie van de overheid op het web aan moet voldoen.
- MODS (Metadata Object Description Schema), een standaard XML-schema voor te gebruiken metadata-velden, eenvoudiger dan MARC, maar uitgebreider dan Dublin Core (MODS, 2009).

Bovenstaande is slechts een zeer beperkte greep van de bestaande standaarden en de ontwikkelingen zijn nog volop gaande. In een apart kader wordt stilgestaan bij het initiatief Schema.org dat zich bezighoudt met de standaardisering van metadata op het web. Inhoudelijke gegevens beschrijven de werkelijke inhoud van een document. Bij tekstdocumenten geven zij antwoord op de vraag: 'waar gaat het over?'.

inhoudelijke metadata

Inhoudelijke metadata worden meestal toegekend door degene die een object of document opneemt in een informatiesysteem of upload op social media. We noemen dat toegekend indexeren (zie hoofdstuk 7.2).

toegekend indexeren

Hij of zij hanteert dan één of meer trefwoorden, een samenvatting, een classificatiecode of een combinatie daarvan om de inhoud van het item te beschrijven.

administratieve metadata

Administratieve metadata hebben betrekking op de beheersmatige aspecten van een item in het systeem waarin het wordt beschreven. Voorbeelden zijn recordnummer, jaar van opname in het systeem en prijs.

Het vastleggen van al deze metagegevens dient verschillende doelen. In de eerste plaats vormen de vastgelegde gegevens in geautomatiseerde omgevingen vaak (maar niet altijd!) zoekingen waarop een item kan worden gevonden. De zoekgang maakt het bijvoorbeeld mogelijk alle documenten van dezelfde auteur in één keer op te vragen en op de sites van webwinkels kunnen we metadata gebruiken om een groep vergelijkbare producten op te vragen. We noemen dit de collocatiefunctie. Vervolgens verschaffen de gegevens bij het document of product (aanvullende) informatie over een gevonden item. Als de metadata niet als zoekgang wordt gebruikt, maar als extra informatie, spreken we van de evaluatiefunctie. De gebruiker identificeert het document dan door middel van de metadata en komt er meer over te weten. Het kan dan voorkomen dat er informatie verschijnt waarop je graag had willen zoeken. De meeste webshops vermelden bijvoorbeeld wel bij een wekkerradio of er een hoofdtelefoonaansluiting aanwezig is, maar bieden dit niet als filter aan. Dit voorbeeld illustreert dat toegekende metadata niet automatisch tot zoek- of filtermogelijkheid leidt, maar dat daar een stapje extra voor nodig is in het informatiesysteem.

collocatiefunctie**evaluatiefunctie****locatiefunctie**

Administratieve gegevens in bijvoorbeeld een bibliotheekcatalogus bevatten onder andere de feitelijke vindplaats van het document. Deze informatie voorziet in de locatiefunctie.

We zien dus dat metadata verschillende functies kunnen hebben en dat we verschillende typen metadata kunnen onderscheiden. Omdat we in dit boek vooral het toegankelijk maken van informatie op inhoud behandelen, houden we de indeling aan die door informatiespecialisten wordt gehanteerd: formele, inhoudelijke en administratieve gegevens. We kunnen hiermee de gegevens die betrekking hebben op de inhoud van een document het makkelijkst onderscheiden.

1.5 Informatietaal

informatietaal

Standaarden voor metadata schrijven dus voor welke kenmerken (velden) van een object of document worden vastgelegd. Als gebruikers vervolgens vrij worden gelaten om de waarden van de velden zelf te bepalen, kunnen er problemen ontstaan omdat er verschillen zijn in interpretatie, woordkeuze, spellingswijze, enzovoort. Het kan in veel gevallen dus handig zijn om hierover afspraken te maken. De afspraken leveren uiteindelijk een vaste lijst met termen op waaruit kan worden gekozen als een document binnen een systeem vindbaar moet worden gemaakt. Als deze lijst termen bevat voor de inhoudelijke aspecten van een document, spreken we van een informatietaal, in tegenstelling tot de taal die wij dagelijks spreken en schrijven, de natuurlijke taal. Een informatietaal kan worden beschouwd als een kunstmatige taal voor het ontsluiten van de inhoud en bepaalde formele aspecten van documenten (Ligthart Schenk, 1993). Evenals een

natuurlijke taal bestaat een informatietaal uit een verzameling termen die gebruikt kan worden om onderwerpen uit te drukken, de vocabulaire, en regels hoe de termen te gebruiken, de grammatica.

Het vocabulaire kan bestaan uit:

**woordsystemen
thesauri
trefwoorden-
lijsten**

- Een selectie van woorden en woordgroepen uit een of meer natuurlijke talen, zoals bijvoorbeeld 'integratie', 'internet', 'Limburg', 'digitale bibliotheken' en 'dagverblijven voor gehandicapten'. In dat geval spreken we van woordsystemen als informatietaal. Voorbeelden van woordsystemen zijn thesauri en trefwoordenlijsten: vaste sets met termen die mogen worden toegekend.

classificaties

- Speciaal gevormde woorden (notaties of codes), die bestaan uit combinaties van letters, cijfers en andere tekens, zoals 316.42(493), AB78945. Informatietalen die hier gebruik van maken heten classificaties of systematieken.
- Combinaties van notaties en woorden uit de natuurlijke taal, bijvoorbeeld 316.32(492Utrecht)"18". Ook deze termen worden gebruikt bij classificaties.
- Woorden die rubrieken vormen waarin onderwerpen of documenten worden geplaatst, zoals bijvoorbeeld in de Open web directory (www.dmoz.org).

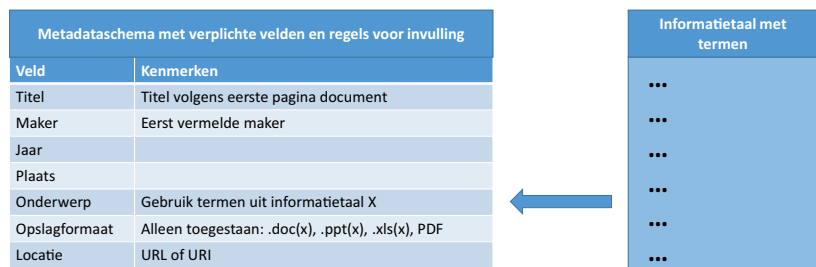
Bij classificaties worden aan de gevormde codes vaak omschrijvingen toegevoegd zodat de gebruikers direct de betekenis van een code weten. De code zelf dient zuiver om de systematiek weer te geven.

Wie de grammatica van een natuurlijke taal bestudeert, onderscheidt de volgende aspecten:

- de semantiek (betekenisleer, waaronder de relaties tussen termen,);
- de morfologie (vormleer, waaronder bijvoorbeeld spelling);
- de syntaxis (leer van de woordvoering/zinsbouw en het combineren van termen);
- de fonologie (klankleer).

Met uitzondering van de fonologie zijn deze aspecten ook bij informatietalen van belang. In hoofdstuk 5 zullen onderwerpen met betrekking tot de semantiek, de morfologie en de syntaxis aan de orde komen. Hoewel we hierboven hebben gezien dat een informatietaal bestaat uit een vocabulaire en grammatica, zien we in het dagelijks gebruik dat de term vocabulaire ook als synoniem voor informatietaal wordt gebruikt. Beheerders van informatiesystemen spreken in dat geval over het gebruik van een vocabulaire om de documenten consequent te beschrijven. Een ander veelgebruikt synoniem voor informatietaal is 'taxonomie'. In hoofdstuk twee worden de verschillende betekenissen van het begrip taxonomie uitgebreid beschreven en toegelicht.

Hiervoor hebben we een aantal standaarden genoemd voor de metadatering van documenten zoals RDA en Dublin Core. Daarmee lijken het op het eerste gezicht misschien informatietalen, maar er is een belangrijk verschil. De genoemde metadatastandaarden beschrijven vooral welke metadatavelden moeten worden gebruikt om een document te beschrijven en hoe deze moeten worden vastgelegd. In een beperkt aantal gevallen worden ook richtlijnen gegeven over de wijze waarop de velden moeten worden ingevuld, zoals de notatie van een datum of naam (bijvoorbeeld: achternaam, voorletters). Een informatietaal gaat echter uitsluitend over de termen die in een veld mogen worden ingevuld, dus de waarden die in het metadataschema worden genoteerd. Daarbij richt een informatietaal zich vooral op het veld van het metadataschema waar trefwoorden kunnen worden vermeld. Je zou dus kunnen zeggen dat een informatietaal termen en richtlijnen bevat om een onderwerpsveld in een metadataschema te vullen. In een aantal gevallen verwijzen metadata-schema's naar bestaande informatietalen die moeten worden gebruikt om het veld te vullen.



Afb. 1.2 - Onderscheid tussen metadataschema en informatietaal.

1.6 Bouwer, indexeerder en zoeker

indexeren

Het toekennen van (al of niet vrije) termen om een document terugvindbaar te maken, noemen we ook wel indexeren. De indexeerder kan bij zijn werkzaamheden vrije termen toekennen zoals de tags op internet, of gebruik maken van een informatietaal als een thesaurus of classificatie. De informatietaal met zijn eigen vocabulaire en grammatica is dan van tevoren ontwikkeld en vastgesteld. Het ontwikkelen (bouwen) van een informatietaal gebeurt meestal door een groep mensen die kennis hebben van:

- het onderwerpsgebied en documentcollectie waarvoor de informatietaal wordt ontwikkeld;
- de regels voor het ontwikkelen van een informatietaal;
- de voorkennis van de zoekers en de zoekmethoden die ze hanteren.

Om de detaillering (specificiteit) van een informatietaal te kunnen bepalen, is het belangrijk dat de bouwer goed op de hoogte is van de collectie die toegankelijk moet worden gemaakt en de terminologie die de gebruikersgroep hanteert. In het verleden werden gebruikers vaak niet of nauwelijks betrokken bij de bouw van een informatietaal. Dit leidde in veel gevallen tot een vocabulaire en systematiek waar de zoekers maar moeizaam mee uit de voeten konden. Tegenwoordig bestaat het risico dat we de techniek laten bepalen hoe we een document op inhoud vindbaar maken zonder dat we ons verdiepen in het zoekgedrag van de gebruikersgroep.

De genoemde rollen brengen verschillende bevoegdheden met zich mee en het is dan ook verstandig om dit onderscheid aan te houden. In veel organisaties kunnen (en moeten!) medewerkers zelf termen aan documenten toekennen als ze deze in een informatiesysteem plaatsen. Daarmee vervullen ze de rol van indexeerder. Als bij het toekennen van termen een informatietaal wordt gebruikt, is het belangrijk deze op orde te houden. Om te voorkomen dat er een wildgroei aan termen in de taal ontstaat, is het belangrijk dat de indexeerders niet de mogelijkheid hebben om zonder meer termen toe te voegen aan de informatietaal. Het recht om daar besluiten over te nemen moet zijn voorbehouden aan de bouwers/beheerders van de informatietaal.

Het indexeren van documenten is arbeidsintensief maar, gezien de besproken beperkingen van een zoekmachine, toch vaak nog de moeite waard. Een deel van de werkzaamheden kan worden geautomatiseerd. Zo kan een deel van de metadata automatisch worden ingevuld op basis van de documentbibliotheek waar het document wordt geplaatst of op basis van de persoon die het document opslaat. Ook bestaat er software die de inhoud van een document kan analyseren en op basis daarvan inhoudelijke metadata kan toekennen. In hoofdstuk 10 wordt deze techniek toegelicht. In de praktijk blijken deze methoden echter kostbaar en niet foutloos waardoor het bij veel informatieomgevingen nog steeds de moeite loont om handmatig metadata toe te kennen. Daarbij geldt de overweging dat een document slechts éénmaal wordt opgeslagen maar wellicht vele malen wordt opgevraagd zodat de tijdsinvestering van de indexeerder ten gunste komt aan vele zoekers.

1.7 Samengestelde onderwerpen

Als we de onderwerpen van documenten in termen willen vatten, komen we vaak omschrijvingen tegen als:

- verhuur van vakantiehuizen;
- Amsterdam in de 17^e eeuw;

- hygiëne onder verpleegkundigen in Nederland.

meervoudige concepten

Je ziet dat deze onderwerpen uit twee of meer onderdelen bestaan die niet zomaar in één woord te vatten zijn. We noemen dit meervoudige concepten. Bij de bouw van een informatietaal zal dan ook besloten moeten worden of deze concepten wel of niet in één term moeten worden opgenomen. De bouwer kan er bijvoorbeeld voor kiezen om ‘verhuur van vakantiehuizen’ als term op te nemen of als variant daarop ‘vakantiehuizen; verhuur’.

Een alternatief is het opnemen van zoveel mogelijk enkelvoudige begripseenheden in een informatietaal. In dat geval neemt de bouwer de termen ‘vakantiehuizen’ en ‘verhuur’ als aparte termen op in de taal. Als daar aanleiding toe is, kunnen deze termen dan worden gecombineerd door ze beiden aan een document toe te kennen.

precoördinatie

Wanneer de bouwer van de informatietaal een combinatie van concepten zoveel mogelijk onderbrengt in één samengestelde term, is er sprake van een zekere mate van precoördinatie. Daarmee heeft de bouwer ook de betekenisrelatie tussen de afzonderlijke concepten vastgelegd. Hij doet dit op basis van regels van de betreffende informatietaal. Bij de meeste classificaties en de traditionele trefwoordenlijsten wordt deze wijze van indexeren toegepast. In het voorbeeld van ‘Amsterdam; 17e eeuw’ schrijven de regels van de gebruikte informatietaal voor dat de tijdsaanduiding achter de onderwerpsaanduiding komt en dat deze worden gescheiden door een puntkomma.

Bij andere informatietalen, zoals een thesaurus, is sprake van termen die een enkel concept uitdrukken, zoals ‘verhuur’ of ‘vakantiehuizen’. Vervolgens wordt door de indexeerder een verzameling enkelvoudige begripseenheden aan het document toegekend, zonder dat de betekenisrelatie wordt vastgelegd. De zoeker legt dan zelf de relatie door in een zoekactie meer dan één term in te tikken en aan te geven dat de combinatie van termen het zoekcriterium is. Het document dat alle ingetikte termen heeft toegekend gekregen valt onder de treffers.

postcoördinatie

Omdat de relatie pas in de zoekfase wordt gelegd, noemen we dit postcoördinatie.

false coördination

In de praktijk blijken postcoördinatieve systemen goed te werken omdat de gebruiker gericht kan zoeken op afzonderlijke concepten en vervolgens naar behoefte allerlei combinaties kan maken. Toch is het risico van verkeerde verbanden (Engels: *false coördination*) niet uitgesloten. Een document met de termen ‘agressie’ en ‘hulpverleners’ kan zowel gaan over agressie ten aanzien van hulpverleners als over agressie die door hulpverleners wordt gepleegd. In hoofdstuk 2 behandelen we de verschillende methoden die kunnen worden gebruikt om informatie te zoeken en termen te combineren, onder andere met booleaanse operatoren.

Overigens is het niet zo dat in een thesaurus alleen maar termen voorkomen die een enkelvoudig concept vertegenwoordigen. Er mogen best termen in voorkomen met een bepaalde mate van precoördinatie. Denk aan termen als 'vrachtschepen', 'goederenvervoer', en 'goudhandel'. In hoofdstuk 4 bespreken we hoe daar in een thesaurus mee wordt omgegaan.

1.8 Samenvattend

In dit hoofdstuk hebben we de basisbegrippen uit de informatiewetenschap en de toegankelijkheidsleer besproken, zoals het DIKW model en het onderscheid tussen directe en indirecte toegankelijkheid. Daarbij zagen we dat ordening en metadata een belangrijke rol spelen in de toegankelijkheidsleer. De begrippen uit dit hoofdstuk zullen regelmatig terugkeren in de volgende hoofdstukken. Voordat we verder ingaan op de verschillende methoden en technieken voor het toegankelijk maken van informatie, zullen we ingaan op het zoekproces. Het is immers belangrijk om te weten welke zoektechnieken worden gehanteerd als we het de zoeker makkelijk willen maken. Hoofdstuk 4, 5 en 7 zijn helemaal gewijd aan het gecontroleerd toekennen van metadata en in hoofdstuk 6 worden de technieken voor ordening verder uitgediept. Omdat veel instellingen gebruik maken van SharePoint, hebben we daar een apart hoofdstuk aan gewijd. Vanaf hoofdstuk 9 bespreken we de nieuwste tools om informatie vindbaar te maken, waaronder het automatisch toekennen van inhoudelijke metadata.

Gebruikte literatuur

AACR2 (2005). *Anglo-American Cataloguing Rules, Second Edition, 2002, Revision: 2005 Update*. American Library Association (ALA), the Canadian Library Association (CLA), and the Chartered Institute of Library and Information Professionals (CILIP). ISBN 978-1-85604-570-4

Buckland, M. (1991), Information as a thing. *Journal of the American society for information Science*, 41(5), p. 351

DCMI (1995). *Dublin Core metadata initiative*. Geraadpleegd op: 10 maart 2010, van: <http://dublincore.org>

Feldman, S. & Sherman, C. (2004). *The high cost of not finding information*. IDC White paper. Opgevraagd 14 april 2016 van: <http://www.ejitime.com/materials/IDC%20on%20The%20High%20Cost%20Of%20Not%20Finding%20Information.pdf>

FRBR (1997). *Functional Requirements for Bibliographic Records: final report, as amended and corrected through February 2009*. Geraadpleegd op: 6 april 2010, van: http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

Gleick, J. (2011). *Informatie*. Amsterdam: Bezige bij. ISBN 9789023466178

Ligthart Schenk, A.H. (1993). *Basisboek documentaire informatieverzorging*. Den Haag: NBLC. ISBN 90-5483-051-4

MARC (1999). *Marc 21 format for bibliographic data*. Washington: Library of Congress. Geraadpleegd op: 10 maart 2010, van: <http://www.loc.gov/marc/bibliographic/>

MODS (2009). *Metadata object description schema*. Geraadpleegd op: 10 maart 2010, van: <http://www.loc.gov/standards/mods/>

Nonaka, I. 1994. A dynamic theory of organizational knowledge creation. *Organisational Science* 5(1) 14–37

RDA (2008). *Resource description and access*. Geraadpleegd op: 10 maart 2010, van: <http://www.rdaonline.org>

Weggeman, M. (1997). *Kennismanagement: Inrichting en besturing van kennisintensieve organisaties*. Schiedam: Scriptum. 9789055940875

Wesch, M. (2007). *Information R/evolution*. Geraadpleegd op 2 april 2016 van: <https://www.youtube.com/watch?v=-4CV05HyAbM>

Verder lezen

Taylor, A. (2004). *The organization of information*, 2nd ed. Westport: Libraries unlimited. ISBN 1-56308-969-6

Weinberger, D. (2007). *Everything is miscellaneous: the power of the new digital disorder*. New York: Times, Holt. ISBN 9780805080438

Hoofdstuk 2

Informatie zoeken

Dit boek gaat over het beter vindbaar maken van informatie. Dat is in zekere zin complementair aan zoeken naar informatie. Om dat vindbaar maken beter in perspectief te zien, moeten we dus ook wat meer over zoeken weten. Dat zoeken gaat soms om heel praktische gegevens. Hoe laat vertrekt morgenochtend een trein naar Maastricht? Waar kan ik deze camera het goedkoopst kopen? Maar vaak heb je veel inhoudelijker informatie nodig. Wie op zoek gaat naar informatie moet voortdurend keuzes maken. De keuze naar welk soort informatie je eigenlijk zoekt, de keuze hoe je vraag af te bakenen, de keuze waar je naar dat soort informatie gaat zoeken, de keuze hoe je vervolgens precies gaat zoeken. Voor sommige informatievragen, bijvoorbeeld voor die trein naar Maastricht, liggen de te maken keuzes nogal voor de hand. Voor andere zul je veel meer moeten weten over informatiesoorten, informatiebronnen, informatiecollections, maar vooral ook over zoekhulpmiddelen en toe te passen zoekstrategieën.

information retrieval

Wat we meestal gewoon “zoeken” noemen, wordt in de wereld van zowel informatieprofessionals als informatici veelal aangeduid als *information retrieval*. De Engelstalige Wikipedia geeft de volgende definitie van dit begrip:

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Daar wordt nog aan toegevoegd dat het zoeken naar die informatie gebaseerd is op fulltext, of andere op inhoud gebaseerde, indexering. Daar worden dus ook metadata mee bedoeld. Belangrijk element in deze omschrijving is ook het verkrijgen van de informatie: *retrieval* betekent onder meer ophalen of terughalen. Denk maar aan de golden retriever, het hondenras dat geacht wordt door jagers geschoten, maar uit zicht geraakte, eenden in de bosjes terug te zoeken én op te halen.

2.1 Informatiecollections en hun zoekhulpmiddelen

Informatiecollections zijn pas nuttig als er een zoekhulpmiddel bij is waarmee je de daarin opgeslagen informatie ook echt kunt vinden en er

1 Delen van dit hoofdstuk zijn ontleend aan een uitgebreider artikel over zoeken van de hand van een van de auteurs van dit boek (Sieverts 2011)

zoeksysteem

toegang toe kunt krijgen. Vrijwel elke informatiecollectie heeft dus ook een zoekstelsel dat daar bij hoort. Een paar voorbeelden:

- Een bibliotheekcatalogus is een collectie beschrijvingen van de boeken en andere objecten in een bibliotheek of mediatheek, met daarbij een zoekstelsel om te kunnen zoeken naar boeken van een bepaalde auteur, over een bepaald onderwerp, uit een bepaald jaar en dergelijke;
- De database van de Nederlandse Vereniging van Makelaars (op Funda.nl) bevat een collectie gegevens en beschrijvingen van huizen die op dat moment te koop zijn, met daarbij een zoekstelsel waarmee je kunt selecteren en zoeken op specifieke kenmerken die voor kopers van huizen van belang zijn;
- De Newsportal van LexisNexis bevat een collectie teksten van Nederlandse krantenartikelen, met daarbij een zoekstelsel dat er vooral op gericht is om artikelen over bepaalde onderwerpen te kunnen vinden, maar waar je je bijvoorbeeld ook tot een bepaalde krant kunt beperken;
- Ook een webzoekmachine als Google is een zoekhulpmiddel dat bij een bepaalde collectie hoort, namelijk de collectie van zoveel mogelijk (intussen al enkele honderden miljarden) webpagina's en andere op het web aanwezige documenten (PDF's, Word-documenten, PowerPoints e.d.) die wereldwijd toegankelijk zijn.

Voor een uitgebreid overzicht van informatiesoorten en hun zoekhulpmiddelen op internet, zie Sieverts (2014).

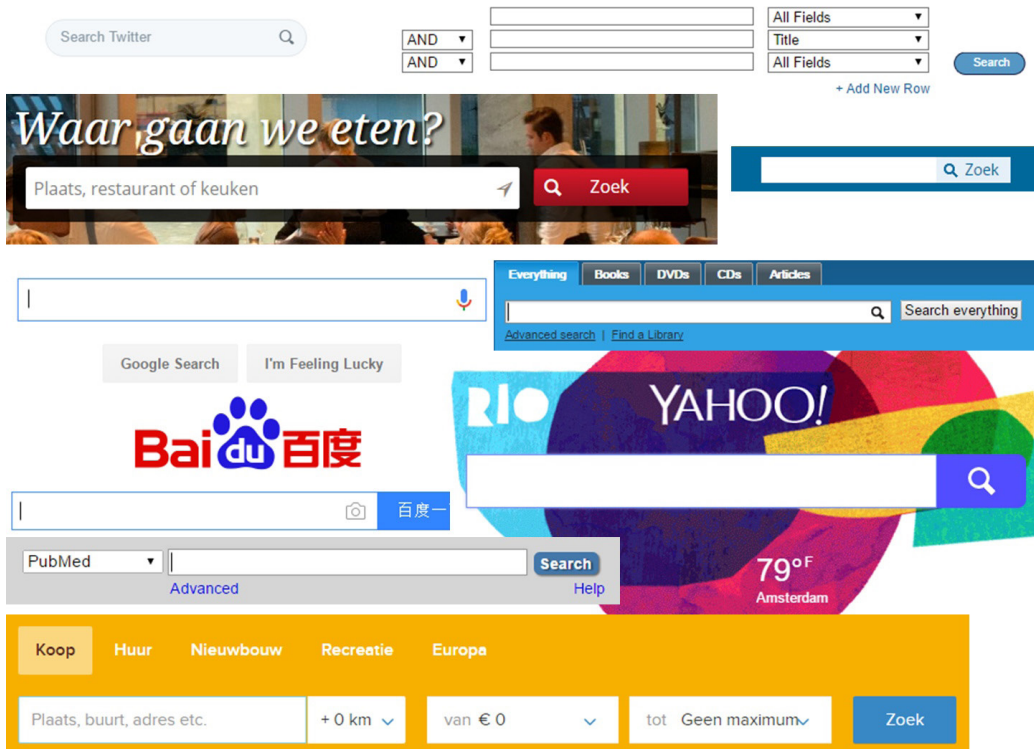
2.2 Aard van de zoekhulpmiddelen

**zoekhulpmid-
delen**

Uit de voorbeelden in de vorige paragraaf kon je al afleiden dat zoekhulpmiddelen hun eigen karakteristieken en specifieke zoekmogelijkheden hebben die direct aansluiten bij de aard en het doel van de informatie die in de bijbehorende collectie zit. Het zoekstelsel van een bibliotheekcatalogus biedt andere mogelijkheden dan een webzoekmachine: het zoekstelsel van Funda.nl zit heel anders in elkaar dan dat van de krantendatabase in Newsportal. Op detailniveau kunnen er dus veel verschillen zijn tussen de diverse zoeksystemen. Toch zijn ze, als je wat meer globaal kijkt, tot een paar basistypen terug te brengen:

1. Waar je bij het woord zoekstelsel waarschijnlijk het eerst aan denkt, zijn de systemen met zoekvensters, waarin je zelf iets moet invullen en waarbij je dus zelf zoekwoorden moet bedenken om je zoekvraag te formuleren. Dat ken je namelijk van Google. Maar ook de meeste bibliotheekcatalogi en een heleboel databases, zoals de al eerder genoemde Newsportal, werken zo. Dergelijke systemen zien er meestal vrij eenvoudig uit. De lege hokjes nodigen als het ware uit om iets in te vullen. Toch blijken zulke systemen in de praktijk vaak minder eenvoudig

en gebruiksvriendelijk te zijn dan ze er uitzien. Het is vaak moeilijk om de juiste vragen te stellen, want hoe weet je welke woorden precies gebruikt zijn in de documenten (of in de beschrijvingen van de documenten) waarnaar je op zoek bent? In principe zoeken deze zoeksystemen namelijk alleen maar letterlijk naar wat je als zoekvraag intikt. De problematiek van het ontwerp van de interfaces voor dit soort systemen wordt op een interessante manier behandeld door Peter Morville in zijn boek "Search Patterns" (Morville, 2010)



Afb. 2.1 - Voorbeelden van de variëteit aan zoekvensters die we in de praktijk tegenkomen.

2. Die problemen heb je vaak minder bij strakker gestructureerde databases. In de daarbij horende zoeksystemen kun je de meeste kenmerken waarop je wilt selecteren uit keuzelijstjes kiezen, zodat je daarvoor niets zelf hoeft te bedenken. In het systeem van Funda.nl kun je bijvoorbeeld uit lijstjes kiezen wat voor type woning je zoekt, in welke prijsklasse, met hoeveel kamers, met of zonder tuin, enzovoort. Daarvoor is wel nodig dat gestandaardiseerd, gestructureerd en consequent metadata voor die kenmerken zijn toegekend. Op de daarbij gebruikte keuzelijsten of autorisatielijsten komen we in de hoofdstukken 4 en 5 nader terug.

**onderwerpsin-
deling**

3. Ook voor nauwelijks gestructureerde informatie bestaan systemen waarin je niet zelf zoekwoorden hoeft te bedenken, omdat de informatie daarin al systematisch op onderwerp is ingedeeld. Zo'n systematische onderwerpsindeling is meestal een hiërarchische structuur, die zich steeds verder vertakt in specifiekere deelonderwerpen, zodat je de gewenste informatie vindt door browsen of navigeren door de hiërarchie van die onderwerpsindeling. Dergelijke systemen zijn dus vooral geschikt in situaties waarin je zelf moeilijk actief kunt omschrijven waarnaar je op zoek bent, maar waarin je wel (passief) in keuzemenu's herkent onder welk onderwerpsgebied de gezochte informatie waarschijnlijk te vinden is. Ook op internet kom je dergelijke indelingen wel tegen in de vorm van zogenaamde onderwerpsgidsen. Aan het zelf bouwen van dergelijke onderwerpsindelingen in de vorm van taxonomieën en classificaties wordt in de hoofdstukken 3 en 6 van dit boek aandacht besteed.

In dit hoofdstuk gaan we vooral in op de eerstgenoemde categorie zoeksystemen.

2.3 Zoekvragen

onderwerpsvraag

known item

Uitgangspunt van elke zoekactie is uiteraard de informatievraag. En informatievragen heb je in vele soorten en maten. Een belangrijke onderscheid is of het gaat om een echte onderwerpsvraag of om een zogenaamde known item vraag. Met dat laatste wordt bedoeld dat je bijvoorbeeld al weet dat een bepaald boek bestaat en nu op zoek bent of het in een bepaalde bibliotheekcollectie aanwezig is, of dat je weet dat er notulen van een bepaalde vergadering zijn en je die in een digitaal archief wilt terugvinden.

Een ander voorbeeld zijn de meeste vragen naar feitelijke gegevens: In welk jaar werd Mozart geboren? Bij welke temperatuur smelt ijs? Hoe ziet de vlag van Nicaragua er uit? Je weet dat het gegeven bekend moet zijn, alleen ken jij het (nog) niet. Je bent dan al tevreden met een getal, met een naam of met een plaatje dat het antwoord op je vraag vormt. Semantische zoeksystemen proberen dergelijke concrete antwoorden direct te geven door slimme interpretatie van zoekvragen. In de paragrafen 9.6.3 en 11.8 komen we daar nog nader op terug.

Bij een onderwerpsvraag ben je meestal op zoek naar informatie over een onderwerp, naar teksten waarin uitleg, meningen en (vaak toch ook wel) allerlei gegevens over dat onderwerp te vinden zijn. Die teksten geven meestal aanleiding tot bestudering, interpretatie, vergelijking en beoordeling. Bij veel van dit soort vragen is tevoren niet bekend of er wel informatie over dat onderwerp bestaat en hoeveel informatie dat dan

zou kunnen zijn, hoeveel je ook werkelijk zult kunnen vinden en hoeveel daarvan nuttig zal zijn.

Een belangrijk element van onderwerpsvragen is hoe diepgaand informatieonderzoek moet worden uitgevoerd. Zo zijn er situaties waarin een snelle oppervlakkige zoekactie volstaat, omdat je je bijvoorbeeld alleen wilt oriënteren op een bepaald onderwerp, bijvoorbeeld als voorbereiding op een diepgaander onderzoek. Je bent dan in feite al tevreden, zodra je maar iets hebt gevonden waaruit wat informatie over het onderwerp te destilleren valt, al is het maar alleen een Wikipedia-beschrijving. Men spreekt daarbij wel van quick & dirty zoekacties. Daarvoor hoef je in de meeste gevallen geen heel ingewikkelde zoekstrategieën toe te passen, al moet je uiteraard wel enige informatie over je onderwerp kunnen vinden. Er zijn echter vaak situaties waarin veel diepgaander onderzoek nodig is. Bijvoorbeeld als je van een bepaald onderwerp allerlei invalshoeken moet achterhalen. Of waar je zo volledig mogelijk alles moet vinden over een bepaald onderwerp om te zien of er nog onbekende aspecten zijn waar je zelf onderzoek naar zou kunnen doen.

2.4 Zoektechnieken

In deze paragraaf kijken we naar de technieken die in diverse zoeksystemen kunnen worden toegepast om informatie te zoeken. Daarbij beperken we ons tot de al eerder genoemde systemen met zoekvensters. Deze technieken kunnen zowel voor diepgaand onderzoek als ook bij een eenvoudige quick & dirty zoekactie worden toegepast.

2.4.1 Zoeken met woorden

Wie informatie zoekt met een zoekmachine doet dit door het intikken van termen. Dit kunnen woorden zijn, maar ook namen, afkortingen, getallen enzovoort. Voor de zoeker hebben deze termen een betekenis, maar voor de software gaat het om verzamelingen codes in een bepaalde volgorde. Zodra er een match is tussen de ingetikte reeks letters of cijfers en een woord of getal in een item in de database, wordt dit als resultaat teruggegeven. Daarbij kan het makkelijk voorkomen dat de zoeker een tikfout maakt of een spellingvariant gebruikt die de zoekmachine niet, of in beperkte mate, herkent, omdat die spelling in de doorzoekbare teksten niet of nauwelijks voorkomt. Dat heeft directe, en soms grote, gevolgen voor het zoekresultaat. Slimme software kan eenvoudige fouten corrigeren (zie paragraaf 11.2), maar dat gaat niet altijd goed. We weten dat één letter een groot verschil kan maken in de betekenis van een woord:

- schijffout, of schrijffout
- aansluiting of aanfluiting
- geen of een

- hard of hart,
- enzovoort.

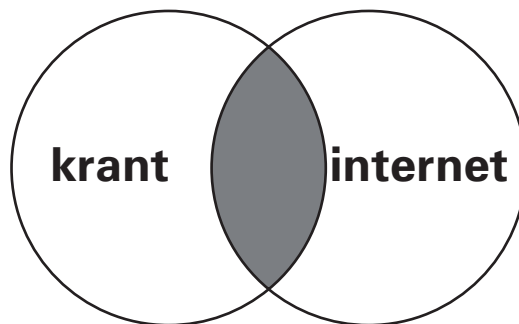
2.4.2 Booleaans zoeken

Bij het formuleren van zoekvragen worden heel vaak combinaties van meer woorden of zoektermen gebruikt. Bij het zoeken worden die woorden dan postcoördinatief met elkaar in verband gebracht. Het zal dan de bedoeling van de zoeker zijn om resultaten te krijgen waarin die woorden of begrippen samen voorkomen. In de zoeksoftware wordt op dat moment de zogenoemde AND-operator toegepast. De meeste zoekmachines op basis van vrije tekst op internet werken standaard op deze manier, ook zonder dat expliciet AND tussen de losse zoekwoorden hoeft te staan. Er zijn echter ook nog andere operatoren tussen de ingetikte woorden denkbaar. Deze operatoren zijn overgenomen uit de wiskunde en genoemd naar de Brit George Boole (1815-1864). We spreken dan ook over Booleaanse operatoren.

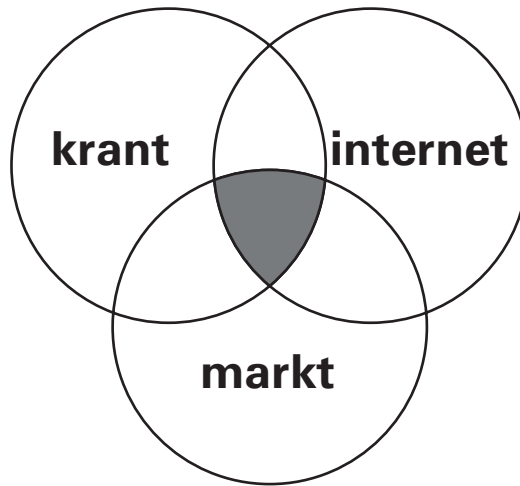
Bij de Booleaanse zoekmethode zijn standaard drie operatoren beschikbaar om zoekwoorden met elkaar te combineren, AND, OR en NOT.

Met de **AND-operator** geef je aan dat beide zoekwoorden in de te vinden documenten moeten voorkomen. Met bijvoorbeeld

krant AND internet



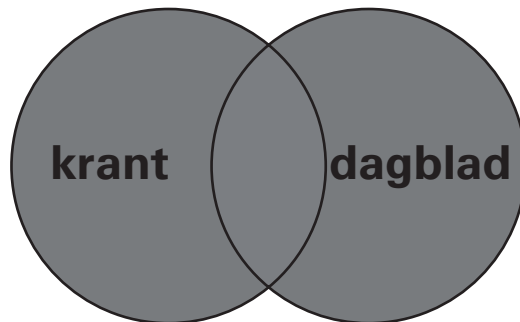
vind je dus alleen documenten waarin die twee woorden samen voorkomen. Een AND-operator zorgt dus voor een sterke inperking van je zoekresultaat. En hoe meer woorden je met AND combineert hoe strengere eisen je aan de te vinden documenten stelt, dus hoe minder je overhoudt.

krant AND internet AND markt

NB: Bij veel zoekmachines mag de AND-operator worden weggelaten. Het intikken van meer zoekwoorden in een zoekscherm voor “eenvoudig zoeken” impliceert daar al automatisch een AND-relatie

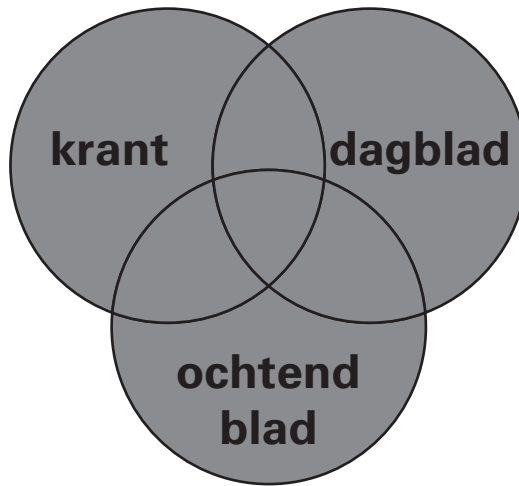
OR-operator

Met de **OR-operator** geef je aan dat maar één van beide woorden in een document hoeft voor te komen (maar beide mag natuurlijk ook). Met bijvoorbeeld

krant OR dagblad

vind je dus alle documenten waarin het woord ‘krant’ voorkomt, plus alle documenten waarin het woord ‘dagblad’ voorkomt (en dus ook die waarin ze allebei voorkomen). Een OR-operator zorgt dus voor uitbreiding van je zoekresultaat: hoe meer woorden je met OR combineert hoe meer je je zoekresultaat uitbreidt. De OR-operator bewijst zijn waarde bij het zoeken op (quasi)synoniemen of verschillende schrijf- of spelwijzen van woorden of namen.

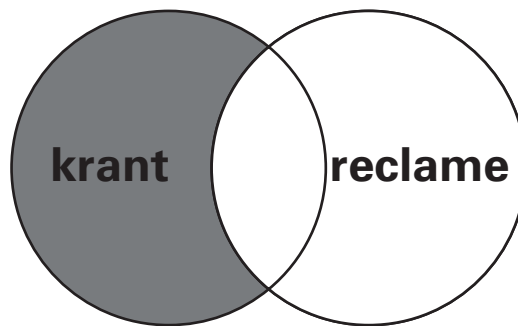
krant OR dagblad OR ochtendblad



NOT-operator

Met de **NOT-operator** geef je aan dat een bepaald woord beslist niet in je zoekresultaten mag voorkomen. Met bijvoorbeeld

krant NOT reclame



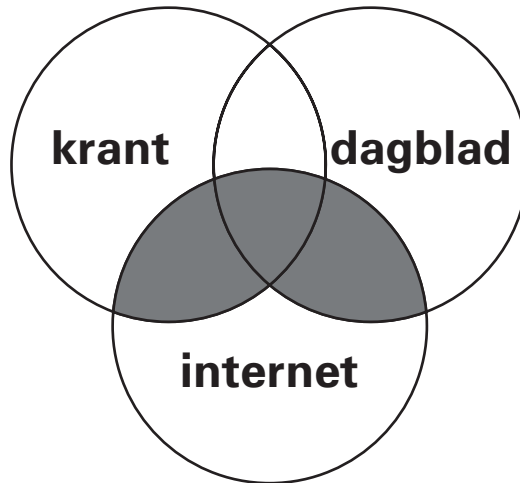
zal geen enkel document meer in je zoekresultaat zitten waarin het woord 'reclame' voorkomt, ook al staat het woord 'krant' er wel degelijk in. Dat is dus een (andere) manier om je zoekresultaat in te perken.

NB: bij sommige zoeksystemen moet je in plaats van NOT de operator ANDNOT gebruiken ("krant ANDNOT reclame"); bij Google en Yahoo moet je in plaats van NOT een minteken direct voor het ongewenste woord zetten ("krant -reclame").

In een zoekopdracht kun je ook verschillende operatoren combineren. Daarbij is het over het algemeen aan te raden haakjes te gebruiken om aan te geven welke zoekwoorden het eerst gecombineerd moeten

worden. Zonder die haakjes is het niet altijd duidelijk hoe een zoekstelsel je zoekvraag interpreteert. Dus bijvoorbeeld

(krant OR dagblad) AND-internet



om documenten te krijgen waarin het woord 'internet' voorkomt samen met het woord 'krant' of het woord 'dagblad'.

Met de Booleaanse methode kun je zoekvragen dus heel precies formuleren. Ook als je voor meer van je oorspronkelijke zoekwoorden synoniemen wilt toevoegen, is dat mogelijk. Zo zou je uiteindelijk bijvoorbeeld de volgende zoekvraag kunnen krijgen:

*(krant OR dagblad OR ochtendblad) AND (internet OR web)
AND markt NOT (reclame OR publiciteit)*

Nadeel van de klassieke Booleaanse zoekmethode is dat het nogal een "zwart-wit" methode is. Een document voldoet wel of niet aan je combinatie van zoektermen. Nuancering wordt daar in principe niet in aangebracht. De presentatievolgorde van zoekresultaten uit databases en bibliotheekcatalogi wordt daarom meestal alleen bepaald door het tijdstip dat de gevonden gegevens aan het systeem zijn toegevoegd; de meest recente gegevens het eerst. Dat laatste geldt niet voor webzoekmachines. Ook op resultaten van Booleaanse combinaties passen die hun reguliere methoden van relevantieordening toe, die we in de volgende paragraaf behandelen.

2.4.3 Best-match en relevantieordening

Webzoekmachines en ook wel sommige andere zoeksystemen passen een zoektechniek toe die wel wordt aangeduid met de term best match

best-match

**relevantievolg-
orde**

zoeken. Daarbij probeert het systeem te bepalen welke resultaten het meest overeenkomen (matchen) met de gestelde zoekvraag. Dat houdt in dat zoekresultaten ook in een bepaalde volgorde gepresenteerd worden. Omdat die mate van overeenkomst met de zoekvraag als een maat voor de relevantie van de gevonden informatie wordt beschouwd, wordt dit wel de relevantievolgorde (Engels: relevance ranking) genoemd. Hiermee wordt het “zwart-wit” schema van het klassieke Booleaanse zoeken doorbroken.

Wanneer je bij zoekmachines als Google maar een enkel woord als zoekvraag intikt, is het voor dergelijke systemen natuurlijk heel moeilijk om zonder verdere context te bepalen wat voor die vraag het meest relevant is. Daarom wordt in zulke gevallen eigenlijk alleen gekeken of je zoekwoord in de titel van een webpagina voorkomt, of wellicht in koppen in de tekst. Zo ja dan hebben die pagina's een grotere kans voor jou relevant te zijn, dan een pagina waar je zoekwoord onopvallend ergens onderaan de pagina staat. Daarnaast wordt ook het vermoedelijk belang van de op je zoekwoord gevonden pagina's afgewogen op grond van het aantal andere webpagina's dat daarnaar verwijst, dat een hyperlink daarheen bevat. Hoe die afwegingen precies geprogrammeerd zijn, is het geheim van elke zoekmachine. Iedere zoekmachine doet dat dan ook een beetje anders.

In de praktijk zul je zoekvragen vaak veel preciezer moeten afbakenen, waarvoor meer zoekwoorden nodig zijn. Dat maakt het voor een zoekmachine al makkelijker om de relevantie van gevonden pagina's op wat meer inhoudelijke gronden in te schatten. Er wordt dan van uitgegaan dat de woorden die een gebruiker als zoekvraag intikt, een soort zinnnetjes vormen. Hoe preciezer een webpagina dat zinnnetje bevat, hoe beter die waarschijnlijk met de zoekvraag overeenkomt en hoe relevanter die dus waarschijnlijk zal zijn. Onderlinge afstand en volgorde van de zoekwoorden in de gevonden webpagina's speelt daar dus een belangrijke rol.

Bij het bepalen van de relevantievolgorde kunnen onderstaande factoren een effect hebben op de positie van een document. Een document scoort hoger:

1. als de zoekterm(en) voorkomt in de titel van het document;
2. als de zoekterm(en) voorkomt in een paragraafkop of een opsomming;
3. als de zoekterm(en) voorkomt in de URL van het document;
4. naarmate de zoekterm(en) meer in het begin van het document voorkomt;
5. naarmate de zoekterm in het document wordt herhaald;
6. naarmate de zoektermen dichter bij elkaar (en in zelfde volgorde) voorkomen;
7. naarmate het bij de voorgaande punten om zeldzame zoektermen gaat;
8. naarmate meer hyperlinks naar een pagina verwijzen.

Als je meer woorden als zoekvraag intikt, vertoont een pagina die niet al die woorden bevat, maar bijvoorbeeld vier van de vijf, ook nog wel enige overeenkomst met de zoekvraag. Toch geven webzoekmachines in bijna alle gevallen alleen resultaten waarin de zoekwoorden allemaal voorkomen. De omvang van het web is namelijk zo groot dat zelfs een zoekvraag op acht woorden meestal nog zoveel resultaten met alle acht die woorden oplevert, dat er weinig aanleiding is ook nog resultaten te laten zien die op één na al die woorden bevatten. Bij zoekmachines voor lokale informatiesystemen, intranetten en kleine collecties, zullen resultaten met op-één-na alle zoekwoorden, vaak ook met lagere relevantie, in het zoekresultaat getoond worden. In paragraaf 11.4 komen we ook nog even op de hier genoemde relevantiefactoren terug.

2.4.4 Nabijheidsoperatoren

Hoe nuttig en noodzakelijk de AND-operator voor het doen van zoekacties ook is, er kleeft ook een nadeel aan. Dat twee woorden samen in een document voorkomen, betekent nog niet dat ze inhoudelijk ook echt iets met elkaar te maken hebben. Toch verwacht je dat meestal wel als je op zo'n combinatie zoekt. Dit is het probleem van bij postcoördinatie zoeken optredende *false coördination* die in paragraaf 1.7 al genoemd werd. Zeker in gevonden documenten met veel tekst kunnen twee met AND gekoppelde woorden ver bij elkaar vandaan staan en niets met elkaar te maken hebben. Bij systemen met relevantieordening probeert één van de ordeningsfactoren uit de vorige paragraaf dit al automatisch te compenseren. In sommige Booleaanse zoeksystemen kun je met nabijheidsoperatoren (Engels: adjacency) ook wat formeler aangeven dat gezochte woorden bij elkaar in de buurt moeten staan. Soms kun je daarvoor een NEAR-operator gebruiken.

NEAR-operator

krant NEAR internet

betekent dan dat de beide woorden niet alleen samen moeten voorkomen, maar dat ze bovendien op niet meer dan ~10 woorden afstand van elkaar mogen staan. Er is dan een veel grotere kans dat er een inhoudelijke relatie tussen die woorden bestaat. Soms kun je zelfs nog preciezer de maximale afstand aangeven. Bij LexisNexis, Ovid en (zelfs) Google kun je dat doen met respectievelijk

krant W/4 internet , krant ADJ4 internet , krant AROUND(4) internet

als je wilt aangeven dat de twee zoekwoorden niet verder dan vier woorden bij elkaar vandaan mogen staan. Deze operatoren zijn dus strenger dan AND en perken je zoekresultaat sterker in, maar ze zijn weer wat minder streng dan wanneer je met een exacte phrase zoekt op samengestelde begrippen.

2.4.5 Truncatie en woordstammen

truncatie

In onze voorbeelden is tot dusverre steeds alleen het enkelvoud van zoekwoorden gebruikt. Je mag er echter beslist niet van uitgaan dat elk zoekstelsel dan ook documenten vindt waarin die woorden toevallig alleen in het meervoud voorkomen. De voorbeeld-zoekvragen zullen in het algemeen dus nog zeer onvolledige resultaten opleveren. Het is natuurlijk onhandig om voor vollediger zoekresultaten zelf alle verschillende woordvormen met OR te moeten combineren. Daarvoor bestaan dan ook verschillende oplossingen. Veel zoeksystemen bieden de mogelijkheid om aan te geven dat je wilt zoeken op alle woorden die met een bepaald woordgedeelte beginnen, zogenaamde truncatie of in computerjargon ook wel “zoeken met wildcards”. Daarmee kun je niet alleen het probleem van enkel- en meervoud voorkomen, maar zoek je ook op andere woorduitgangen van je zoekwoord, zoals daarvan afgeleide werkwoordsvormen, bijvoeglijke naamwoorden of bijwoorden en hun vervoegingen of verbuigingen, alsook op samenstellingen met andere woorden. Een in veel systemen gebruikt symbool voor truncatie (de wildcard) is de asterisk *. Met

*krant**

zoek je dan dus ook op ‘kranten’, ‘krantjes’, ‘krantenknipsel’, ‘krantenkop’, ‘krantenpapier’, ‘krantenwijk’, enz.

spellingsvarianties

In sommige zoeksystemen kunnen ook binnenin een woord één of meer letters gemaskeerd worden, om op spellingsvarianties te kunnen zoeken. Met bijvoorbeeld

publi?aties

vind je dan beide schrijfwijzen van het woord, met een “c” of met een “k” (en in principe natuurlijk ook met andere letters op de plaats van het vraagteken, maar die zullen in dit geval niet voorkomen). Helaas bieden de bekende webzoekmachines niet dit soort mogelijkheden tot truncatie. Daar zul je dus zelf woordvarianten, met OR gecombineerd, in je zoekvraag moeten meenemen.

woordstam

Met truncatie kun je als zoeker zelf bepalen wat de woordstam is waarop je wilt zoeken, dus welk deel van het woord je wilt afkappen en door een truncatieteken vervangen. Er zijn ook systemen die dit automatisch proberen te doen. De eenvoudigste variant hiervan is, dat bij een in enkelvoud ingetikt zoekwoord ook automatisch op het meervoud wordt gezocht en omgekeerd. Een stapje verder gaan systemen die elk zoekwoord tot hun taalkundige woordstam reduceren, door standaard achtervoegsels van die woorden te verwijderen. Daardoor kan in één keer gezocht worden op de zelfstandig naamwoorden, de werkwoordsvormen, de bijvoeglijk naamwoorden en/of de bijwoorden

die op dezelfde woordstam gebaseerd zijn. Voor allebei die methoden heeft het zoekstelsel dus enige kennis nodig van de taal waarin gezocht wordt: het moet weten wat in een bepaalde taal de regels zijn voor meervoudsvormen of welke standaard uitgangen in die taal achter een woordstam kunnen voorkomen. Systemen die automatisch op woordstammen zoeken passen in de praktijk vaak alleen Engelse taalregels toe. Bij LexisNexis wordt zo automatisch op Engels enkel- en meervoud gezocht en zelfs op verschil tussen Engelse en Amerikaanse spelling: 'bicycle' vindt ook 'bicycles', 'behaviour' vindt ook 'behavior' en omgekeerd. Ook zoekmachines zoeken steeds vaker automatisch op woordstammen, waardoor je minder last hebt van het ontbreken van de mogelijkheid te trunceren. Zo vindt Google met de zoekterm 'vaccination' ook webpagina's waarin de woorden 'vaccinations', 'vaccine' of 'vaccinated' voorkomen. Voor steeds meer Nederlandse woorden gebeurt dat intussen ook. In hoofdstuk 11 komen we hier nog iets uitgebreider op terug.

2.5 Velden, filteren en facetzoeken

Bij veel zoeksystemen zul je standaard in alle digitaal beschikbare tekst zoeken. Bij zoekmachines is dat de volledige tekst van de webpagina's en andere soorten documenten die daarmee doorzoekbaar zijn. In dit soort documenten valt ook weinig verder te specificeren omdat er geen standaard inhoudelijke structuur in zit. Hooguit zijn er de formele titel en het URL als apart te doorzoeken onderdelen.

velden

Bij gegevens in bibliotheekcatalogi en andere databases ligt dat heel anders. Daarin zijn gegevens veel gestructureerder opgeslagen en van die structuur in zogenaamde velden kun je bij het zoeken gebruik maken. Voor de metadata in die velden wordt in veel gevallen gebruik gemaakt van gestandaardiseerd vocabulaire dat afkomstig is uit keuzelijsten, autorisatielijsten en thesauri. De gegevens in bibliotheekcatalogi en in bibliografische databases hebben een min of meer vergelijkbare structuur om de oorspronkelijke publicaties met metadata te beschrijven. Hoewel het om veel meer gegevens gaat dan alleen de titels van die publicaties, wordt ook wel gesproken van titelbeschrijvingen. In die titelbeschrijvingen vind je in elk geval altijd velden voor (uiteeraard) de titel, voor de namen van de auteurs en voor het jaar van publicatie. Afhankelijk van het beschreven soort document kunnen er ook velden zijn voor de uitgever van een boek, voor het aantal pagina's daarvan, voor de naam van het tijdschrift waarin een artikel is verschenen en voor de datum waarop, het afleveringsnummer waarin en de paginanummers waar dat artikel stond. Daarnaast is er vaak ook nog een veld voor inhoudelijke termen die aangeven waarover de publicatie gaat, voor rubrieken of categorieën op basis waarvan de collectie systematisch is ingedeeld en voor een samenvatting van de inhoud van de publicatie.

Voor het toekennen van metadata voor de formele kenmerken van publicaties wordt in veel gevallen gebruik gemaakt van keuzelijsten of autorisatielijsten: voor gestandaardiseerd toekennen van inhoudelijke kenmerken kunnen informatietalen als thesauri worden toegepast. In de hoofdstukken 4 en 5 komt dit nog uitgebreid aan de orde.

Veel van deze velden kunnen als zoekingang worden gebruikt. Dat betekent dat je bij het zoeken kunt aangeven dat je specifiek alleen in de inhoud van een bepaald veld wilt zoeken. De inhoud van sommige van deze velden, zeker waar het om gestandaardiseerde formele metadata gaat, wordt ook wel gebruikt voor het filteren van zoekresultaten.

Veel van wat hier over het gebruik van metadata gezegd is, geldt ook voor documentsystemen zoals SharePoint.

2.5.1 Zoeken in auteursveld

Bij auteursnamen is de achternaam de primaire zoekterm om op te zoeken. De voornaam of voorletter komt daar pas achter. Bij achternamen met voorvoegsels is niet altijd duidelijk wat daarmee gebeurt. In Nederland worden die meestal niet als onderdeel van de achternaam beschouwd (zoals Piet de Vries in het telefoonboek ook onder de V gezocht moet worden). In Engels- of Franstalige systemen en ook in België worden voorvoegsels meestal wel bij de achternaam gerekend en daar vaak ook aan vast geschreven. Het is in zulke gevallen dus altijd zaak verschillende mogelijkheden uit te proberen (Vries, P. de / De Vries, P. / Devries, P.).

Of je voor de voornaam alleen voorletters of hele voornamen moet gebruiken, kan per systeem heel verschillend zijn. Binnen een systeem zou wel naar standaardisatie gestreefd moeten worden. Toch blijkt dat bij veel bibliografische databases niet het geval te zijn, zodat je als zoeker zelf moet gokken hoe je alle publicaties van Eric Sieverts bij elkaar moet krijgen (E. / E.G. / Eric / Eric G. / ...?).

Bij bibliotheekcatalogi wordt er vaak naar gestreefd om controle uit te oefenen op de namen van (bekende) auteurs, zodat verschillende naamvormen en pseudoniemen van een zelfde auteur als zodanig herkend worden (bijvoorbeeld G.K. van het Reve = Gerard Reve = Simon van het Reve) en het dus ook niet uitmaakt op welke vorm je zoekt. Anderzijds kan daar onderscheid gemaakt worden tussen verschillende auteurs die toevallig exact dezelfde naam hebben.

2.5.2 Zoeken op titelwoorden

Bij bibliotheekcatalogi kun je meestal kiezen of je wilt zoeken op de titel of op titelwoorden. In het eerste geval moet je de precieze titel van een publicatie weten, of op zijn minst het begin daarvan. Een eventueel lidwoord aan het begin van een titel moet hierbij worden weggelaten.

Deze methode heeft uiteraard alleen zin voor een zogenoemde *known item search*. Veel handiger is het vaak om te zoeken op titelwoorden. Daarvoor hoef je alleen één of meer losse woorden in te tikken. Voor een *known item search* moeten dat natuurlijk een paar zeer karakteristieke woorden uit de (bekende) titel zijn. Voor een onderwerpsvraag gebruik je gewoon de zoekwoorden die je onderwerp omschrijven. Een reden om hierbij specifiek op titelwoorden te zoeken, is dat zo gevonden publicaties een veel grotere kans hebben werkelijk over het gezochte onderwerp te gaan.

Ook bij zoekmachines is het mogelijk om, als enige specifieke zoekingang, op titelwoorden te zoeken. De zoeksyntax om bij Google naar het woord 'krant' in paginatitels te zoeken, luidt:

intitle:krant

Bedenk daarbij wel dat het begrip titel voor webpagina's een veel minder formeel gegeven is dan bij boeken of tijdschriftartikelen. Het is het stukje tekst dat toevallig een HTML code <title> heeft meegekregen.

2.5.3 Zoeken op trefwoordvelden

De meeste catalogi en databases hebben een speciaal veld voor toegekende termen. Voordeel om daarvan bij het zoeken gebruik te maken, is dat je redelijk zeker weet dat daarmee gevonden documenten ook echt over het onderwerp gaan waarnaar je op zoek bent. Een niet onbelangrijk nadeel is dat de manier waarop dit soort termen wordt toegekend, en dus ook de manier waarop je daarmee moet zoeken, per systeem nogal kan verschillen.

In veel systemen zijn de gebruikte termen ontleend aan een informatietaal. Ze komen dan uit een thesaurus of een andere standaard lijst met toegestane trefwoorden. Als zoeker moet je dan weten op welke woorden je wel en niet moet zoeken. In sommige systemen worden trefwoorden veel meer willekeurig toegekend. Wel wordt dan meestal besloten om daarbij alleen meervoudsvormen van woorden te gebruiken (of juist alleen enkelvoud).

Naast de standaardisatie van de gebruikte termen, heeft beschikbaarheid van een thesaurus met hiërarchische relaties nog een voordeel voor de zoeker. Al dan niet navigerend, kan hij in principe door de termen grasduinen en zo op ideeën komen, onder meer voor het vereiste niveau van specificiteit van te gebruiken termen. Ook kan hij vaak termen aanwijzen of aanvinken die hij bij de zoekactie wil gebruiken, waarmee tikfouten worden voorkomen. De software kan eventueel ingetikte niet-voorkeurstermen automatisch omzetten in de gewenste voorkeurstermen (vervangend zoeken). En last but not least is er de mogelijkheid van generiek zoeken. (Zie hoofdstuk 4 bij stap 4). In dat geval kan aan het

**vervangend
zoeken**

generiek zoeken

zoeksysteem opdracht worden gegeven om, behalve op de gekozen term, tevens te zoeken op één, op meer of op alle onderliggende specifiekere termen. Dit stelt echter wel hoge eisen aan de zuiverheid van de hiërarchische verwijzingen binnen de thesaurus.

Een voorbeeld hiervan is te zien in afbeelding 2.2. Daar kan de zoeker aangeven of hij alleen op de term ‘Addiction’ wil zoeken of dat hij één of meer onderliggende termen (Engels: *narrower terms*), of zelfs de bovenliggende term (Engels: *broader term*) ‘Behavior disorders’, mee wil nemen in zijn zoekactie. Door een vinkje in de kolom ‘explode’ worden alle onderliggende termen van een gekozen term in de zoekactie meegenomen.

Thesaurus for **Addiction**

Database: **PsycINFO**

Combine selections with:

OR

Continue >>

Select Term(s)	Subject Heading	Hits	Explode	Focus	Scope Note
[Back up in List]					
☒ Addiction		3890	☐	☐	
	[Broader Terms]				
☐	Behavior Disorders	7035	☐	☐	
	[Narrower Terms]				
☒	Alcoholism [+NT]	21264	☒	☐	
☐	Drug Addiction [+NT]	7380	☐	☐	
☐	Internet Addiction	228	☐	☐	
☐	Sexual Addiction	420	☐	☐	
	[Related Terms]				
☐	Craving	1239	☐	☐	
☐	Drug Abuse [+NT]	27189	☐	☐	
☐	Drug Usage [+NT]	11470	☐	☐	
☐	Pathological Gambling	1919	☐	☐	
☐	Workaholism	157	☐	☐	

Afb. 2.2 - Mogelijkheid tot generiek zoeken met behulp van de “Thesaurus of Psychological index terms”.

2.5.4 Andere velden om op te zoeken of filteren

Op welke velden je verder nog kunt zoeken is voor een deel afhankelijk van de specifieke catalogus of database. Ook zijn voor een gemiddelde zoekactie niet alle verdere velden altijd even nuttig. Een paar velden die soms wel nog van belang zijn, zeker als de inhoud daarvan via autorisatielijsten gestandaardiseerd is, zijn onder meer:

- Publicatiejaar. Zowel voor een *known item search* als voor onderwerpsvragen biedt dit de mogelijkheid om je resultaat, in combinatie met onderwerpstermen, in te perken op de meest waarschijnlijke of meest gewenste tijdperiode.
- Taal. Biedt de mogelijkheid om zoekresultaten in te perken op de gewenste taal of talen. Zeker bij databases waar onafhankelijk van de oorspronkelijke taal van een document, de metadata in één standaardtaal (meestal Engels) worden toegekend is dit een nuttige optie.
- Documentsoort. Biedt de mogelijkheid om zoekresultaten in te perken op alleen boeken, video's, rapporten, scripties, peer-reviewed artikelen, PowerPoints en dergelijke. Welke inperkingen geboden worden hangt af van de aard van de betreffende collectie.

The screenshot displays the Ovid search system's filter interface for the Medline database. The filters are organized into two main columns. The left column includes filters for Age Groups, Animal Types, CheckTags, Clinical Queries, Subject Subsets, and Year Published. The right column includes filters for Journal Subsets, Languages, Publication Types, Status, and Star Ranking. Each filter category has a list of options and a scroll bar to navigate through the list.

Age Groups

-
- All Infant (birth to 23 months)
- All Child (0 to 18 years)
- All Adult (19 plus years)
- Newborn Infant (birth to 1 month)
- Infant (1 to 23 months)

Animal Types

-
- Cats
- Cattle
- Chick Embryo
- Dogs
- Goats

CheckTags

-
- Animals
- Female
- Humans
- Male

Clinical Queries

-
- Reviews (maximizes specificity)
- Reviews (best balance of sensitivity and specificity)
- Therapy (maximizes sensitivity)
- Therapy (maximizes specificity)
- Therapy (best balance of sensitivity and specificity)
- Diagnosis (maximizes sensitivity)

Subject Subsets

-
- AIDS
- Bioethics
- Cancer
- Complementary Medicine
- Dietary Supplements

Year Published

-
- Last Year
- Last 2 Years
- Last 3 Years
- Last 4 Years
- Last 5 Years

Journal Subsets

-
- AIDS/HIV Journals
- Core Clinical Journals (AIM)
- Bioethics Journals
- Biotechnology Journals
- Communication Disorders Journals

Languages

-
- Czech
- Danish
- Dutch
- English
- Esperanto
- Estonian

Publication Types

-
- Letter
- Meta Analysis
- Multicenter Study
- News
- Newspaper Article
- Observational Study

Status

-
- In Data Review
- In Process
- MEDLINE
- Publisher

Star Ranking

-
- ***** Five Stars (10781)
- **** Four Stars (0)
- *** Three Stars (0)
- ** Two Stars (0)

Afb. 2.3 - Voorbeeld van de filters die het zoekstelsel Ovid aanbiedt voor de medische database Medline

filteren

Bij veel zoeksystemen kun je zoekresultaten filteren op dit soort gegevens. De inperkingen worden daartoe in aparte zoekhokjes met keuzelijstjes

aan de zoekvraag toegevoegd. Ook sommige webzoekmachines bieden deze mogelijkheid. In sommige zoeksystemen worden dit soort formele inperkingen in aparte filterschermen aangeboden, waarin de gewenste filters aangeklikt kunnen worden. Afbeelding 2.3 geeft daar een voorbeeld van.

facetzoeken

Filters kunnen meestal zowel voor- als achteraf op een zoekactie worden toegepast. Er is ook nog een speciaal soort filtering, dat alleen toe te passen is op een al verkregen zoekresultaat en dat facetzoeken, of ook wel parametrisch zoeken, wordt genoemd. Daarbij worden alleen die filters getoond die binnen het zoekresultaat nog tot resultaten leiden. Bovendien wordt vooraf in het filteroverzicht al vermeld hoeveel treffers nog overblijven als dat filter wordt toegepast. In afbeelding 11.4 (in hoofdstuk 11) is daar een voorbeeld van te zien.

2.6 Zoekresultaten

2.6.1 Kwaliteit van het resultaat

Bij zoekacties streef je uiteraard altijd naar een bevredigend eindresultaat. Om te kunnen bepalen wat in dit kader bevredigend is, moet je weten wat de aard van de informatiebehoefte was in de betreffende situatie, en bovendien moet je eigenlijk objectieve criteria hebben om dat eindresultaat te beoordelen. Over het algemeen willen gebruikers snel informatie vinden die:

- zoveel mogelijk compleet is: cruciale informatie mag niet gemist worden;
- relevant is: dat wil zeggen dat de gevonden informatie tegemoet komt aan de informatiebehoefte van de gebruiker;
- niet teveel is ondergesneeuwd door heel veel resultaten die niet relevant zijn;
- betrouwbaar is: in open omgevingen zoals het vrije web kan heel veel informatie worden gevonden die absoluut onbetrouwbaar is, dit in tegenstelling tot gecontroleerde omgevingen waar vaak redactie, controle en selectie wordt toegepast op de inhoud.

De zoeker die zijn zoekresultaten evalueert, kan tot de conclusie komen dat het resultaat tekortschiet ten aanzien van één of meer van deze criteria. Dit falen kan veel verschillende oorzaken hebben. Zo kan het aan de zoeker zelf liggen. De zoekvraag is dan te vaag geformuleerd of er zijn verkeerde of te beperkte bronnen, zoekingen en -termen gebruikt. Ook de gebruikte applicatie speelt een belangrijke rol. Hoe indexeert de zoekmachine, op welke velden kan worden gezocht en hoe wordt het resultaat gerangschikt en gepresenteerd aan de gebruiker? Tenslotte zijn er de taalkundige aspecten. De spelling, betekenis en context van woorden hebben een cruciale invloed op het zoekresultaat.

relevantie

Methoden om gevonden resultaten te beoordelen op relevantie en betrouwbaarheid vallen voor een groot deel buiten de scope van dit boek. We richten ons immers op het goed ontsluiten van informatie, niet zozeer op de beoordeling van zoekresultaten. In het kader van het onderwerpsgebied retrieval (zoeken en vinden van informatie) besteden bijvoorbeeld de auteurs Kulthau (1994), Saracevic (1996) en recenter Hjørland (2010) en (opnieuw) Saracevic (2007a, 2007b, 2015) aandacht aan het fenomeen relevantie. Voor wat betreft de betrouwbaarheid kunnen we stellen dat de redactie van een informatieomgeving die serieus werk maakt van een goede toegankelijkheid, zich ook verantwoordelijk zal voelen voor het aanbieden van betrouwbare, juiste informatie.

2.6.2 Opbrengst, vangst en precisie

We kunnen proberen te berekenen in hoeverre de resultaten van een zoekactie voldoen aan de eerder genoemde criteria compleetheid en relevantie. We hanteren hierbij de termen opbrengst, vangst en precisie.

opbrengst

De opbrengst (Engels: *yield*) is het totale aantal treffers dat een zoekactie oplevert. Op zich zegt de opbrengst zelf nog niets over de kwaliteit van de zoekactie.

vangst

De vangst (Engels: *recall*) is het aantal relevante items dat een zoekactie oplevert in verhouding tot het totaal aantal relevante items dat in het systeem aanwezig is. Vangst is dus dat deel van de aanwezige relevante documenten dat daadwerkelijk is gevonden, of in een formule: gevonden relevant/totaal relevant. Wanneer de vangst gering is, betekent dat dat veel relevante informatie gemist is en dus niet aan het criterium van compleetheid voldaan wordt.

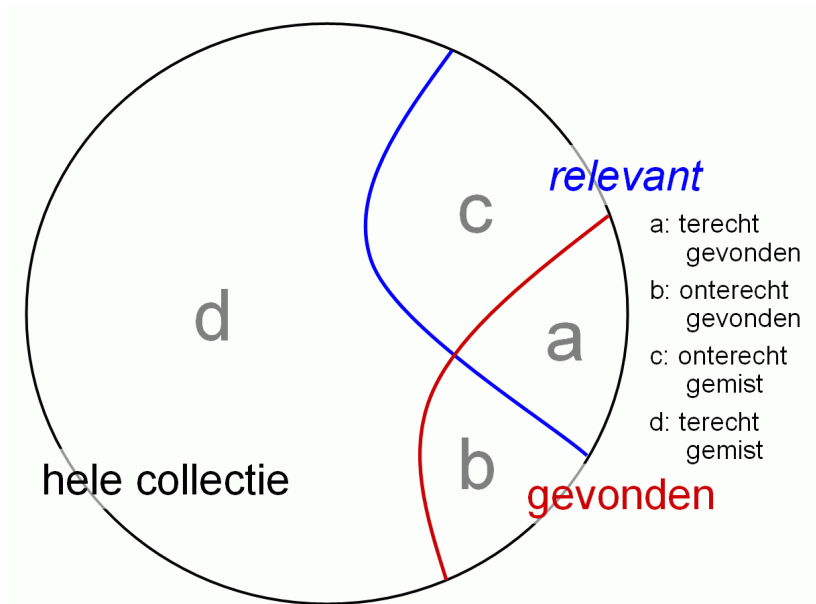
precisie

De precisie (Engels: *precision*) is het aantal relevante treffers in verhouding tot de opbrengst. Precisie is dus dat deel van de gevonden documenten dat ook daadwerkelijk relevant is, of in een formule: gevonden relevant/opbrengst. Wanneer de precisie klein is, betekent dat dat de relevante resultaten ondergesneeuwd raken tussen niet-relevant materiaal, waardoor tijdrovende handmatige na-selectie nodig is.

Aangezien het bij vangst en precisie om verhoudingen gaat, worden deze meestal in percentages uitgedrukt. Een rekenvoorbeeld:

Een intranet bevat 500 webpagina's. Hiervan zijn 50 items relevant voor een medewerker die op zoek is naar informatie over 'kwaliteitscontroles'. Een zoekactie levert 40 items op (opbrengst), waarvan er 25 over 'kwaliteitscontroles' blijken te gaan. Dat is de helft van het aantal relevante items in het hele intranet. In dat geval is de vangst dus: 25/50 of wel 50%. De precisie komt uit op: 25/40 of wel 62,5%. De 37,5% niet relevante items (15 stuks) die gevonden zijn, worden ruis genoemd.

In een schema ziet dat er uit als in afbeelding 2.4.



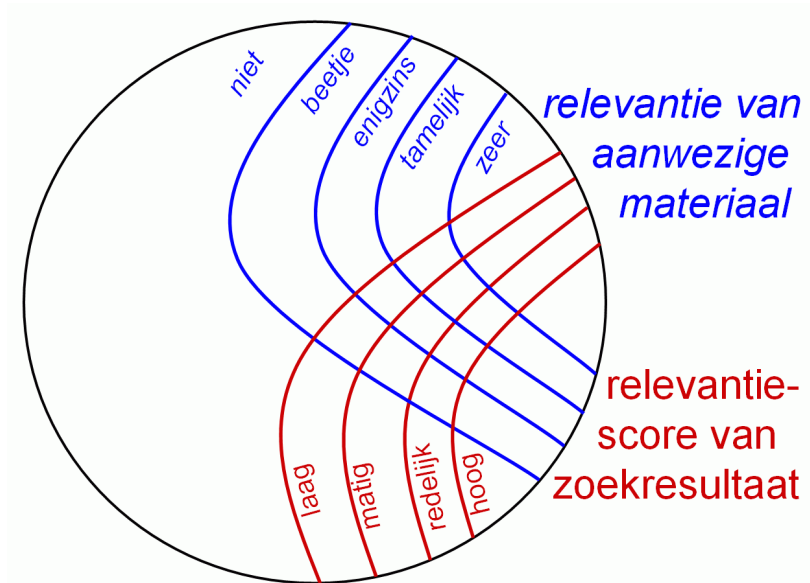
Afb. 2.4 - Vangst en precisie. De vangst is de verhouding tussen a en $(a+c)$. De precisie is de verhouding tussen a en $(a+b)$.

Het feit dat zowel vangst als precisie veelal in percentages worden uitgedrukt, suggereert dat het oordeel over zoekresultaten kwantificeerbaar is. In de praktijk zitten daar echter allerlei haken en ogen aan. Zo is relevantie een relatief begrip waarin nuances zijn aan te brengen. Dat geldt zowel voor het oordeel van de vraagsteller, als de door sommige zoekalgoritmes aan zoekresultaten toegekende relevantiescores. Afbeelding 2.5 illustreert dat. Vangst en precisie zijn dan niet meer in een eenvoudige formule uit te drukken.

Daarnaast is vooral vangst zeer slecht te kwantificeren, omdat die wordt bepaald door datgene wat relevant is, maar niet is gevonden. In de praktijk is echter vrijwel nooit bekend hoeveel relevante informatie bij een zoekactie is gemist of hoe groot de totale hoeveelheid relevante informatie $(a+c)$ is die over het gezochte onderwerp bestaat. Alleen in een gecontroleerde laboratoriumsituatie zijn voor testdoeleinden weleens gegevens daarover bekend. Redenen om het belang van het begrip vangst te relativeren zijn bovendien:

- In de huidige tijd van informatieovervloed is het lang niet altijd zo belangrijk dat werkelijk alles wordt gevonden.

- In de definitie van opbrengst en vangst wordt geen rekening gehouden met de uniciteit van de gevonden informatie: een zoekresultaat van tien relevante documenten die vrijwel identieke informatie bevatten, is in de praktijk niet waardevoller dan een resultaat met drie unieke, heel verschillende, elkaar aanvullende (of misschien zelfs tegensprekende) documenten. Van die drie is het dan veel erger als je er eentje mist dan van die tien.



Afb. 2.5 - Nuance in de mate waarin de treffers relevant zijn en door het zoekstelsel op relevantie geordend worden.

Wat voor een zoeker wel nuttig is om te weten, zijn de oorzaken waarom relevante informatie gemist wordt en welke technieken er zijn om daar (bij voorbaat of achteraf) iets aan te doen. Dat maakt het mogelijk een zoekactie aan te passen als het resultaat niet bevredigend lijkt te zijn. Veel van de in dit boek behandelde methoden voor het toegankelijk maken van informatie zijn er ook op gericht om het voor gebruikers makkelijker te maken tot een betere vangst te komen. Verder kan het in de praktijk vaak nuttig zijn om de vangst van twee verschillende zoekacties of zoekmethoden te vergelijken, zonder die in absolute zin vast te stellen. Er is dan dus sprake van de relatieve vangst van die twee zoekacties, op basis waarvan kan worden vastgesteld wat een betere aanpak voor de betreffende zoekvraag is geweest.

De precisie van een zoekactie is in theorie beter vast te stellen. Daarvoor hoef je namelijk alleen te kijken naar het materiaal dat daadwerkelijk is gevonden. Toch biedt ook dat begrip vooral schijnzekerheid:

- De situatie is niet altijd zo zwart-wit dat een gevonden document alleen maar OF relevant OF niet-relevant kan zijn.
- Bij relevantie (en dus precisie) wordt geen rekening gehouden met de kwaliteit en/of de betrouwbaarheid van de gevonden informatie.
- In de definitie van precisie wordt geen rekening gehouden met de kennis of de documenten die de aanvrager van de informatie al bezit; dus met de vraag hoeveel gevonden informatie voor die aanvrager werkelijk nieuw en/of nuttig is. Zo'n nieuwheidsfactor (Eng: novelty) is volledig gebruikersafhankelijk.
- Huidige methoden van relevantieordening maken het moeilijker eenduidig te bepalen wat je eigenlijk met precisie bedoelt: gaat het erom hoeveel van "de eerste 10" relevant zijn, of hoeveel van het totale zoekresultaat?
- In deze tijd van informatieovervloed is het in de praktijk vaak onmogelijk om alle, of zelfs maar een representatief deel van, gevonden documenten op hun relevantie te beoordelen (zoals de 100.000'en die Google beweert op een zoekactie te hebben gevonden).

Ondanks deze beperkingen geldt ook hierbij dat het voor een zoeker heel nuttig kan zijn inzicht te hebben in de oorzaken waarom een zoekresultaat ook niet-relevante informatie bevat, en in de technieken die je kunt toepassen om dat (bij voorbaat of achteraf) zo veel mogelijk te voorkomen (zie hoofdstuk 11). Bovendien speelt ook de precisieproblematiek een belangrijke rol bij vrijwel alle in dit boek besproken methoden voor het inhoudelijk toegankelijk maken van informatie.

2.7 Samenvattend

Om de toegankelijkheid in een informatiesysteem goed te organiseren, is het belangrijk om kennis te hebben van de verschillende zoekmethoden en -technieken die in de praktijk toegepast kunnen worden. We zagen dat er veel soorten zoekhulpmiddelen zijn, waarbij onder meer verschil bestaat tussen de mate van gestructureerdheid van de opgeslagen gegevens en de mate waarin die gegevens gestandaardiseerd zijn. Het zoeken zelf gebeurt met behulp van woorden, of we deze nu aanvinken of intikken. We proberen onze zoekvraag daarbij zo specifiek mogelijk te maken door woorden en velden te combineren. Hierbij vormen de Booleaanse operatoren handige hulpmiddelen. Geavanceerde zoeksystemen ondersteunen de gebruiker door zoveel mogelijk te interpreteren wat de zoeker bedoelt, zodat spellingvarianten en bijvoorbeeld meervoudsvormen automatisch in de zoekactie worden meegenomen. Ook proberen zoekmachines de mate van relevantie van documenten te bepalen. Om de kwaliteit van zoekacties te meten, kunnen we de opbrengst, vangst en precisie proberen vast te stellen, al kent dat wel een aantal beperkingen.

Gebruikte literatuur

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, 61(2), 217-237. Geraadpleegd op 21 augustus 2016 van: http://www2.hawaii.edu/~donnab/lis670/hjorland_relevance_2010.pdf

Kuhlthau, C.C. (1994). *Seeking meaning: a process approach to library and information services*. Norwood: Ablex. ISBN 1-56750-019-6

Morville, P., Callender, J. (2010). *Search Patterns; Design for Discovery*. O'Reilly Media. Geraadpleegd op 21 augustus 2016 van: <http://searchpatterns.org/>

Saracevic, T. (1996). Relevance reconsidered. *Information Science: Integration in Perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*. Copenhagen, 14-17 oktober 1996. Geraadpleegd op 21 augustus 2016 van: https://comminfo.rutgers.edu/~tefko/CoLIS2_1996.doc

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13), 1915-1933. Geraadpleegd op 21 augustus 2016 van: http://www2.hawaii.edu/~donnab/lis670/saracevic_relevance_part_II_2007.pdf

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13), 2126-2144. Geraadpleegd op 21 augustus 2016 van: https://ils.unc.edu/courses/2015_fall/inls509_002/papers/Saracevic07Part3.pdf

Saracevic, T. (2015). Why is relevance still the basic notion in information science. In *Re: inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015)* (pp. 26-35). Geraadpleegd op 21 augustus 2016 van: <https://comminfo.rutgers.edu/~tefko/Saracevic%20Why%20relevance%20-%20ISI2015.doc>

Sieverts, E. (2011). De informatie vinden die je zoekt; in: *Handboek Informatiewetenschappen, IWAbase, I-VF-660*. VakMediaNet, Alphen aan de Rijn. Geraadpleegd op 21 augustus 2016 van: <http://sieverts.pbworks.com/f/zoeken-en-vinden-2011.pdf>

Sieverts, E. (2014). Zicht op de veelheid aan bronnen op internet; in: *Handboek Informatiewetenschappen, IWAbase, I-VF-670*. VakMediaNet, Alphen aan de Rijn. Geraadpleegd op 21 augustus 2016 van: <http://sieverts.pbworks.com/f/internetbronnen.pdf>

KADER

SEO en SEA

Het zoeken binnen bedrijfsmuren of op een website (*enterprise search*) is een andere tak van sport dan het gevonden worden via internetzoekmachines zoals Google. Bij die laatste is het voor bedrijven belangrijk om zo hoog mogelijk in de resultaten te komen. Dat kan grofweg op twee manieren: de eerste is door middel van SEO (*search engine optimization*), de tweede optie is door middel van SEA (*search engine advertising*). Bij SEO probeer je onder andere door middel van goede content, links en metadata zo hoog mogelijk in de zoekresultaten te komen. Daarna probeer je gebruikers door middel van aansprekende titels te verleiden om op jouw link te klikken en je website te bezoeken. Het toepassen van deze SEO-technieken kost je niks. Bij SEA wordt gebruikt gemaakt van de advertenties boven en onder de natuurlijke zoekresultaten. Deze advertenties worden getoond als een gebruiker zoekt op de woorden die aan die specifieke advertentie gekoppeld zijn. De adverteerder betaalt een bepaald bedrag als er ook daadwerkelijk op een advertentie (die leidt naar de webpagina van de adverteerder) geklikt wordt. Verreweg de grootste aanbieder van advertenties bij zoekmachines is Google AdWords (<https://www.google.nl/adwords/>).

Of je nou kiest voor SEO of SEA, als je een website hebt, wil je dat je doelgroep jouw website (door een plek hoog in de zoekresultaten of via een betaalde advertentie) vindt op specifieke woorden. Als ik een kapper in Amersfoort ben, wil ik graag dat personen die zoeken op 'kapper' en 'Amersfoort' op mijn website uitkomen. Het is bij SEO/SEA dan ook de kunst om erachter te komen hoe jouw doelgroep zoekt. Want misschien zoeken je klanten wel op 'haarstylist' of 'nieuw kapsel' in plaats van 'kapper'. En mogelijk zoeken ze niet op 'Amersfoort', maar op een specifieke wijk ('Vathorst' of 'Kattenbroek').

Om erachter te komen hoe je doelgroep zoekt, kun je gebruik maken van een zoekwoordenplanner. De meest bekende en gebruikte is Google KeywordPlanner. <https://adwords.google.nl/KeywordPlanner>. Als je hier bijvoorbeeld 'kapper' intikt, dan krijg je een lijst met combinaties van woorden (ook wel zoekwoord-ideeën genoemd).

Naast elk zoekwoord wordt ook het 'gem. maandelijkse zoekopdrachten' getoond. Dat is een overzicht dat aangeeft hoeveel mensen per maand zoeken op dat woord. Met de juiste woorden kun je dan SEO-teksten schrijven of SEA advertenties laten plaatsen.

Zoekwoordplanner		Zoekwoorden vinden		Planning beoordelen	
Targting ¹⁾ Nederland ²⁾ Nederlands ³⁾ Google en zoekpartners ⁴⁾ Uitsluitingszoekwoorden ⁵⁾		Uw product of service kapper Meer opties Zoekopdracht aanpassen		Op deze pagina ziet u berekeningen voor zoekvolumes. Als u een uitgebreidere vergelijking wilt, stelt u een campagne in en voert u deze uit. Nagaan	
Periode ⁶⁾ Gem. aantal maandelijke zoekopdrachten voorgegeven voor afgelopen 12 maanden		Gemiddeld aantal maandelijke zoekopdrachten voor alle ideeën 100K – 1M		Adverteerdersprofielen 1) Zoekwoordideeën Kosten Downloaden Alle toevoegen (164)	
Zoekwoorden		Gem. maandelijke zoekopdrachten ¹⁾		Voorgeschat bid ²⁾	
kapper		10K – 100K Laag		€ 0,82	
				Rijen weergeven: 26 van 1 van 1 zoekwoorden	
Zoekwoord (op relevantie)		Gem. maandelijke zoekopdrachten ¹⁾		Voorgeschat bid ²⁾	
kapsalon		10K – 100K Laag		€ 1,05	
kapper zonder afspraak		100 – 1K Laag		€ 0,72	
kapsalon kapper		100 – 1K Laag		€ 0,80	
haar ontkleuren		1K – 10K Hoog		€ 0,61	
pruifelijt kapper		100 – 1K Normaal		€ 0,63	

Afbeelding: Zoekwoordideeën die de zoekwoordenplanner toont voor ‘kapper’ zijn o.a. ‘kapsalon’ en ‘kapper zonder afspraak’.

Hoewel de Google KeywordPlanner in eerste instantie bedoeld is voor SEO/SEA, kan het ook gebruikt worden bij het verzinnen van termen of synoniemen in een synoniemenlijst of thesaurus. Het kan je, naast het bevragen van je doelgroep of het analyseren van content, inspiratie geven voor relevante trefwoorden.

Verder lezen

Google Search Engine Optimization Starter Guide
<http://static.googleusercontent.com/media/www.google.com/nl/webmasters/docs/search-engine-optimization-starter-guide.pdf>

SEA: <https://nl.wikipedia.org/wiki/Zoekmachinemarketing>

Hoofdstuk 3

Ordening

In hoofdstuk 1 hebben we onderscheid gemaakt tussen directe en indirecte toegankelijkheid. Als er sprake is van directe toegankelijkheid dan is de informatie waarin we zoeken op een logische manier geordend, zodat we zonder hulp van bijvoorbeeld een zoekmachine snel bij de juiste informatie komen. Ordening is hierbij dus een belangrijk hulpmiddel. Bij indirecte toegankelijkheid gebruiken we een instrument als een catalogus of zoekmachine om te achterhalen welke informatie beschikbaar is en waar we deze kunnen vinden. Ook hierbij kan ordening een rol spelen. Bij een aantal van deze instrumenten kan namelijk een informatietaal worden ingezet om de vangst en precisie te vergroten. Deze informatietaal bestaat vaak uit termen die worden geordend door ze in relatie tot elkaar te presenteren. Denk bijvoorbeeld aan een overzicht met hoofd- en subtermen. In dit hoofdstuk gaan we dan ook dieper in op de verschillende technieken die we kunnen gebruiken om informatie en termen te ordenen.

3.1 Ordening en hiërarchie

Zowel in een informatieomgeving als in de fysieke wereld hebben mensen de behoefte om materialen te ordenen. De term hokjesgeest heeft een negatieve bijklank, maar onze hersenen hebben nu eenmaal de behoefte om alle informatie die tot ons komt op de één of andere manier overzichtelijk te houden. Dat gebeurt vrijwel continu, bijvoorbeeld als we een ruimte binnenlopen en we alle objecten die we zien onmiddellijk (en gedeeltelijk onbewust) indelen in categorieën zoals meubels, verlichting, decoratie en personen. Als gevolg van de ordening kunnen we spullen vervolgens makkelijker terugvinden. We zorgen daarmee voor een goede directe toegankelijkheid van de objecten of documenten. Tenslotte willen we niet te veel tijd kwijt zijn aan het zoeken naar onze spullen. Een bouwvakker zal zijn gereedschap dus zo goed mogelijk ordenen en de garderobemedewerker hangt de jassen op nummer.

Ook in ons huishouden proberen we onze spullen op een logische manier op te bergen zodat we alles snel terug kunnen vinden. We constateren dat iedereen er zo zijn eigen methode op nahoudt. De één vindt het volstrekt logisch om alle stropdassen bij elkaar in de kast te hangen, een ander kiest er voor om iedere stropdas op te hangen bij het pak waar deze het beste bij past. Wel bestaat er over het algemeen consensus over het feit dat de stropdassen bij de kleren horen.

Ook bij digitale informatie houdt iedereen er zijn eigen systeem op na om zijn bestanden te ordenen. Soms wordt een bepaalde ordening gestimuleerd door de software, maar binnen die kaders kiest iedereen zijn eigen logica. Dat is natuurlijk geen probleem zolang iedereen zijn eigen ordening maar snapt en min of meer consequent toepast. Zodra verschillende personen gebruik maken van hetzelfde informatiesysteem kunnen er problemen ontstaan. Wie bijvoorbeeld spullen via een webwinkel aanbiedt, wil natuurlijk dat gebruikers snel terechtkomen bij het product dat ze zoeken. Een zoekmachine kan daarvoor zorgen, maar in een aantal gevallen geven gebruikers er de voorkeur aan om zodanig door de producten te browsen dat ze met enkele muisklikken bij het gezochte zijn. De gebruiker heeft daarbij het voordeel dat hij geen zoekwoorden hoeft te bedenken en toch een set complete resultaten kan vinden. De bouwer van de site zal zich dus goed moeten verplaatsen in het zoekgedrag van de klant en zich niet te veel laten leiden door de processen of afdelingen van de eigen organisatie.

Voor informatieomgevingen als een intranet, netwerkschijf of documentmanagementsysteem geldt hetzelfde. Deze zijn opgezet om informatie met elkaar te delen. Dat betekent kortweg dat iedere gebruiker, of die nu informatie plaatst of zoekt, overweg moet kunnen met de gekozen systematiek. Er moeten dus afspraken komen die op de eerste plaats logisch zijn ten aanzien van het beoogde doel en op de tweede plaats voor iedereen te begrijpen zijn.

In een kort filmpje in de Dik Voormekaarshow van André van Duin (Duin, 2008) wordt een bibliothecaris opgevoerd die alle boeken op kleur gesorteerd in de kast heeft staan. Het ziet er leuk uit en iedereen begrijpt de systematiek, maar het dient het doel van een bibliotheek niet; tenzij de gebruikers een voorkeur hebben voor een bepaalde kleur boeken. De meeste bibliotheken bieden hun boeken aan op onderwerp of auteur zodat gebruikers snel hun weg vinden. Zo zullen de meeste mensen de boekenkast thuis ook hebben ingericht. We zien dan overigens wel dat ordening vaak een eenzijdige vindbaarheid van documenten oplevert, vooral in fysieke omgevingen. Ter aanvulling zijn er vaak instrumenten voor indirecte toegankelijkheid, zoals een catalogus, voor nodig om de materialen op verschillende wijzen terug te kunnen vinden.

In een digitale omgeving zijn we gewend om documenten te ordenen in mappen, waarbij het eenvoudig is om submappen aan te maken zodat er een soort boomstructuur ontstaat. Zo zal een map 'vakanties' al snel een onderverdeling op jaar krijgen. Hetzelfde zien we op websites waar het menu is opgedeeld in submenu's. In dat geval spreken we van een hiërarchie in de ordening. Begrippen die onder dezelfde noemer vallen, worden bij elkaar gezet onder een overkoepelende term. Voor deze hiërarchische ordening wordt vaak het begrip taxonomie gehanteerd. Ook informatietalen als classificaties en thesauri worden opgezet op basis

hiërarchie

van een hiërarchische onderverdeling van begrippen. We zagen al eerder dat het gebruik van categorieën eigen is aan de menselijke geest. Het onderverdelen van categorieën in een hiërarchie maakt daar vrijwel altijd deel van uit.

3.2 Betekenis van het begrip taxonomie

In de loop van de jaren is de term taxonomie in verschillende betekenissen gebruikt. Het is dan ook nuttig om deze verschillende betekenissen te bespreken voordat we ingaan op de regels die gelden bij het ontwikkelen van een goede ordening. Gemeenschappelijk aan alle betekenissen van een taxonomie is de toepassing van hiërarchie bij de ordening van begrippen, onderwerpen of goederen. Het onderscheid tussen de verschillende soorten taxonomieën is te maken naar de wijze waarop deze ordening plaatsvindt of naar het doel waarvoor de taxonomie wordt ingezet.

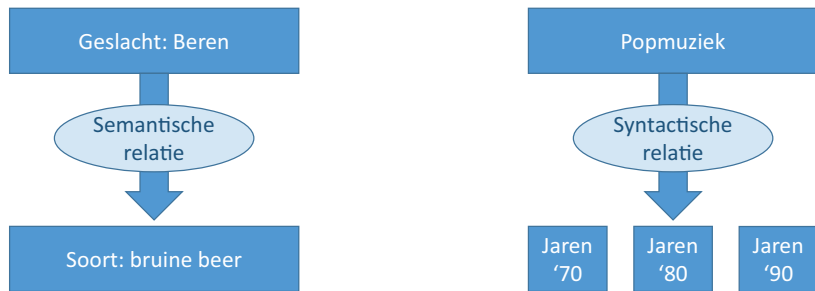
In de biologie is de taxonomie een klassiek systeem om het planten- en dierenrijk in kaart te brengen. Leidend daarbij is de onderverdeling vanaf het hoogste niveau 'rijk' (Engels: *Kingdom*), waaronder de dieren, mensen en planten vallen, tot aan het gedetailleerde niveau van 'geslachten' en 'soorten'. Zo behoort de bruine beer (*ursus arctos*) als 'soort' tot het 'geslacht' van de beren (*ursus*). Er bestaat dus een directe relatie tussen de overkoepelende term 'geslacht' (Engels: *genus*) en het onderliggende niveau 'soort' (Engels: *species*).

Een dergelijke onderverdeling zien we ook bij termen waarmee de inhoud van een document kan worden aangeduid. De bovenliggende en onderliggende termen hebben dan een relatie in betekenis. Zo kan het begrip 'onderwijs' worden verdeeld in 'lager', 'middelbaar' en 'hoger onderwijs'. Deze betekenisrelatie tussen termen wordt aangeduid met het begrip semantische relatie. In afbeelding 3.1 zie je dat weergegeven.

semantische
relatie

Naast de semantische (betekenis-)relatie kan er bij een hiërarchische onderverdeling ook sprake zijn van een syntactische relatie. Bij het ordenen van kleding in een webshop kan een onderverdeling gemaakt worden op kleur. Deze kleuren hebben geen betekenisrelatie met de kleding omdat ze net zo goed kunnen slaan op auto's of legoblokjes. We spreken hierbij van een syntactische relatie (Engels: *syntagmatic relation*) tussen de boven- en onderliggende begrippen. In de wereld van de informatievoorziening zijn er dus overeenkomsten met de biologische taxonomie (onderverdeling op basis van semantiek), maar zijn er ook verschillen aan te wijzen.

syntactische
relatie



Afb. 3.1 - Semantische en syntactische relaties.

Als we de taxonomie beschouwen tegen het licht van de toegankelijkheidsleer voor informatie, dan kunnen we de definitie aanhouden die Heather Hedden (2010) aangeeft in haar boek “The accidental taxonomist”:

“A hierarchically structured system of organizing names of concepts”

Zij laat dus buiten beschouwing of het om semantische of syntactische relaties gaat. Wij hanteren dat onderscheid wel: bij thesaurusbouw (hoofdstuk 4) is met name sprake van semantische relaties terwijl het bij autorisatielijsten (hoofdstuk 5) met name om syntactische relaties gaat.

Hiërarchie is dus het kenmerk van een taxonomie. In zijn boek over taxonomieën schrijft Evert Jagerman (Jagerman, 2006) dat een taxonomie een hiërarchische onderverdeling van begrippen is. Hij maakt daarbij onderscheid tussen de verschillende toepassingen van een taxonomie. Op de eerste plaats kan de onderverdeling van begrippen rubrieken opleveren waaronder documenten worden geplaatst. Dit is bijvoorbeeld het geval bij de mappenstructuur op een computer en de menustructuur op een website.

navigational taxonomy

Jagerman noemt dit de *navigational taxonomy*. Als deze taxonomieën worden ingezet om informatie binnen bedrijven te organiseren, bijvoorbeeld op een intranet of documentmanagementsysteem, dan sluiten ze vaak direct aan bij de afdelingen of bedrijfsprocessen. De begrippen die in een hiërarchische structuur worden ondergebracht en de rubrieken vormen, dekken als het goed is de lading van de documenten die in de betreffende rubrieken zijn ondergebracht. Het zijn de namen van mappen, rubrieken of internetpagina's.

descriptive taxonomy

Bij de tweede toepassing van de taxonomie gaat het om begrippen die als metadata aan een document kunnen worden toegekend. Degene die het document plaatst, kiest één of meer begrippen uit de lijst en de gebruiker kan het document vinden door op de metadata te zoeken via een zoekmachine of via de lijst. Deze lijst met begrippen die kunnen worden toegekend zijn vaak hiërarchisch geordend. Jagerman noemt deze lijst met begrippen de *descriptive taxonomy*. Dit kan een eenvoudige autorisatielijst

facetten

zijn, maar ook een uitgebreide informatietaal. In SharePoint is het bijvoorbeeld mogelijk om in de metadata store een *descriptive taxonomy* in te richten als informatietaal: het is dus niet de mappenstructuur, maar een verzameling termen die in een logische structuur zijn geplaatst. Het resultaat is dat een zoeker die een term aanklikt of intikt alle documenten ziet die de betreffende term toegekend hebben gekregen. Een groot voordeel ten opzichte van een mappenstructuur is, dat een document meer dan één term toegekend kan krijgen en dus via verschillende ingangen kan worden gevonden. Een document dat gaat over België, Nederland en Frankrijk, kan deze drie landen als term krijgen. Er kan zelfs gekozen worden om termen vanuit verschillende invalshoeken (facetten) toe te kennen. Naast het land kan dat bijvoorbeeld de tijdspanne zijn die in het document wordt besproken. Op deze manier kan het document vanuit verschillende invalshoeken worden teruggevonden, terwijl een mappenstructuur een indeling op één facet afdwingt.

We geven het voorbeeld van een ICT opleiding. Deze kan van zijn docenten vragen om al het lesmateriaal dat in SharePoint wordt geplaatst van termen te voorzien waarmee het zowel het vak als het documenttype wordt aangeduid. De taxonomie kan er dan als volgt uit zien:

- Vak

- Informatica
 - Programmeren basis
 - Modelleren
 - Java
 - ...
- IT alignment
 - Procesanalyse
 - Informatiesystemen
 - Business modellen
 - ...
- Contentbeheer
 - Basis toegankelijkheidsleer
 - Informatietalen
 - Zoektechnologie
 - ...

- Documenttype

- Presentatie
- Syllabus
- Opdracht
- ...

In bovenstaand voorbeeld kan een document een term toegekend krijgen uit de hiërarchie met vakken *en* uit de lijst met documenttypen. Zo kunnen bijvoorbeeld lesopdrachten over informatietalen via beide begrippen worden teruggevonden. Ook is het vaak mogelijk om begrippen met elkaar

te combineren. Zo kan een setje worden gecreëerd van alle presentaties voor een bepaald vak. In het volgende hoofdstuk zullen technieken worden besproken om termen bij een zoekactie te combineren of uit te sluiten.

De termen uit het voorbeeld zouden ook alfabetisch kunnen worden gepresenteerd, maar voor de docent die de juiste term zoekt om het document te typeren of een documentgroep te zoeken, is de hiërarchische indeling veel overzichtelijker. Zowel de *navigational taxonomy* als de *descriptive taxonomy* kan worden beschouwd als een systeem op zichzelf om informatie te ontsluiten. Beiden kunnen echter ook de basis vormen voor meer geavanceerde systemen als een classificatie of thesaurus.

boomstructuur

Voor een hiërarchische onderverdeling van begrippen wordt ook nog weleens de term boomstructuur gehanteerd, hoewel dat strikt genomen niet hetzelfde is: in een hiërarchie 'overerven' de onderliggende begrippen alle kenmerken van de bovenliggende term zoals dat in de biologische taxonomie ook het geval is. In de boomstructuur is de relatie iets vrijer: termen vallen weliswaar onder een categorie, maar hoeven niet alle kenmerken van de bovenliggende term te bezitten (Kwasnick, 1999). Eerder zagen we al de mogelijkheid van syntactische relaties tussen termen, maar ook bij semantische relaties kan de onderliggende term kenmerken hebben die afwijken van de bovenliggende term. Een voorbeeld daarvan is de onderverdeling van bijvoorbeeld een computer in de onderdelen scherm, toetsenbord en dergelijke. Er is weliswaar een semantische relatie, maar voor een scherm geldt niet dat het een soort computer is en voor een toetsenbord evenmin.

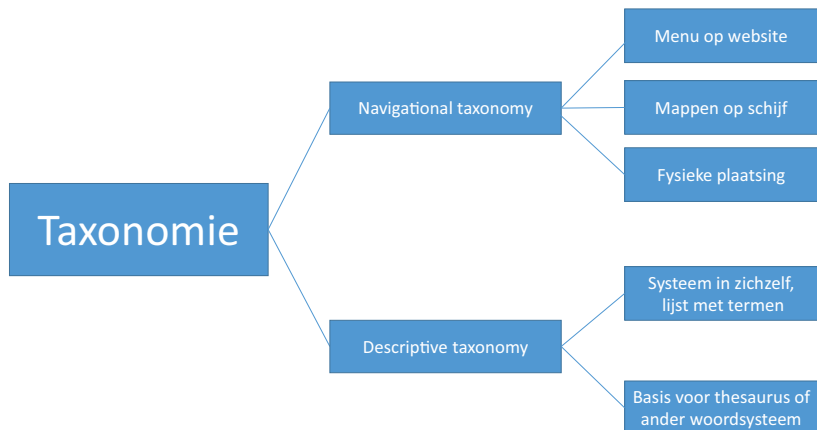
Of het nu om een strikte hiërarchie of een boomstructuur gaat, in beide gevallen geldt dat als deze structuur aan bepaalde voorwaarden voldoet en wordt uitgebreid met een codering voor de afzonderlijke rubrieken, er sprake is van een classificatie. Als de hiërarchie geavanceerd wordt vormgegeven met vaste verwijzingen, spreken we over een thesaurus.

taxonomie

We kunnen dus stellen dat een taxonomie een hiërarchische onderverdeling van begrippen is die tot rubrieken leidt waar materiaal in kan worden geplaatst of die als termen kunnen worden gebruikt om aan documenten toe te kennen. In het Angelsakisch taalgebied is de term inmiddels een containerbegrip geworden voor alle typen informatietalen of ordeningssystemen, zelfs als daar geen hiërarchie aan te pas komt. De eerdergenoemde Heather Hedden geeft in haar boek dan ook een tweede definitie voor het begrip taxonomie waarmee ze deze alomvattendheid uitdrukt:

"Any knowledge organization system, whether hierarchical or not, involving controlled names of concepts".

Leveranciers van software waarmee systemen voor toegankelijkheid kunnen worden gebouwd houden er ook vaak een brede opvatting van het begrip taxonomie op na. Zij spreken dan ook over “*taxonomy management tools*” als containerterm voor applicaties waarmee verschillende soorten informatietalen kunnen worden beheerd. In dit boek houden we vast aan het hiërarchische aspect als kenmerk van de taxonomie, maar het is goed om te weten dat de term met name in buitenlandse literatuur en in andere contexten als breder concept kan worden gehanteerd.



Afb. 3.2 - Verschillende varianten van de taxonomie.

3.3 Werkwijze en regels

Of we nu een *navigational taxonomy* of een *descriptive taxonomy* ontwerpen, het is altijd belangrijk dat het resultaat eenduidig, compleet en goed werkbaar is. Een mappenstructuur waarbij het niet duidelijk is waar je een document kwijt kunt of waar je het kunt vinden, schiet zijn doel voorbij en een lijst met termen die niet compleet is of waar termen verschillend kunnen worden geïnterpreteerd, is al snel onwerkbaar. Om tot een goede taxonomie te komen, is het belangrijk dat in ieder geval goed wordt gekeken naar de collectie met documenten en dat de gebruikers worden betrokken bij het bepalen en onderverdelen van de begrippen. Hieronder beschrijven we een stappenplan dat voorziet in de belangrijkste aandachtspunten bij het creëren van een werkbare taxonomie. Deze werkwijze is mede gebaseerd op het schema “How to build a taxonomy” van Straits Knowledge (2012) en voert gedeeltelijk terug op oude principes uit de classificatietheorie. Een aantal principes uit de methodieken van bibliotheken om boeken op onderwerp bij elkaar te plaatsen, kunnen goed worden ingezet bij het ordenen van digitale informatie.

3.3.1 Stap 1: Doel bepalen

Bij veel activiteiten is het zinnig om het doel te bepalen voordat je van start gaat. Dat geldt ook voor het organiseren van je informatie. Met welk doel wil je een ordening of informatietaal inzetten? Daaruit volgt namelijk de keuze voor de juiste informatietaal. Als we boeken gaan ordenen hebben we bijvoorbeeld de keuze uit een ordening op onderwerp waarbij een numerieke classificatie van waarde is, of een ordening op alfabet waarbij dat niet nodig is. In een SharePointomgeving kan een mappenstructuur worden aangemaakt, maar het gebruik van mappen kent een belangrijk nadeel; de documenten in een mappenstructuur zijn slechts via één weg terug te vinden. Wie op het hoogste niveau het verkeerde pad kiest, komt nooit bij het juiste document. Bij een mappenstructuur is het daarom belangrijk dat er sprake is van een strakke, logische indeling volgens bepaalde afspraken. In hoofdstuk 6 gaan we in op deze regels die combinatievolgorde worden genoemd. Als documenten via verschillende wegen teruggevonden moeten worden is een *descriptive taxonomy* logischer. Vervolgens bepaalt het doel hoe de betreffende ordening eruit gaat zien. De eigenaar van een supermarkt heeft zich bij de ordening van zijn producten ten doel gesteld dat de consumenten op zoveel mogelijk artikelen worden geattendeerd, omdat dat het koopgedrag kan beïnvloeden. Daarmee worden concessies gedaan aan een ander doel, namelijk consumenten zo snel mogelijk naar het gezochte artikel brengen. Als de snelheid het doel zou zijn, dan waren supermarkten anders ingericht. Winkelketens experimenteren regelmatig om te achterhalen wat de juiste ordening voor hun doel is. In een interview in de Volkskrant van 4 december 2015 geeft de directeur van een warenhuisketen bijvoorbeeld aan dat de badlakens in de Nederlandse vestigingen bij het linnengoed liggen, terwijl deze in Frankrijk als experiment bij de badkamerartikelen zijn geplaatst. Het lijkt erop dat deze keuze in Frankrijk leidt tot betere verkopen.

Het doel van een website is meestal wel vast te stellen, omdat de meeste websites worden opgezet met een duidelijk doelstelling. Lastig is wel dat het doel of de doelgroep meerledig kan zijn. Wie vormt bijvoorbeeld de doelgroep van de website van een basisschool? Zijn het de ouders van de leerlingen of zijn het de leerlingen zelf?

Een intranet heeft bij de meeste bedrijven de functie om medewerkers te informeren. De laatste jaren is het intranet echter interactiever geworden in de zin dat het ook gebruikt wordt om kennis en documenten te delen die door de medewerkers zelf worden geplaatst. De consequentie hiervan is dat de ordening niet alleen bepaald wordt door een redactie die publiceert, maar dat de logica ook wordt bepaald vanuit de medewerkers van de organisatie zelf.

Het doel van een website is dus van invloed op de gekozen ordening. Zaken die voor het doel van belang zijn, zullen een prominente plek moeten krijgen en de gehanteerde terminologie moet aansluiten bij de

doelgroep. Een specialistische site met medische informatie voor artsen kent andere begrippen en een andere ordening dan een site die is gericht op patiënten.

3.3.2 Stap 2: Begrippen verzamelen: analyse van de content en de processen

Bij het verzamelen van begrippen baseer je je op twee belangrijke bronnen: de informatie die je aanbiedt en de begrippen die door de gebruikers worden gehanteerd in de dagelijkse praktijk. Bij het analyseren van de informatie kun je handmatig te werk gaan, maar ook software gebruiken. De eenvoudigste methode is een *wordcloud* waarbij veelgebruikte woorden groot worden weergegeven (zie bijvoorbeeld tagcrowd.com). Er bestaat ook geavanceerdere software die op basis van een documentverzameling een voorstel voor een taxonomie kan doen. In hoofdstuk 10 wordt deze software nader besproken. Het is daarnaast erg belangrijk dat de gebruikers worden betrokken bij de taxonomie die je wilt ontwikkelen. Zij zijn niet alleen de toekomstige zoekers in je systeem, maar in veel gevallen ook diegenen die documenten gaan plaatsen. Om te zorgen dat documenten goed vindbaar worden geplaatst, moeten de gebruikers zich goed herkennen in de gekozen termen. Een effectieve methode om de terminologie van de doelgroep in kaart te brengen, is het organiseren van een workshop. In deze workshop kun je enkele grote thema's noteren op grote vellen waarna gebruikers de gelegenheid krijgen om gerelateerde begrippen rondom deze thema's noteren. Op deze manier ontstaan *maps* waarbij termen rondom een thema zijn weergegeven. Bij meertalige instellingen is het overigens aan te raden om direct met de verschillende talen te werken. Het achteraf vertalen van een taxonomie die "af" is, leidt vaak tot interpretatieverschillen. In de praktijk blijkt dat een combinatie van deze werkwijzen het meest aan te bevelen is. Na het verzamelen van de begrippen zul je moeten kiezen welke begrippen je daadwerkelijk gaat gebruiken. Zo zul je in de verzameling vaak synoniemen tegenkomen. Dan is het in de eerste plaats belangrijk om te bepalen of er werkelijk sprake is van synoniemen of dat begrippen qua betekenis dicht bij elkaar liggen, de zogenoemde quasi-synoniemen. Bij echte synoniemen maak je een keuze voor een voorkeursterm. Als je informatiesysteem de mogelijkheid biedt om door te verwijzen, neem je ook de niet-voorkeursterm op en verwijs je van daaruit door. Bij deze beslissingen houd je het doel en de doelgroep van je site zo goed mogelijk voor ogen. Op de site van de jeugdprogrammazer Zapp (www.zapp.nl) is bijvoorbeeld gekozen voor de term 'programma's', terwijl de website van de NOS (www.nos.nl) spreekt over 'uitzendingen'. Bij quasi-synoniemen zul je moeten bepalen of je ze als synoniemen behandelt en een voorkeursterm kiest, of dat je beide begrippen naast elkaar hanteert omdat er nu eenmaal een nuanceverschil in betekenis is. Zo kun je de begrippen 'reader' en 'syllabus' beschouwen als synoniemen, maar in een specifieke documentcollectie kunnen ze naast elkaar bestaan omdat er onderscheid is in herkomst (verzamelde fragmenten van anderen of een door de docent zelf geschreven document).

3.3.3 Stap 3: Facetten en groepen

Door de verzamelde begrippen te bestuderen, kan een aantal facetten worden herkend waarop de documenten kunnen worden ingedeeld. Zo kunnen documenten bijvoorbeeld worden ingedeeld op het thema dat ze behandelen, maar bijvoorbeeld ook op de doelgroep waarop ze gericht zijn, het proces waar ze bij horen, de klant die het betreft of hun redactionele vorm zoals bijvoorbeeld offerte, notulen of verslag. Bij een documentcollectie is het raadzaam om het aantal facetten binnen de perken te houden. Voor ieder facet moet immers een term door de gebruiker worden ingevuld (voor zover dat niet automatisch kan). Bij producten kan de aanbieder van een website heel ver gaan. Zo kunnen vakantiehuizen metadata van veel verschillende facetten toegekend krijgen die allemaal als filter kunnen dienen. Wie een vakantiehuis zoekt, kan zijn vakantiebestemming niet alleen selecteren op basis van locatie en aantal personen, maar bijvoorbeeld ook op de aanwezigheid van wifi, zwembad, afstand tot het strand en de supermarkt, enzovoort.

Na het bepalen van de facetten kunnen de verschillende termen bij deze facetten worden ondergebracht. Dit proces noemen we *cardsorting* (zie kader). Tijdens het *cardsorten* zal blijken dat de benoemde facetten wellicht aanpassing behoeven. Uiteindelijk is er dan ook sprake van een iteratief proces waarvan het resultaat vervolgens wordt getest. Uiteraard kunnen uit de test ook zaken naar voren komen die tot aanpassingen leiden.

In een webomgeving kunnen de producten of pagina's, in tegenstelling tot een mappenstructuur, volgens verschillende facetten worden gegroepeerd. Het gevolg is dat pagina's ook via verschillende paden kunnen worden benaderd. Zo kunnen op de pagina van een kledingwinkel de verzameling jassen voor dames gevonden worden via >dameskleding>jassen, maar ook via >jassen>dames. Die verschillende invalshoeken moeten dan wel bij dezelfde informatie uitkomen, anders ontstaan er twee plekken waar (ongeveer) dezelfde informatie staat (en zijn er dus twee plekken waar dezelfde informatie moet worden bijgehouden). Een goed content-managementsysteem voorziet in deze mogelijkheden en zorgt ervoor dat de informatie slechts op één plek staat.

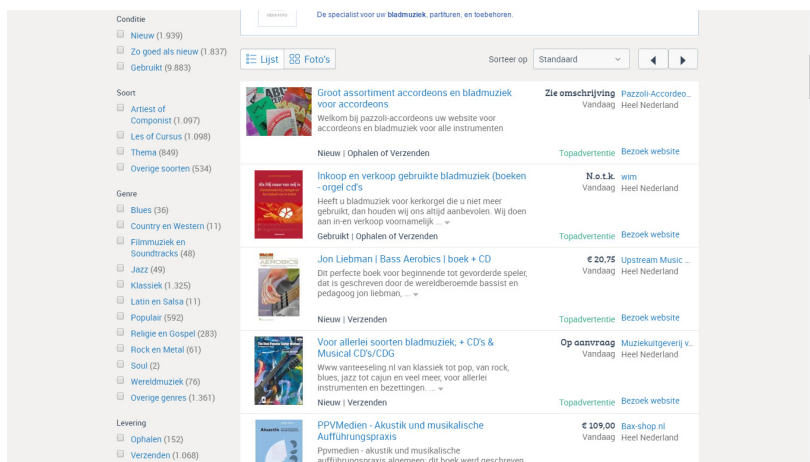
Eenheid van verdelingskarakteristiek

Welke wegen je binnen een website ook aanbiedt naar de pagina's, de menu's moeten logisch en ondubbelzinnig zijn voor de gebruiker. Een website waarbij de gebruiker in hetzelfde dropdownmenu de keuze krijgt uit: 'dameskleding' en 'buitenkleding' schept verwarring omdat de gebruiker niet weet wat hij moet kiezen als hij op zoek is naar een damesjas. In het beste geval leiden ze naar dezelfde content, maar dat weet de gebruiker op dat moment niet. In dit voorbeeld hebben we te maken met een belangrijke regel binnen de classificatietheorie, de eenheid van verdelingskarakteristiek. Hieronder verstaan we dat de onderverdeling van een klasse moet zijn gebaseerd op één en hetzelfde facet. Er moeten

eenheid van
verdelings-
karakteristiek

dan ook niet twee of meer verdelingskarakteristieken op één niveau worden toegepast. Kleding onderverdelen in dames-, heren-, tennis- en voetbalkleding betekent een schending van de regel, omdat er sprake is van twee facetten: kleding naar sekse en naar functie (sport). De eenheid van verdelingskarakteristiek blijft overeind wanneer kleding onderverdeeld wordt naar 'herenkleding' en 'dameskleding', of naar 'tenniskleding' en 'voetbalkleding'. Ook kunnen we voorwerpen onderverdelen naar onderdelen: fietsen naar fietswielen, fietssturen, fietskettingen enzovoort.

De rubriek bladmuziek op Marktplaats kent verschillende facetten (afbeelding 3.3). Het facet 'genre' is netjes opgedeeld in verschillende muziekgenres en daarmee eenduidig. Bij het facet 'soort' ontstaat echter verwarring. Het ontbreken van de eenheid van verdelingskarakteristiek blijkt onder andere uit het feit dat de onderdelen elkaar niet uitsluiten. Verschillende componisten hebben muziek met een bepaald thema geschreven, zoals Kerst, huwelijk e.d., en een aantal heeft ook zeer gerichte etudes voor muzieklessen gecomponeerd. De facetten sluiten elkaar dus niet uit en de zoeker raakt in verwarring: moet hij voor het Weihnachtsoratorium van Bach nu zoeken bij 'componisten' of bij het thema 'Kerst'?



Afb. 3.3 - Muziekrubriek op Marktplaats.

Co-extensie en modulatie/gradatie

Zodra je binnen een homogene groep een bepaalde hiërarchie zichtbaar kunt maken, is het raadzaam ook te letten op de regels ten aanzien van de co-extensie en de modulatie/gradatie.

co-extensie

De officiële regel voor co-extensie luidt: de gezamenlijke omvang (of co-extensie) van het totaal van de subklassen moet even groot zijn als de overkoepelende klasse. Deze regel houdt in dat alle informatie over alle

aspecten van het overkoepelende onderwerp een plaats moeten kunnen krijgen. Als binnen de klasse 'sportkleding' alleen de subklassen 'tennis-' en 'voetbalkleding' bestaan, waar vind je dan de atletiekoutfit? Dit is op te lossen door de klasse 'atletiekkleding' in te voeren, of, als oplossing die altijd bruikbaar is, de klasse 'overige sportkleding'. Een rubriek of map 'overig' kan dus een goede oplossing zijn. Het wordt echter problematisch als deze rubriek vervolgens te vol wordt met allerlei onderwerpen. In dat geval is het zaak om van vaak voorkomende onderwerpen een eigen rubriek te maken.

modulatie/ gradatie

Met modulatie/gradatie bedoelen we dat er moet worden gezorgd voor een geleidelijke overgang bij het onderverdelen van een klasse. De bouwer slaat dan geen niveau over, maar laat de tussenstappen in de ordening zien. Hij plaatst de categorie 'tennissokken' niet meteen onder de klasse 'kleding', maar maakt de indeling kleding > sportkleding > tenniskleding > tennissokken.

Websitebouwers moeten een balans zien te vinden tussen de regel van modulatie/gradatie enerzijds en een beperkt aantal klikken anderzijds. Lang was de vuistregel dat een gebruiker niet meer dan drie of vier keer zou moeten klikken om bij de gewenste pagina te komen. Inmiddels wordt deze vuistregel genuanceerd. Belangrijker dan het aantal kliks is de helderheid van de structuur, zo blijkt uit onderzoek (zie bijvoorbeeld

The screenshot shows the Intratuin website's 'DIEREN' (Animals) section. The navigation bar includes links like 'Voor het beste groenadvies', 'Alles voor huis, tuin en dier', and '51 winkels in Nederland'. The main menu lists categories: KAMERPLANTEN, TUINPLANTEN, BINNEN- EN BUITENPOTTEN, TUIN & TERRAS, TUINMEUBELEN, BARBECUES, WONEN, DIEREN, and MERKEN. The 'DIEREN' section is expanded, showing a grid of product categories:

- DIEREN IN JE TUIN**
 - Buitenvogels
 - Eekhoorns
 - Egels
 - Bijen
 - Insecten
 - Vlinders
- DIERENBENODIGDHEDEN**
 - Dieren speelgoed
 - Dierenverzorging
 - Kooi accessoires
 - Kooien, hokken en benches
 - Kussens, manden en dekens
 - Riemen en halsbanden
 - Voer en snacks
 - Voer- en drinkbakken
- HONDEN**
 - Benchs en Hokken
 - Honden verzorging
 - Hondenpoepscheppen en -zakjes
 - Hondenspeelgoed
 - Hondentassen en reisartikelen
 - Hondenvoer en -snacks
 - Kussens, manden en dekens
 - Riemen en halsbanden
 - Trainingshulpen
 - Voer- en drinkbakken
 - Medicijnen voor honden
- KATTEN**
 - Katten draagtas en reisartikelen
 - Katten verzorging
 - Kattenbak en accessoires
 - Kattenluiken
 - Kattenspeeltjes
 - Kattenvoer en -snacks
 - Krabpalen
 - Kussens, manden en dekens
 - Riemen en halsbanden
 - Voer- en drinkbakken
 - Medicijnen voor katten
- KNAAGDIEREN**
 - Cavia benodigdheden
 - Fretten benodigdheden
 - Hamster benodigdheden
 - Knaagdieren speelgoed
 - Knaagdierenkooi toebehoren
 - Knaagdierenkooien
 - Knaagdierenvoer en -snacks
 - Knaagdieren manden
 - Knaagdieren verzorging
 - Muizen benodigdheden
- KONIJNEN**
 - Konijnen transportbox en reismanden
 - Konijnen verzorging
 - Konijnenhok toebehoren
 - Konijnenhokken
 - Konijnenvoer
- VOGEL**
 - Vogelkooi toebehoren
 - Vogelkooien en volières
 - Vogel speelgoed
 - Vogelvoer

At the bottom, there are promotional banners for 'KOM JE SPEUREN in de winbol', 'SPEUR TUIN', 'selective+', and a cat product.

Afb. 3.4 - Webshop van Intratuin.

Frankwatching, 2014). Bezoekers willen vooral niet het gevoel hebben verdwaald te zijn. Deze bevinding wijst dus ook op het nut van een doordachte ordening. Voor de omvang van een homogene groep, dus het aantal items of onderdelen op hetzelfde niveau waaruit een gebruiker kan kiezen, is er geen duidelijk maximum. Zolang de verdelingskarakteristiek uniform is en de volgorde van de items logisch, kunnen gebruikers hun weg in een lijst met items wel vinden. In onderstaande afbeelding is de webshop van Intratuin weergegeven. In het openklapmenu worden de subrubrieken op alle niveaus direct weergegeven. Dat maakt het voor de gebruiker overzichtelijk. Bovendien zijn er niet meer dan drie niveaus. Ook het aantal items op hetzelfde niveau blijft met maximaal tien binnen de perken. Over de eenheid van de verdelingskarakteristiek valt nog wel wat op te merken: 'konijnen' zou natuurlijk eigenlijk onder de 'knaagdieren' moeten worden geplaatst.

Twee voorbeelden

Bij een aantal vakantiesites is goed te zien hoe lastig het is om informatie goed in groepen op te delen. Zo komen we bij de reisorganisatie TUI onder

The screenshot displays the 'Muziek' (Music) section of the ECI website, which has 144,659 titles. The interface includes a sidebar on the left for filtering by 'Genres' and 'Kenmerk' (Feature).

Genres:

- Pop (62385)
- Klassiek cd's (2694)
- Jazz, Blues, Country & Folk (2181)
- Rock & Metal (17436)
- Dance, House, Hip Hop & Rap (6763)
- Amusement cd's (4822)
- Wereldmuziek (4484)
- Verzameld cd's (2862)
- R&B & Soul cd's (2304)
- Musical, Soundtrack & Cabaret cd's (588)
- Muziekaccessoires (140)

Kenmerk:

- Verschenen laatste 30 dagen (1000)
- Verschijnt komende 30 dagen (562)
- Meer Kenmerk ▾
- Talen ▾
- Engels (36194)
- Duits (1549)
- Meer Talen ▾

Muziek (144659 titels)

Top deals
 20% EXTRA KORTING
 gebruik actiescode **topdeal**

Best verkocht

Rank	Title	Artist	Format	Price
1.	Qmusic Top 500 Van H...	CD -	€ 18,99	
2.	25	CD - Adele	€ 17,99	
3.	538 Hitzone 77	CD -	€ 17,99	
4.	100X Zomer 2016	CD -	€ 17,99	
5.	Onderweg Naar Jou (...)	CD - Jannes	€ 17,99	
6.	...	CD	€ 17,99	

Alvast te reserveren

The 'Alvast te reserveren' section shows a grid of album covers, including 'HIT ZONE 78', 'ROYAL NEW TOWN', 'henk bernard', 'ROY ARISON COLLECTED', and 'Django Wagner'.

Afb. 3.5 - Facetten op de website van ECI.

andere de categorieën ‘vliegereizen’ en ‘last minutes’ tegen. Zijn de meeste last-minutes dan geen vliegereizen en hoe weet ik dat ik niets mis? Zolang klikken op elk van die kenmerken dezelfde informatie oplevert, is er weinig aan de hand, maar dat is vaak niet het geval.

De grote aanbieders van boeken, cd's etc. sleutelen regelmatig aan de structuur van hun site, omdat men in de gaten heeft dat het lastig is om producten voor de consument goed vindbaar te presenteren. Bij de online boekhandel ECI (www.eci.nl, afbeelding 3.5) was de rubriek ‘muziek’ in het verleden onderverdeeld in onder andere de rubrieken ‘Algemeen’, ‘Kerst’, ‘Jeugd’ en ‘Nederlandstalig’. Het was daardoor lastig om bijvoorbeeld een Nederlandstalige kerst-cd te vinden. De site heeft dat nu mooi opgelost door de taal als een apart facet te presenteren. Onder het facet genre staan keurig de verschillende muziekstijlen bij elkaar en onder het facet ‘talen’ staan ook alleen maar talen. Je zou verwachten dat je deze facetten kunt combineren om bijvoorbeeld een Duitse Jazz cd te vinden. Dat lukt echter niet, de software lijkt een combinatie van facetten jammer genoeg niet aan te kunnen.

3.3.4 Stap 4: Collocatievolgorde

De werkzaamheden uit stap drie leiden als het goed is tot een heldere indeling in facetten of rubrieken die ieder een aantal items of subrubrieken bevatten. In de vierde stap wordt bepaald in welke volgorde deze items worden weergegeven. We noemen dit de collocatievolgorde. Collocatie betekent letterlijk “bij elkaar zetten” en is een mooie term voor iets wat vaak vanzelfsprekend is. Als het goed is, vindt een gebruiker in de opsomming al snel wat hij zoekt. Bij een rubriek met historische onderwerpen verwachten we een indeling in tijdperken van vroeger naar het heden en bij het facet ‘prijs’ is het logisch dat de bedragen oplopen. Toch zijn er ook wel situaties waarbij wat langer moet worden nagedacht over de volgorde. Bij webshops wordt er vaak voor gekozen om binnen een rubriek de meest verkochte producten bovenaan te plaatsen. We noemen dat een pragmatische volgorde. Daarmee wordt een grote groep gebruikers prima bediend, maar wie op zoek is naar een minder populair product moet de hele lijst aflopen. Met name bij rubrieken met meer dan 10 items kan het dan logischer zijn om een andere systematiek aan te houden. Natuurlijk kan er ook voor gekozen worden om de items in alfabetische volgorde te plaatsen. Dat kan goed werken, al is er dan van echte collocatie (bij elkaar zetten) geen sprake meer. In bovenstaand voorbeeld van de webshop van Intratuin is het gissen naar de collocatievolgorde. Zo bevat de rubriek ‘honden’ meer dan 10 items die op het eerste gezicht alfabetisch worden gepresenteerd. Het onderdeel ‘medicijnen voor honden’ staat dan echter verkeerd. Het lijkt erop dat er in principe is gekozen voor een alfabetische volgorde en dat de ‘medicijnen’ later zijn toegevoegd.

In documentsystemen zal de software vaak een alfabetische volgorde aanhouden. Als deze niet wenselijk is, kan de volgorde worden beïnvloed door bijvoorbeeld nummers of letters voor de naam van het item of rubriek te plaatsen.

dynamische inhoud

Bij de genoemde voorbeelden van websites is meestal sprake van een vaste websitestructuur en een dynamische inhoud: de websitebouwer bepaalt de structuur, bepaalt de onderwerpen en maakt de pagina's, maar de inhoud van de pagina's is dynamisch. De informatie wordt als het ware aan een pagina gehangen door de informatie verschillende termen van verschillende facetten te geven. Deze werkwijze wordt vooral toegepast bij winkelketens en handelssites. Als men het over taxonomieën heeft, dan wordt in eerste instantie meestal de achterliggende indelingsstructuur van deze sites met dynamische inhoud bedoeld: de structuur wordt vastgelegd in een taxonomie (categorieën); de producten (of informatie-items) worden voorzien van een categorie aanduiding. Voor de gebruiker werkt het als een *navigational taxonomy*: hij browsert door de verschillende menu's en submenu's. Om dit mogelijk te maken, is er op de achtergrond eigenlijk een *descriptive taxonomy* toegepast omdat de gebruiker via verschillende invalshoeken bij dezelfde content kan komen. Dit wordt duidelijker, wanneer de gebruiker ook daadwerkelijk kan filteren op de verschillende facetten, zoals in het genoemde voorbeeld van vakantiehuizen en de site van de ECI.

vaste inhoud

Niet alle websites die gestructureerd informatie willen aanbieden, hebben dit dynamische karakter. Soms hebben ze nog gewoon paginagebonden inhoud. De websitebouwer bepaalt dan de structuur en de onderwerpen, maakt de pagina's en voorziet die van inhoud. In beide gevallen (vaste of dynamische inhoud) moet de websitebouwer letten op de regels van eenheid van verdelingskarakteristiek, co-extensie en modulatie. Bij websites met vaste inhoud zal er weinig behoefte bestaan om verder te voldoen aan de regel van co-extensie, door bijvoorbeeld de rubriek 'overige' te introduceren. Alle belangrijke items zijn immers tevoren bedacht en opgenomen in de website.

3.4 Samenvattend

In dit hoofdstuk zijn we uitgebreid ingegaan op het aspect van de ordening en het fenomeen taxonomie dat daarbij hoort. Omdat we grote hoeveelheden informatie graag onderverdelen, is bij een taxonomie altijd sprake van een hiërarchische onderverdeling van begrippen, al of niet volgens strenge regels. We zagen daarbij het onderscheid tussen de *navigational taxonomy* en de *descriptive taxonomy*. In het verleden maakten informatiespecialisten als archivariissen of bibliothecarissen nogal eens indelingen die volgens de strakke regels misschien wel klopten, maar waar de gebruiker niet goed mee uit de voeten kon. Een goede

taxonomie kan dan ook alleen worden gemaakt door een vast stappenplan te gebruiken en de gebruikers bij het proces te betrekken. Aanbieders van webwinkels kijken goed naar het zoekgedrag van de gebruiker en passen hun sites daar op aan. Wie een informatiesysteem inricht waar gebruikers zelf content kunnen plaatsen, kan bij de bouw van de taxonomie al niet om de toekomstige gebruikers heen.

Gebruikte literatuur

Duin, A. van (2008). *Vreemde bibliotheek met Ron Brandsteder*.

Geraadpleegd op: 18 maart 2010, van:

<http://www.youtube.com/watch?v=sg2KRGkNWA>

Frankwatching, 2014. *5 hardnekkige UX-mythes: waarom moeten we er vanaf?*. Opgevraagd 20 april 2016 van:

<http://www.frankwatching.com/archive/2014/01/13/5-hardnekkige-ux-mythes-waarom-moeten-we-er-vanaf/>

Hedden, H. (2010). *The accidental taxonomist*. Information today. ISBN 978-1-57387-397-0

Jagerman, E. (2006). *Creating, maintaining and applying quality taxonomies*. Zoetermeer: E.J. Jagerman. ISBN 978-90-811274-1-7

Kwasnick, Barbara H. (1999). *The role of classification in knowledge representation and discovery*. Library Trends, 48 (1), 22-47

Straits Knowledge (2012). *How to build a taxonomy*. Opgevraagd 20 april 2016 van: <http://www.greenchameleon.com/gc/guides/>

Verder lezen

Magrijn, H; Linden, M. van der (2008). De actualiteit van 19e-eeuwse classificatietheorie. In Trier, G. van (red.), *Handboek Informatiewetenschap voor bibliotheek en archief*. (Aanvulling 46, IV F 800, p. 1-34). Alphen aan den Rijn: Kluwer

KADER

Cardsorting

De navigatiestructuur van een website bepaalt in hoge mate de vindbaarheid van de informatie. Staat de content onder de categorie waar je die verwacht? En is het label wel helder? Wie een navigatiestructuur ontwerpt, gaat vaak te rade bij de gebruikers van de informatie. Want als je weet waarnaar en op welke manier de gebruiker zoekt, is het mogelijk om een navigatiestructuur te ontwerpen die daar goed bij aansluit. Dit soort gebruikersonderzoek of *User Experience Design* kan gedaan worden door middel van *cardsorting*.

Cardsorting is een methodiek waarbij onderwerpen door respondenten gegroepeerd worden om zo inzicht te krijgen in wat voor de gebruiker een logische hiërarchie is. Er zijn twee soorten *cardsorting* te onderscheiden. De eerste heet open *cardsorting* en daarbij kunnen de categorieën zelf aangemaakt worden. Bij gesloten *cardsorting* worden de items geplaatst onder “al gegeven” categorieën.

Voordelen van kaartjes	Voordelen van software
Laagdrempelig: kaartjes zijn altijd voorhanden en goedkoop.	Er hoeft geen tijd te worden besteed aan het schrijven van kaartjes
Overleg tussen deelnemers is gemakkelijk te organiseren	Meer onafhankelijk doordat mensen individueel invullen
Je leert veel van het (hardop) denken van de deelnemers tijdens het sorteren	In korte tijd kun je veel sessies afnemen zonder dat je daar zelf bij hoeft te zijn
Grote flexibiliteit tijdens de sessie, want kaartjes en labels kunnen op meerdere plekken worden neergelegd of bijgemaakt worden	Online analysetools beschikbaar
...	...

Het proces verloopt in een aantal stappen:

Stap 1:

Maak een overzicht van termen. Gebruik termen waarvan je weet dat ze gebruikt zullen gaan worden. Schrijf ze op kaartjes of type ze in het

programma. Bij de gesloten variant maak je op dit moment ook de categorieën aan.

Stap 2:

Nodig mensen uit om de items te sorteren. Dat sorteren kan digitaal gebeuren of tijdens een bijeenkomst waarbij kaartjes worden gebruikt.

Beide keuzes hebben hun voor-en nadelen, die hierboven staan weergegeven.

Als voorbeeld voor software hebben we afbeeldingen genomen van het programma OptimalSort (www.optimalsort.com), een programma waarmee gratis een beperkt aantal deelnemers items kan sorteren. Deze software genereert drie grafieken: het standardisation grid, de similarity matrix en het dendrogram.

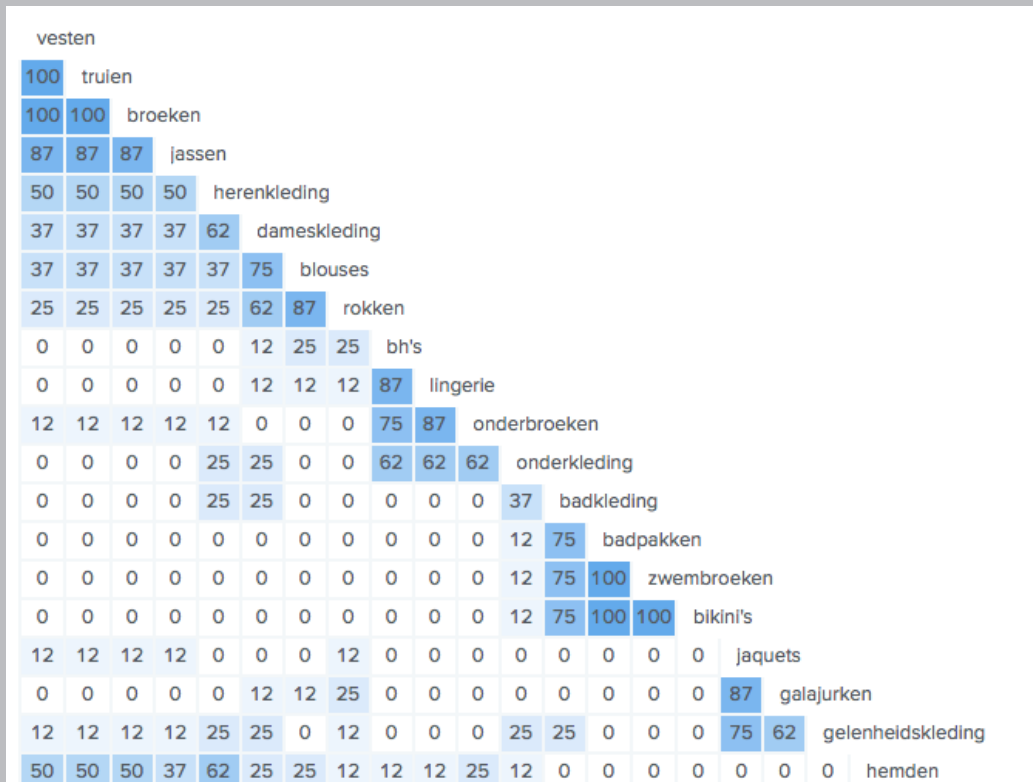
	Badkleding	Bovenkleding	Gelegenheidskle...	Kleding	onderkleding
badkleding	5			2	
badpakken	7				
bh's				2	5
bikini's	7				
blouses				7	
broeken		2		4	
dameskleding				7	
galajurken			6	1	
gelegenheidskleding			4	2	
hemden		1		5	1
herenkleding				7	
jaquets			6		

Afb. 2 - Een gedeelte van het standardisation grid.

In het standardisation grid staat hoe vaak een item in een bepaalde categorie gesorteerd wordt. In afbeelding 2 wordt 'badpakken' altijd (7x) onder de 'badkleding' gezet. Het item 'hemden' blijkt lastiger om te plaatsen: dat blijkt in drie categorieën terecht te komen.

In de similarity matrix wordt van paren van termen bekeken hoe dicht ze bij elkaar in de buurt komen. In afbeelding 3 is zichtbaar dat 'truien' en 'vesten'

worden steeds in elkaars buurt geplaatst (een score van 100%), terwijl de eerder genoemde 'hemden' ook hier veel variëteit vertonen: het vaakst komen ze bij de 'herenkleding', maar ook bij de 'broeken' en 'truien' worden ze in de buurt geplaatst.



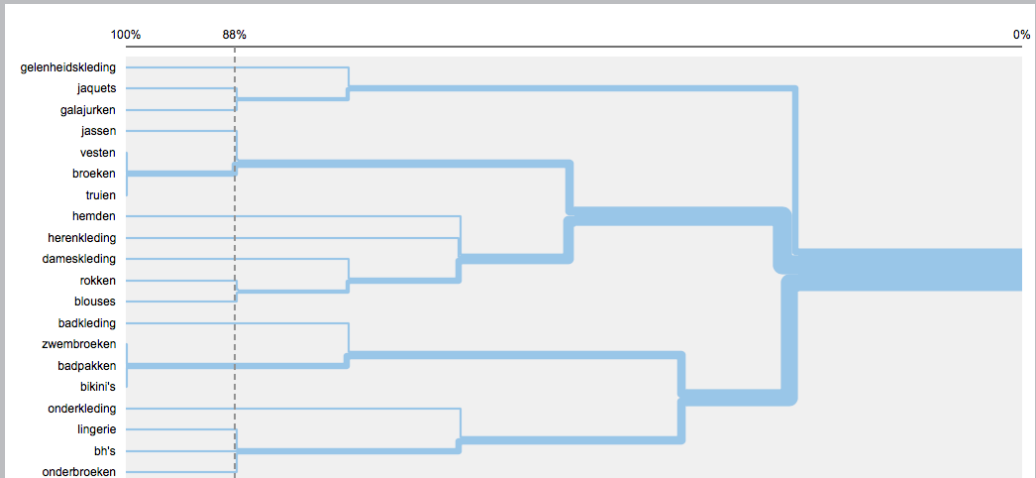
Afb. 3 - Similarity matrix

Uit het dendrogram tenslotte (zie afbeelding 4) valt op te maken hoeveel procent van de deelnemers bepaalde items groepeerde. In onderstaande uitsnede valt bijvoorbeeld te zien dat 88% van de deelnemers de items 'truien', 'broeken', 'vesten' en 'jassen' bij elkaar in de buurt plaatst. In het algemeen geldt in het dendrogram: hoe dikker de lijn, hoe meer deelnemers het met elkaar eens zijn.

Alle drie deze hulpmiddelen dragen bij tot het resultaat van *cardsorting*. Een *cardsorting* sessie of een aantal online sessies levert nog geen uitgewerkte structuur op, maar geeft wel inzicht in de volgende punten:

- Of de termen op zichzelf duidelijk zijn;
- Hoe de intuïtieve ordening van de informatie is;
- Hoe breed en diep de structuur wordt ;

- Welke categorieën er worden onderscheiden;
- Of verschillende doelgroepen op dezelfde indeling en/of categorieën uitkomen.



Afb. 4 - Dendrogram

Uiteindelijk kun je daarmee komen tot het resultaat van *cardsorting*: een boomstructuur van onderwerpen en subonderwerpen.

Hoofdstuk 4

Van keuzelijst tot thesaurus

In hoofdstuk 3 is aandacht besteed aan structurering, een vorm van directe toegankelijkheid. Dat is een praktische manier van ordenen: de zaken die bij elkaar horen bij elkaar in de buurt zetten. Toch zijn we er daarmee niet. We willen vaak ook met woorden collecties doorzoeken en de documenten die we hebben, vindbaar maken voor onszelf of anderen. In dit hoofdstuk maken we daarom de overstap naar de indirecte toegankelijkheid: het vinden van documenten door als tussenstap woorden in te tikken in een zoekmachine of catalogus.

In dit hoofdstuk bespreken we de uitgangspunten van woordenlijsten zoals die gebruikt worden in het informatievak om informatie toegankelijk te maken en terug te kunnen vinden. Daarbij gaan we in op de verschillende soorten woordenlijsten die daarvoor bestaan. Daarna richten we ons op de meest uitgewerkte vorm van een woordenlijst: een thesaurus. De regels waaraan een thesaurus voldoet zijn namelijk ook toepasbaar op alle andere lijsten. Eerst gaan we in op het doel van een thesaurus, wat een thesaurus precies is en wat een term of concept is. In de rest van het hoofdstuk leggen we aan de hand van vijf stappen uit hoe een goede thesaurus gemaakt wordt. Dat zijn achtereenvolgens:

1. Bepalen van het doel van een thesaurus.
2. Verzamelen van de benodigde begrippen en vaststellen van de termen. Hier spelen de begrippen termredactie (hoe moeten woorden geschreven en gespeld worden), termdefinitie (welke lading dekt de term) en de keuze van termen een rol.
3. Maken van homogene groepen. Daarbij worden de termen die aan elkaar verwant zijn gegroepeerd.
4. Leggen van relaties tussen de termen: een essentieel onderdeel van een thesaurus. Daarbij gaan we in op hiërarchische, associatieve en gelijkwaardigheidsrelaties.
5. De laatste stap, het ordenen van de termen, betreft de presentatievormen van de uiteindelijke lijst met termen.

4.1 Woordenlijsten

Wie in Funda zoekt op de term 'Hilversum', vindt alle woonhuizen die in Hilversum te koop zijn. Wie in een beeldbank als Shutterstock zoekt op de term 'tuinen', vindt allemaal beeldmateriaal met tuinen. Dat materiaal wordt gevonden doordat de term 'Hilversum' of 'tuinen' toegevoegd is aan

de items die gevonden worden. Daarbij wordt soms gebruik gemaakt van een (standaard) woordenlijst.

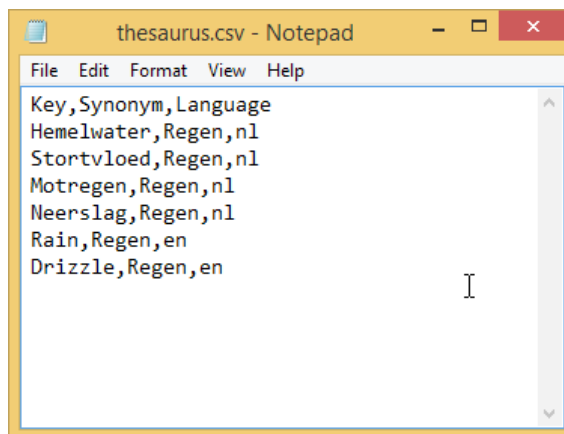
Woordenlijsten kunnen enorm variëren in omvang en mate van complexiteit. In oplopende volgorde van complexiteit onderscheiden we de volgende mogelijkheden:

- **Keuzelijsten:** Dit zijn korte lijsten die er voor zorgen dat één schrijfwijze gebruikt wordt, zowel voor het zoeken als voor het invullen (geslacht, dagen van de week, etc.). Denk aan allerlei formulieren die digitaal ingevuld worden, zoals in afbeelding 4.1.



Afb. 4.1 - Keuzelijst.

- **Synoniemenlijsten:** Hierbij worden termen geassocieerd met een aantal synonieme of quasi-synonieme woorden, zodat bij het gebruik van één van die woorden automatisch op alle geassocieerde woorden wordt gezocht. Zo kunnen de termen 'regen', 'hemelwater', 'stortvloed', 'motregen' en 'neerslag' allen gebruikt worden en aan elkaar gekoppeld zijn. Wie zoekt op één van de termen, blijkt automatisch ook te zoeken op de daaraan gekoppelde termen.



Afb. 4.2 - Synoniemenlijst.

- **Autorisatielijsten:** Het verschil tussen deze lijst en de synoniemenlijst, is dat hier een voorkeursterm of een voorkeurschrijfwijze is gekozen. In het bovenstaande voorbeeld zou dat bijvoorbeeld 'neerslag' kunnen zijn. Die voorkeur kan op verschillende manieren gebruikt worden, maar kenmerkend ervoor is dat als een gebruiker bijvoorbeeld 'VK' intypt, de computer dit herkent en er 'Volkskrant' of zelfs 'De Volkskrant' van maakt. Als

Term	gevleugelden
Use	vogels

Afb. 4.3 - Autorisatielijst.

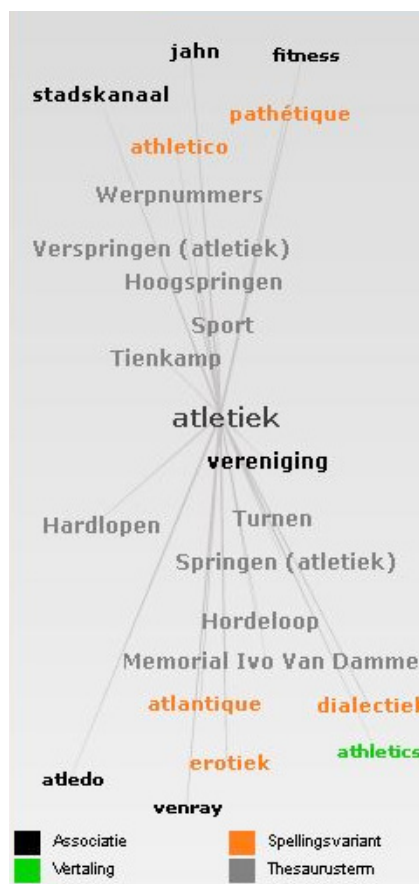
vervangend zoeken

dit soort lijsten wordt gebruikt om termen toe te voegen aan content, is een voorkeursterm of voorkeursschrijfwijze prettig. Als indexeerder kies je dan “vanzelf” de juiste term. Bij het maken ervan spreekt men wel van een use/used for-relatie en van voorkeurstermen en verwijstermen. Bij het zoeken heet deze functionaliteit vervangend zoeken omdat de door de zoeker ingetypte niet-voorkeursterm vervangen wordt door de voorkeursterm.

- **Hiërarchische lijsten:** Een andere term hiervoor is de in hoofdstuk 3 besproken *descriptive taxonomy*. Daarbij is sprake van overkoepelende (Engels: *broader*) en onderliggende (Engels: *narrower*) termen (BT/NT). In hoofdstuk 8 geven we een voorbeeld hoe SharePoint dit doet.

- **Associatieve lijsten:** Woorden- of termenwolken zijn de meest voorkomende vorm hiervan. Een term kan op verschillende manieren (handmatig of automatisch gegenereerd) in verband worden gebracht met andere termen. Dat kan eruitzien als in afbeelding 4.4. Zo had de term Kroatië jarenlang een associatie met oorlog. Op het moment van schrijven heeft dit land voornamelijk een associatie met toerisme. Associaties kunnen dus ook veranderen. In de termenwolken kunnen ook weer allerlei andere relaties zichtbaar worden.

- **Thesauri:** Als de woordenlijst synoniemen, hiërarchische structuren en associatieve verwijzingen bevat en deze worden gebruikt in het toegankelijk maken van en het zoeken naar informatie, dan spreken we van een thesaurus. Die kan zichtbaar of achter de schermen ingezet worden om optimale precisie en optimale vangst te bereiken. Kortom: hij helpt je te vinden waarnaar je zoekt. Een thesaurus voldoet aan een aantal regels die omschreven zijn in de internationale standaard ISO 25694 (zie ook het betreffende kader).



Afb. 4.4 - Associatieve lijst (aquabrowser).

- **Ontologieën:** Een ontologie is een strikt en uitputtend schema voor een bepaald onderwerpsdomein, meestal in een hiërarchische structuur, dat alle relevante grootheden en hun relaties bevat en ook de regels waaraan die grootheden en relaties binnen dat domein voldoen. Omdat een ontologie kennis beschrijft en niet primair bedoeld is om informatie toegankelijk te maken, gaan we er in dit hoofdstuk niet verder op in. In hoofdstuk 9 komt de ontologie wel aan de orde.

In de bovenstaande opsomming is te zien dat termen op verschillende manieren toegepast kunnen worden. Maar wat komt er nu kijken bij het maken van zo'n termenlijst? Wie een lijstje maakt van de dagen van de week, komt daar meestal nog vanzelf uit, maar als de lijst wat langer wordt, komen er ook meer problemen om de hoek kijken. Want moet een term nu in het meervoud staan, of niet? Wat doen we met namen en synoniemen? De regels voor het bouwen van een thesaurus zijn vastgelegd in een ISO-norm. Deze regels zijn ook nuttig voor eenvoudiger vormen van lijsten en daarom gebruiken we de principes van thesaurusbouw als basis voor dit hoofdstuk. Voor wie nog niet eerder met een thesaurus gewerkt heeft, zijn de voorbeelden in hoofdstuk 13 wellicht behulpzaam.

In "Organiseer je informatie" (2010) zijn aan de uitleg over termen, relaties en het bouwen van een thesaurus drie hoofdstukken gewijd. Ieder die zich werkelijk wil verdiepen in deze materie, raden we aan om deze hoofdstukken te raadplegen. Voor diegenen die een meer globale indruk willen krijgen of met meer eenvoudige toepassingen rondom woorden, tags en termen te maken krijgen, gaan we ervan uit dat dit hoofdstuk volstaat.

4.2 Doel van een thesaurus

Een thesaurus is bedoeld om zowel de indexeerder als de zoeker de juiste term te laten gebruiken voor hetzelfde concept. Om dit doel te kunnen bereiken, is het nodig dat de thesaurus een overzicht biedt van alle concepten die nuttig zouden kunnen zijn voor een bepaald doel in een bepaalde omgeving. De concepten worden weergegeven als termen en voor ieder concept is één van de mogelijke representanten van het concept weergegeven als voorkeursterm. Daarnaast zou een thesaurus de voorkeurstermen op een dusdanige manier moeten presenteren dat mensen gemakkelijk de termen daaruit kunnen vinden die ze nodig hebben. Dat wordt bereikt door gebruik te maken van relaties tussen de termen en tussen concepten. Deze relaties worden gebruikt om de termen gestructureerd te tonen.

Als een thesaurus wordt gebruikt tijdens het indexeren en zoeken, dan kunnen de termen en hun relaties gebruikt worden voor:

- het uitbreiden van de zoekvraag;
- het geven van suggesties voor alternatieve (zoek)termen;
- het presenteren van specifieke termen om een zoekvraag te specificeren;
- het opsporen van spelfouten;
- de ondersteuning van automatisch indexeren.

Een goede voorkeursterm aan een concept toekennen is soms lastig, want een concept kan soms door meerdere termen worden beschreven (synoniemen), maar ook kan een term meerdere betekenissen hebben (homografen, homoniemen). Een gecontroleerde informatietaal zoals een thesaurus helpt dan door de betekenis van een concept of term ondubbelzinnig vast te leggen. Wie een woord in een woordenboek zoekt, weet dat een woord vaak meerdere betekenissen heeft. Zo heeft de term 'ontsluiting' te maken met de geboorte, maar ook met het toegankelijk maken van informatie. In een thesaurus heeft elke term één specifieke betekenis binnen de context van het onderwerpsdomein waar het voor gemaakt wordt. De structuur van een thesaurus, vooral de hiërarchische relaties, geeft vaak al een idee van die betekenis van een term. De term 'bank' in een context van financiën, levert een ander beeld op dan hetzelfde woord in een context van meubels.

4.3 Wat is een thesaurus?

Nu we een globaal beeld hebben van wat een thesaurus doet en kan, is het handig aandacht te besteden aan de definitie van een thesaurus. Het beschrijven van de relaties tussen termen is een belangrijk kenmerk van de thesaurus. Hieronder staat een voorbeeld van de thesaurus van de Politieacademie. Hier staat vermeld dat 'politie-uitrusting' de voorkeursterm is (en 'uitrusting politie' de niet-voorkeursterm). 'Politie-uitrusting' is een topterm, want deze term heeft geen bredere termen. Wel is er een aantal meer specifieke termen zoals 'politie-uniformen' en 'politiehonden'. Wie belangstelling heeft voor het zoeken naar 'politie-uitrusting', is misschien ook geholpen met de termen 'politieorganisatie', 'portofoons' en 'veiligheidsvesten'. Termen die niet hetzelfde betekenen, maar wel een associatie oproepen met de term 'politie-uitrusting'. Je ziet dit weergegeven in de thesaurus van de Politieacademie.

thesaurus

De officiële definitie van een thesaurus luidt:

Een geordende verzameling van termen die bestaat uit zoveel mogelijk enkelvoudige begrippen op basis van natuurlijk taalgebruik, waarvan de onderlinge semantische relaties zijn vastgelegd en welke gebruikt wordt voor indexeren en postcoördinatief zoeken.

We zullen de onderdelen uit deze definitie afzonderlijk behandelen:

- De thesaurus bevat een geordende verzameling. Meestal worden de termen van de thesaurus alfabetisch gepresenteerd. Dit heeft als gevolg dat termen dus niet op betekenis bij elkaar staan. ‘Zee’ en ‘zeep’ hebben inhoudelijk weinig met elkaar te maken, maar komen in elkaars buurt voor, door de vaak gekozen alfabetische presentatie.
- Het gaat om zoveel mogelijk enkelvoudige begrippen. We komen daar later in dit hoofdstuk nog op terug.
- De termen zijn gekozen of samengesteld uit de natuurlijke taal. Het is dus niet de bedoeling dat gekunstelde woordcombinaties zoals ‘personenautoproductiegeschiedenis’, omschrijvingen zoals ‘vervoermiddelen naar gebruik’ of codes zoals ‘845.3’ als term voorkomen.

semantiek

- De semantiek is de studie die gaat over de betekenis van woorden. Semantische relaties zijn dus relaties tussen termen die iets over de betekenis zeggen. Waar het alfabet als ordeningsmechanisme wordt gebruikt, komt de betekenis van de termen terug in de vermelde relaties. Een auto is een soort vervoermiddel. ‘Auto’ en ‘vervoermiddel’ hebben derhalve een semantische relatie met elkaar en deze relatie is in de thesaurus zichtbaar.
- Indexeren is de handeling die we verrichten als we termen toekennen aan informatie. Wie een foto opslaat en er een tag ‘oma’ aan toevoegt, indexeert op dat moment.

postcoördinatief

- Tenslotte het postcoördinatief zoeken. Dat klinkt ingewikkeld, maar iedereen die in Google twee of drie woorden intypt, is al postcoördinatief aan het zoeken. Het houdt in dat het leggen van een (syntactische) relatie (de coördinatie) tussen die woorden pas na de bouw- of indexeerfase (vandaar post) bij het zoeken gedaan wordt. De thesaurus is een postcoördinatieve informatietaal.

politie-uitrusting

Gebruikt voor:

- [uitrusting politie](#)

Meer specifiek:

- [politie-uniformen](#)
- [politiebewapening](#)
- [politiehonden](#)
- [politiepaarden](#)
- [politievoertuigen](#)

Zie ook:

- [politieorganisatie](#)
- [portofoons](#)
- [veiligheidsvesten](#)

Afb. 4.5 - Thesaurus Politieacademie.

Een thesaurus maakt gebruik van een aantal afkortingen die we hieronder kort beschrijven:

TT - <i>Top term</i>	de hoogste term in de hiërarchie van een term.
BT - <i>Broader term</i>	de term waaronder de huidige term valt.
NT - <i>Narrower term</i>	de meer specifieke termen van een term.
RT - <i>Related term</i>	de associatieve relaties van een term.
UF - <i>Used For</i>	de verwijsterm die aan de term is verbonden.
USE	de verwijzing naar een voorkeusterm.

In de uitleg van wat een thesaurus is, zijn we over een woord heengestapt: term. Wat is een term nu eigenlijk? En wat is het niet? Alvorens de stappen voor het bouwen van een thesaurus te behandelen, willen we met meer nadruk naar dit begrip kijken.

4.4 Concepten en termen

woord

Een term, zo stellen we, is niet gelijk aan een woord. Een woord is een te isoleren spraakgeluid dat op zichzelf een betekenis heeft. Een woord is niet altijd een term, maar een term bestaat wel uit een of meer woorden.

Een term is een woord, of woordgroep, dat je kiest om te dienen als metadata. Het moet dus aan een aantal eisen voldoen. De belangrijkste eis is die van ondubbelzinnigheid in betekenis. Zo is 'bank' een woord, maar het kan niet meteen dienen als term. Daarvoor is een context nodig, waaruit blijkt of het om een meubelstuk of financiële instelling gaat.

concept

De ISO 25964 stelt dat thesauri de relaties tussen concepten leggen, waarbij een concept een cognitieve eenheid is (Engels: *unit of thought*), die uitgedrukt kan worden in termen. Een term is dan een label voor het concept, maar een concept kan beschreven worden door middel van meerdere termen, denk maar aan het concept 'regen' dat beschreven kan worden door de term 'regen', 'rain' of 'pluie', afhankelijk van de taal.

Een concept kan ook in meerdere woorden beschreven worden, zoals 'allergie bij kinderen', waarbij in een thesaurus dit ene concept uitgedrukt wordt in meerdere eenvoudiger concepten, zoals 'allergie' en 'kinderen'.

term

Die eenvoudige concepten zijn dan hetzelfde als termen. Een term is te omschrijven als: "een woord of woordgroep dat een concept kenmerkt". Een term kan dus van alles zijn: dingen, materialen, activiteiten, processen, evenementen, eigenschappen van mensen, acties, disciplines, onderwerpen of hoeveelheden.

In een thesaurus streven we naar zoveel mogelijk enkelvoudige begrippen. Voorbeelden van enkelvoudig begrippen zijn 'fietsen', 'wetenschappers'

of ‘schilderijen’. De Nederlandse taal bestaat echter niet alleen uit enkelvoudige, maar ook uit samengestelde begrippen: ‘onderzoeksleders’, ‘railvervoer’, ‘klassieke muziek’ of ‘ecologische kleding’.

Wanneer is een woord of woordgroep dat een concept kenmerkt dan geschikt om als term te functioneren en wanneer niet? In de volgende paragraaf kijken we meer in detail naar het proces van thesaurusbouw.

4.5 Proces thesaurusbouw

De stappen die je doorloopt bij het bouwen van een thesaurus komen sterk overeen met de stappen die we al kennen bij het ordenen van informatie zoals in hoofdstuk 3 aan de orde is gekomen. Wie een eenvoudige woordenlijst wil gebruiken zal sommige van de hier volgende stappen sneller af kunnen handelen en naar eigen inzicht die theorie toepassen die van nut is in de eigen situatie:

1. Bepaal het doel van de thesaurus.
2. Verzamel de termen en stel ze vast:
 - 2a verzamel en selecteer de begrippen
 - 2b stel de termen vast.
3. Maak homogene groepen:
 - o geheel/deel
 - o geslacht/soort.
4. Bepaal relaties en maak verwijzingen:
 - o gelijkwaardigheid
 - o hiërarchie
 - o associatief.
5. Orden alfabetisch.

Stap 1: Bepaal het doel van de thesaurus

Bij elk stappenplan kom je deze stap tegen en dit is dus zo langzamerhand een open deur. Toch blijft het een essentieel aspect; hieronder valt immers de omschrijving van het domein en de gebruikers. Daarnaast is het noodzakelijk te besluiten wat voor lijst precies nodig is en of er wellicht al elders een gebruikt wordt. Allereerst is het aan te raden om te kijken wat er al beschikbaar is. Als er thesauri bestaan op het onderwerpsgebied die geheel of gedeeltelijk, al dan niet aangepast, overgenomen kunnen worden dan is dat aan te raden, omdat het waarschijnlijk zowel tijd als kosten bespaart. In hoofdstuk 13 geven we een paar voorbeelden van in Nederland gebruikte thesauri.

Daarnaast is het handig om te bedenken welke vorm de lijst moeten hebben en wat er precies getoond moet worden. Als een eenvoudige autorisatielijst kan volstaan voor het doel, dan scheelt ook dat veel tijd en geld.

Wie hierover nadenkt kan iets hebben aan de lijst die in het kader “Checklist” na dit hoofdstuk is opgenomen. Hierin wordt een opsomming gegeven van een aantal rubrieken. Ook is van belang om vast te stellen of de lijst geïntegreerd moet worden met andere systemen (en of dat mogelijk is). Ook is het handig al in dit stadium te besluiten of, en zo ja, hoe vaak de lijst een update nodig heeft en wie daar de rechten voor krijgt om dat te doen. Een lijst met dagen van de week zal ongewijzigd blijven, maar een lijst van mogelijke vergoedingen, onderwerpen of projecten heeft regelmatig updates nodig.

Dat betekent dat het in dit stadium ook handig is om na te denken over de volgende zaken:

- Wie is de verantwoordelijke voor dit project? Dat hoeft niet de uitvoerder van het project te zijn, maar wel iemand die beslissingsbevoegdheid heeft.
- Wie werken er mee? Het is essentieel om tijdens het maken van grotere lijsten de hulp van mensen te krijgen die met een systeem werken, bijvoorbeeld om de juiste zoektermen te gebruiken en te blijven gebruiken.
- Hoe vaak en met wie hebben we overleg? Deze momenten zijn essentieel omdat daar de discussie over termen kan plaatsvinden. Dat houdt aanstaande gebruikers betrokken bij het ontwerp, waardoor de kans op een succesvolle implementatie ook nog eens toeneemt.
- Welke informatieprofessional is beschikbaar voor het maken en bijhouden van de termenset die gebouwd wordt? En welke IT-professional is bereid om te helpen bij problemen die ontstaan op technisch gebied?
- Welke software gebruiken we? Niet alle software ondersteunt alle functies die mogelijk zijn voor woordenlijsten. Basisvereisten voor het maken van een echte thesaurus zijn in ieder geval dat er geen limiet is aan het aantal op te nemen termen, noch aan de lengte van termen, *scope notes* of andere velden. Ook moet er geen gestelde limiet zijn aan het aantal hiërarchische niveaus. Tenslotte moet het mogelijk zijn om gelijkwaardigheidsrelaties, hiërarchische relaties en associatieve relaties, eenvoudig uit te wisselen met andere software (en evt. zoeksystemen).

Stap 2a: Verzamel en selecteer de begrippen

Bij het verzamelen van de begrippen hoeft de bouwer zich nog niet zo druk te maken over de juiste schrijfwijze, betekenis, enzovoort. In dit stadium is het een kwestie van de woorden vinden waarmee de onderwerpen van het domein benoemd kunnen worden.

Er zijn verschillende manieren om achter de woordenschat van een domein te komen. Als een reeds bestaande thesaurus op het onderwerpsgebied niet geheel kan worden overgenomen, dan kan er wel

gekeken worden of toch een gedeelte van de terminologie interessant kan zijn. Ook kunnen er al historisch gegroeide trefwoordenlijsten of een mappenstructuur op een netwerkschijf aanwezig zijn die als basis voor de nieuw te bouwen thesaurus kunnen dienen.

Als het niet mogelijk is om van bestaande lijsten gebruik te maken, dan zal de thesaurusbouwer zelf op zoek moeten naar de te gebruiken termen. Als er al een retrievalsysteem is, dan kan gekeken worden naar de daarin toegekende tags of termen. Als er een lijst van gebruikte termen bestaat, dan is er een (begin van) een woordenlijst beschikbaar. Ook kunnen beschrijvingen van representatieve documenten uit de collectie of die documenten zelf nader bekeken worden. Wellicht is er een lijst beschikbaar van zoekgedrag of de log van een website met zoektermen. Een andere optie is om naslagwerken zoals speciale encyclopedieën als bron van de te gebruiken termen te raadplegen. In alle gevallen is het echter belangrijk dat uiteindelijk een lijst ontstaat die dicht bij de belevingswereld van de gebruiker staat. Houd daarom ook contact met de gebruikers en leg in twijfelgevallen termen aan hen voor. Tegelijkertijd is ook de selectie van belang: welke soorten termen nemen we wel op en welke niet?

Stap 2b: Stel de termen vast

termkeuze

termredactie

termdefinitie

Natuurlijk is het wel zaak kritisch te zijn ten aanzien van de al gebruikte terminologie. Daarvoor kun je de eisen toepassen van termkeuze, termredactie en termdefinitie. Termkeuze gaat, zoals het woord al zegt, over welk woord we gebruiken om een bepaald concept te beschrijven, want vaak zijn er meer mogelijkheden. Termredactie gaat over de juiste schrijfwijze van de term. Termdefinitie gaat over de betekenis van de term, zodat een term een uniek concept binnen de thesaurus weergeeft. Om er gemakkelijk mee te kunnen werken, hebben we ervoor gekozen om de eisen ten aanzien van termkeuze, termredactie en termdefinitie weer te geven in de vorm van een aantal vragen.

Termkeuze

- Maak je gebruik van de officiële naam?
Wil je een term voor spinnen hebben, gebruik je dan 'spinnen' of de wetenschappelijke naam 'arachnia'? Het antwoord is dat er geen vaststaande keuze is, maar neem die term waarvan je verwacht dat de gebruikers hem zullen hanteren bij hun zoekacties.
- Mogen buitenlandse woorden (leenwoorden) ook?
Ook hier geldt: neem de meest voorkomende term. We kennen in het Nederlands een heleboel leenwoorden die oorspronkelijk uit een andere taal komen. Die vormen intussen deel van het Nederlands. In het algemeen stellen we voor een Nederlandse term te nemen, tenzij de buitenlandse term meer bekend is. Dus liever 'auteursrecht' dan 'copyright' maar liever 'laptop' dan 'shootcomputer'.

- Wat als er twee termen hetzelfde begrip beschrijven?
Veel concepten kunnen beschreven worden met meerdere termen. Zo beschrijven de termen 'rijwiel' en 'fiets' allebei hetzelfde concept. Gebruik de meest gangbare term. Dat wordt de voorkeursterm (zie stap 4).
- Mogen afkortingen voorkomen?
Wel als het een woord geworden is. Horeca is formeel een afkorting van HOfel, REstaurant, CAFé, maar wordt als woord gebruikt. Over het algemeen geldt echter: kies liever de uitgeschreven vorm. 'Ondernemingsraad' heeft dus de voorkeur boven 'OR'. Een extra argument om woorden uit te schrijven is dat afkortingen meestal niet uniek zijn. Zo staat VK bijvoorbeeld voor de Volkskrant, maar ook voor VKontakte, een Russische social networking site.
- Mogen namen voorkomen?
In principe biedt een thesaurus geen plaats aan namen. De reden daarvoor is dat er te veel namen zijn. Wel kan er een aparte plaats- of persoonsnamenlijst worden aangelegd als dat gewenst is. Als het gaat om een beperkt en vaststaand aantal namen, zoals de namen van afdelingen, of een overzicht van aandoeningen die een zorgverzekeraar vergoedt, kan hierop een uitzondering worden gemaakt. Hoofdstuk 5 gaat hier meer specifiek op in.

Termredactie

- Mag een term spaties bevatten?
Ja. Een term mag bestaan uit meerdere woorden, wanneer ze als geheel maar één afzonderlijk concept beschrijven. 'Klassieke muziek' is de beschrijving van een specifieke muzieksoort en kan als zodanig als term functioneren.
- Mag een term bestaan uit een samenstelling?
Ja. Samenstellingen zijn twee of meer woorden die aan elkaar geschreven worden. We gebruiken ze veel in het Nederlands, meer dan bijvoorbeeld in het Engels. 'Kunstgeschiedenis' en 'ziekenhuis' zijn voorbeelden van samenstellingen die kunnen worden gebruikt als term. Van belang is dat het nog wel een "enkelvoudige begripseenheid" blijft. Een complex concept als: 'niet-westerse allochtone vrouwen' zul je altijd splitsen. Deze term kan uiteenvallen in twee delen: 'allochtone vrouwen' en 'niet-westerse vrouwen'.
- Is elke samenstelling automatisch geschikt als term?
Nee, dat is niet het geval. Gekeken moet worden of de samenstelling als losse termen dezelfde betekenis houdt. We bekijken dit aan de hand van het woord 'automotor'. Het gaat hier om de motor van een auto. Samengesteld betekenen de woorden precies hetzelfde als wanneer ze los gebruikt worden. In dat geval kun je twee dingen besluiten:
 - a. Twee termen opnemen: 'auto' en 'motor' en deze door de zoeker laten combineren.
 - b. Opnemen als één term. Wanneer de termen niet gesplitst worden en een hele lange lijst ontstaat met thesaurustermen die beginnen

met ‘auto’, zoals ‘automotoren’, ‘autostoelen’, ‘autowielen’, enzovoort, dan kan dat als signaal gebruikt worden om een term alsnog te splitsen.

- Hoe worden complexe termen verwerkt?

Naast termen die in twee delen gesplitst kunnen worden, komen ook zogenaamde complexe termen voor. Deze zullen we niet uitputtend behandelen, hoewel ze regelmatig voorkomen. Voorbeelden hiervan zijn ‘deeltijd beroepsonderwijs’, ‘Victoriaanse zilveren theepotten’, ‘biologische diepvriesmaaltijden’, ‘ecologische dameskleding’, enzovoort. Deze zijn als volgt te beschouwen: het kernwoord kan op meerdere delen slaan, bij ‘deeltijd beroepsonderwijs’ gaat het feitelijk over de twee delen, namelijk ‘deeltijd’ en ‘beroeps’ die samen de kern ‘onderwijs’ bepalen. Deze termen splitst men doorgaans zodat er twee termen ontstaan. In dit geval is dat ‘deeltijdonderwijs’ en ‘beroepsonderwijs’. Deze samenstellingen zullen we doorgaans behouden. Meer over complexe termen is te vinden in hoofdstuk 3 van “Organiseer je informatie” of in de ISO norm 25964-1 op pagina 40.

- Staat een term altijd in het meervoud?

We hebben hiervoor al een aantal voorbeelden gezien waarin termen in het meervoud staan. Als het om concrete begrippen gaat, kiezen we inderdaad voor de meervoudsvorm. Bij abstracte begrippen doen we dat doorgaans niet. Beide soorten begrippen vinden we terug in een thesaurus.

Concrete termen zijn: ‘fietsen’, ‘tafels’, ‘activisten’.

Abstracte termen zijn: ‘hindoeïsme’, ‘vermoeidheid’, ‘wereldoorlog’; ‘onderwijs’.

concrete termen

abstracte termen

Termdefinitie

Soms is het wenselijk of zelfs noodzakelijk om de betekenis van een term toe te lichten om ervoor te zorgen dat hij uniek is in de termenlijst. Dat is het gebied dat bestreken wordt door termdefinitie:

- Wat doe ik met een term die twee betekenissen heeft?

De betekenis van een term kan op zichzelf staand onduidelijk zijn. Bijvoorbeeld omdat het een homoniem is: termen met verschillende betekenis, die gelijk geschreven worden. Door een qualifier of kwalificatie te gebruiken is het mogelijk om onderscheid te maken tussen homoniemen. Een voorbeeld hiervan is ‘depressie’. In een thesaurus van de GGZ zal deze term niet in de context van een lagedrukgebied of laagconjunctuur voorkomen. Als een term binnen een vakgebied ondubbelzinnig gekoppeld kan worden is een kwalificatie niet nodig. Wanneer toch meerdere betekenissen mogelijk zijn, omdat de thesaurus een breed gebied bestrijkt, dan kan verwarring voorkomen worden door gebruik te maken van een qualifier. De qualifier vormt onderdeel van de term, dus als besloten is om ‘depressie (stemmingsstoornis)’ op te nemen, zal in de thesaurus de term ‘depressie (stemmingsstoornis)’, naast ‘depressie (laagconjunctuur)’ of ‘depressie (lagedrukgebied)’ voorkomen.

qualifier

scope note

- Kan ik ook méér uitleg kwijt bij een term?

Ja, door middel van een zogenaamde *scope note*. Een *scope note* stelt de grenzen van een term vast. In een thesaurus wordt dit weergegeven als een apart veld, vaak met de afkorting SN. Een voorbeeld hiervan is:

Ontsluiting

SN Ontsluiting is het vindbaar of toegankelijk maken van informatie. Ontsluiting omvat alle activiteiten en faciliteiten om documenten en de daarin vervatte informatie terugvindbaar te maken; kortom, ontsluiting faciliteert het zoeken naar informatie.
([https://nl.wikipedia.org/wiki/Ontsluiting_\(informatiewetenschap\)](https://nl.wikipedia.org/wiki/Ontsluiting_(informatiewetenschap)))

Wanneer alle termen bekend zijn, zijn ze intussen ook ergens genoteerd. Wie begint met een thesaurus of een thesaurus maakt met een beperkt aantal termen, kan baat hebben bij een start door middel van papieren kaarten. Deze kaart kan de vorm hebben van een eenvoudige tabel zoals in afbeelding 4.6 is weergegeven. Hierop is een aantal van de afkortingen die in paragraaf 4.3 zijn genoemd te herkennen.

Term	Notatie
Datum invoer/wijziging	
SN	NT
UF	RT
TT	Bron
BT	Opmerkingen

Afb. 4.6 - Voorkeurstermkaart.

thesaurussoftware

Bij een grotere hoeveelheid termen wordt een papieren opzet onoverzichtelijk. Dan is de overstap naar speciale thesaurussoftware handig om daarmee de thesaurus verder uit te bouwen en te onderhouden. Deze software biedt de mogelijkheid termen in te voeren, de verwijzingen op te nemen en deze ook meteen te controleren. Daarnaast zijn met behulp van deze software overzichten uit te draaien en administratieve kenmerken toe te voegen. Er zijn verschillende softwareprogramma's, maar ze moeten allemaal minimaal bieden wat ook op de papieren kaarten staat.

Tenslotte kan men in Excel of Word lijstjes aanmaken. Deze bieden natuurlijk geen extra faciliteiten, maar een alfabetische rangschikking van termen is dan wel mogelijk.

Stap 3: Maak homogene groepen

Nu de termen bekend zijn en geanalyseerd op hun geschiktheid voor opname in een lijst, worden ze als eerste in grove groeperingen gedeeld.

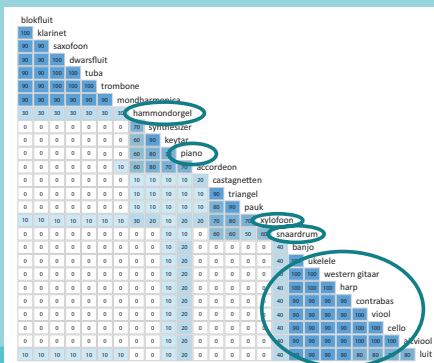
Op basis van die groepering kan bepaald worden welke relaties er straks kunnen worden gelegd. Het gaat dan zowel om gelijkwaardigheidsrelaties, als om hiërarchische- en associatieve relaties. Dit groeperen heeft als voordeel dat ook te zien is of er zogenaamde verweesde termen zijn. Deze hebben immers geen groep waar ze bij horen.

Termen kunnen handmatig of automatisch worden gerangschikt naar hun betekenis, zodanig dat er groepen van termen ontstaan die bij elkaar horen, vooral uitgaande van geslacht-soort en geheel-deel relaties tussen termen. De betekenis van de termen is daarbij leidend. Een vrachtwagen is een soort vervoermiddel, dus de termen 'vrachtwagens' en 'vervoermiddelen' behoren tot dezelfde homogene groep (semantische relatie). Een vrachtwagen kan melkpakken vervoeren, maar dat wil niet zeggen dat daarom de termen 'vrachtwagens' en 'melkpakken' tot dezelfde homogene groep behoren, want hier gaat het om een syntactische relatie.

Voor wie handmatig te werk gaat, is het eenvoudig. De kaarten die, gelet op hun betekenis, bij elkaar horen, worden bij elkaar gelegd. In een geautomatiseerde situatie is dit soms wat lastiger, hoewel er ook software bestaat die deze stap kan vergemakkelijken. Het bij elkaar leggen wat bij elkaar hoort kan ook als een spel worden uitgevoerd door verschillende (groepen) gebruikers. Het eerder besproken *cardsorting* is een voorbeeld van software die hiervoor gebruikt wordt.

Similarity Matrix:

- hammondorgel, piano, xylofoon en snaardrum blijken lastig te plaatsen
- snaar- en strijkinstrumenten zijn door de meesten bij elkaar geplaatst, op de piano na dan...



Afb. 4.7 - Similarity matrix uit Cardsorting gemaakt door HvA studenten Emy Groot-Bluemink en Toby Kok.

Stap 4: Bepaal de relaties en maak verwijzingen

Relaties zijn essentieel in een thesaurus. Ze leiden de indexeerder en de gebruiker naar de beste term om een concept uit te drukken. Nu is het zaak om te bepalen welke semantische relaties de termen onderling hebben en tot welke verwijzingen dat leidt. In paragraaf 3.2 is het onderscheid tussen syntactische en semantische relaties al aan de orde geweest.

De verwijzingen noteer je bij de term op de daarvoor bestemde plaats (zie afbeelding 4.6 en 4.8).

Gelijkwaardigheidsrelaties

gelijkwaardig- heidsrelaties

Thesauri, maar ook sommige andere lijsten, bieden de mogelijkheid om gelijkwaardigheidsrelaties of equivalentierelaties aan te maken. Dit vergroot immers het zoekgemak van de gebruiker. Gelijkwaardigheidsrelaties zijn relaties tussen twee termen met gelijke (of gelijkwaardig toegekende) betekenis. Als de bouwer het mogelijk wil maken om op beide termen te zoeken kan hij de ene term verheffen tot voorkeursterm en de andere tot verwijstern. Zo kan besloten worden om de termen 'auteursrecht' en 'copyright' beide op te nemen, maar de ene (bijvoorbeeld 'auteursrecht') te zien als voorkeurstern. Dat is de term die gebruikt wordt bij het toekennen van metadata en waarvoor de termkaart wordt ingevuld (zie afbeelding 4.6). De term 'copyright' wordt automatisch de niet-voorkeurstern waarvoor een beperktere kaart wordt ingevuld (zie afbeelding 4.8). Deze term wordt niet zelf gebruikt, maar iedereen die op die term zoekt, krijgt (bij voorkeur automatisch) het zoekresultaat van 'auteursrecht'.

Wie gebruikt maakt van thesaurussoftware regelt dit door bij 'copyright' het veld USE 'auteursrecht' in te vullen. Automatisch vult het systeem bij het record van 'auteursrecht' het veld UF 'copyright' (Used For).

Niet-voorkeurstern
Datum invoer/wijziging
USE
Bron
Opmerkingen

Afb. 4.8 - Niet-voorkeursternkaart.

Het grote voordeel van het kiezen van een van de termen is natuurlijk dat bij deze ene term ‘auteursrecht’ alle informatie over auteursrecht te vinden is, en dat deze niet verspreid is over de termen ‘auteursrecht’ en ‘copyright’, waarbij de zoeker naar informatie zelf moet bedenken deze beide termen te gebruiken. Bij het gebruik van thesaurussoftware blokkeert de software de niet-voorkeursterm, waardoor alleen de voorkeursterm gekozen kan worden bij het indexeren. De zoeker die de niet-voorkeursterm intypt, wordt automatisch naar de voorkeursterm geleid (vervangend zoeken).

Overigens gelden deze regels ook voor de quasi-synoniemen of zelfs voor tegenstellingen. ‘Werkeloosheid’ en ‘werkgelegenheid’ zijn feitelijk tegenstellingen, maar wellicht is bij beide termen dezelfde soort informatie te vinden en dat kan de keuze voor een van de termen als voorkeursterm rechtvaardigen. Tenslotte: een niet-voorkeursterm heeft altijd maar één voorkeursterm, maar een voorkeursterm kan meerdere verwijzingen hebben vanuit niet-voorkeurstermen.

Pop-up store

USE Winkel

Shop

USE Winkel

Winkel

UF Pop-up store

UF Shop

Hiërarchische relaties

Hiërarchische relaties vormen de kern van een thesaurus, maar ook andere lijsten kennen vaak wel hiërarchieën waar soms ook mee gezocht kan worden.

generieke

In de thesaurusbouw kennen we twee soorten hiërarchische relaties. De eerste is de generieke relatie of geslacht/soort relatie. Daarbij is de onderliggende term een soort van een overkoepelende term. ‘Onderwijs’ is bijvoorbeeld het geslacht waarbij ‘Basisonderwijs’ en ‘Middelbaar onderwijs’ soorten daarvan zijn.

partitieve

De andere soort hiërarchische relatie is de partitieve relatie. ‘Handen’ en ‘ellebogen’ vormen allebei onderdeel van de ‘armen’. De overkoepelende term is het geheel en de onderliggende termen zijn delen daarvan.

generiek zoeken

Software die hiërarchische relaties (zowel generieke als partitieve) ondersteunt, biedt soms de mogelijkheid om tegelijk te zoeken op zowel een gekozen term als op de hiërarchisch onderliggende termen. We noemen dit generiek zoeken. Generiek zoeken is alleen mogelijk als de onderliggende termen in de thesaurus zijn aangegeven en gekoppeld aan de bovenliggende term. Voor alle duidelijkheid: dat geldt dus voor zowel generieke als partitieve relaties. Indien zowel vervangend als generiek zoeken mogelijk is, spreken we van een volledige thesaurusfunctionaliteit.

De op hiërarchische relaties gebaseerde verwijzingen laten niveaus van bovenschikking en onderschikking zien. De gebruikte terminologie daarvoor is BT (*Broader Term*) en NT (*Narrower Term*). Onderling verwijzen deze termen naar elkaar. Bij het gebruik van thesaurussoftware worden bij het invullen van de *narrower terms* bij term1, automatisch de spiegelende records van de *narrower terms* aangemaakt waarin “BT term1” vermeld wordt.

Twee voorbeelden:

Record met de term: Onderwijs

NT Basisonderwijs

NT Middelbaar onderwijs

Record met de term: Basisonderwijs

BT Onderwijs

Record met de term: Middelbaar onderwijs

BT Onderwijs

zuiverheid

Belangrijk in een thesaurus is de zuiverheid van de relatie. Dat is met name van belang als generiek zoeken mogelijk is. Anders zullen ongewenste zoekresultaten (ruis) het gevolg zijn.

alle/sommige toets

Om te beoordelen of een relatie zuiver is, kan de alle/sommige toets gedaan worden (Engels: *all-and-some-test*). Hierbij wordt tweemaal een vraag gesteld en het antwoord moet tweemaal ja zijn, voordat de relatie kan worden gelegd. Wanneer men bijvoorbeeld wil vaststellen of een relatie tussen ‘fietsen’ en ‘racefietsen’ of tussen gebouwen en deuren kan worden gemaakt, stellen we de volgende vragen:

Zijn **alle** ‘racefietsen’ ‘fietsen’? (een onderliggende term moet *volledig deel* uitmaken van de bovenliggende term). Het antwoord is JA;

Zijn **sommige** ‘fietsen’ ‘racefietsen’? (een bovenliggende term omvat *sommige* onderliggende termen). Het antwoord is wederom JA.

Dan mag deze hiërarchische relatie gelegd worden.

Zijn **alle** ‘deuren’ deel van een ‘gebouw’? Nee, er zijn ook deuren in kasten, in vliegtuigen, in ovens, enzovoort;

Hebben **sommige** ‘gebouwen’ ‘deuren’? Ja, dat wel.

Deze hiërarchische relatie mag dus niet gelegd worden. Een associatieve verwijzing (hierover later meer) kan wel gelegd worden.

Als er foutieve relaties gelegd, worden, kan er bij het zoeken ruis ontstaan. Stel dat iemand de (foute) relatie maakt:

Gebouwen

NT Deuren

We hebben net gesteld dat deuren op meer plekken voorkomen dan alleen in gebouwen. Stel, we hebben twee documenten:

Installatiebrochure IKEA - Koelkastdeur wisselen van rechts naar links en omgekeerd.

en

Man doet poging uit het vliegtuig te stappen en gebruikt daarvoor de nooddeur van het vliegtuig.

ruis

Beide documenten krijgen onder andere de term 'deuren' mee. Een zoeker gaat later zoeken op de term 'gebouwen'. Hij vindt niet wat hij zoekt en gaat generiek nogmaals zoeken, in de hoop dan meer informatie te krijgen. Hij krijgt dat hopelijk, maar daarnaast ook ruis. Beide bovenstaande documenten worden immers gevonden, maar zijn voor de zoeker niet interessant; de zoeker wilde namelijk informatie vinden over deuren in gebouwen.

Het bovenstaande voorbeeld maakt al duidelijk dat niet alle termen een hiërarchische relatie met elkaar kunnen hebben, omdat ze als ze dat wel hebben, ruis genereren in de zoekactie.

Zo mag een concrete term niet met een abstracte term gecombineerd worden en is 'gebouwen' NT 'beton' dus een niet-toegestane relatie. Overigens helpt ook in deze gevallen de alle/sommige toets.

polyhiërarchische relaties

Termen kunnen ook in twee of meer categorieën ondergebracht worden. Dit noemt men polyhiërarchische relaties. Deze termen hebben meerdere *broader terms*. Twee voorbeelden hiervan zie je in de afbeeldingen 4.9 en 4.10.

Social Media

Scope Note: [Web-based systems or techn multimedia.](#)

Category: [Communications Media](#)

 [Search collection using this descriptor](#)

Broader Terms

[Computer Mediated Communication](#)

[Mass Media](#)

Narrower Terms

N/A

ducklings

Technical Categories

P Plural Form

Subject Category

OG Organism Groups

Term Definition

Young ducks.

Non-preferred Term

duckling

Broader Term

ducks 

young birds 

Afb. 4.9 - Polyhiërarchie Social Media.

Afb. 4.10 - Polyhiërarchie ducks.

De ene thesaurus maakt meer gebruik van polyhiërarchische relaties dan de andere. De MeSH thesaurus (zie ook hoofdstuk 13) is een voorbeeld van een thesaurus waarbij de polyhiërarchische relatie vaak voorkomt. In afbeelding 4.11 zie je dat 'nose' in de MeSH thesaurus in verschillende hiërarchieën voorkomt. Overigens ondersteunen niet alle systemen deze relatie, dus het is zinvol dit in de eigen omgeving te checken alvorens polyhiërarchische relaties aan te maken.

[All MeSH Categories](#)

[Anatomy Category](#)

[Body Regions](#)

[Head](#)

[Face](#)

Nose

[All MeSH Categories](#)

[Anatomy Category](#)

[Respiratory System](#)

Nose

[Nasal Bone](#)

[Nasal Cartilages](#)

[Nasal Cavity](#)

[Nasal Mucosa](#)

[Goblet Cells](#)

[Olfactory Mucosa +](#)

[Nasal Septum](#)

[Vomer](#)

[Paranasal Sinuses](#)

[Ethmoid Sinus](#)

[Frontal Sinus](#)

[Maxillary Sinus](#)

[Sphenoid Sinus](#)

[Turbinates](#)

[All MeSH Categories](#)

[Anatomy Category](#)

[Sense Organs](#)

Nose

[Olfactory Mucosa](#)

[Olfactory Receptor Neurons](#)

[Vomeronasal Organ](#)

Afb. 4.11 - MeSH thesaurus.

Wie thesaurussoftware gebruikt voor het bouwen van een woordenlijst, merkt dat deze een aantal foutcontroles uitvoert. Zo wordt gecontroleerd of een ingevoerde term uniek is. Ook zal meer dan één soort verwijzing tussen dezelfde termen niet worden toegestaan. De onderstaande relatie zal dus niet opgenomen kunnen worden als men met thesaurussoftware werkt:

Wielervedstrijden

NT Racefietsen

RT Racefietsen

Ook worden fouten in de trapsgewijze verwijzingen gesignaleerd. Handmatig zou de onderstaande fout gemaakt kunnen worden. Thesaurussoftware zal deze verwijzing niet toestaan:

Fouten in thesaurus

Vogels

NT kanaries

In een hiërarchische afdruk zie je de term “kanaries” op twee niveaus terugkomen:

Vogels

NT zangvogels

Vogels

...Kanaries

...Zangvogels

.....Kanaries

Zangvogels

NT kanaries

Afb. 4.12 - Zangvogels hiërarchie.

Associatieve relaties

Eerder werd al gesteld dat een hiërarchische relatie alleen gemaakt mag worden als hij zuiver is en dat dat is te testen door gebruik te maken van de alle/sommige toets. Als een hiërarchische relatie niet mogelijk is, dan kan mogelijk een associatieve relatie gemaakt worden. Associatieve relaties zijn semantische relaties die niet als gelijkwaardigheidsrelatie of hiërarchische relatie worden aangemerkt. Nu zijn deze relaties alleen interessant als ze leiden naar voor de gebruiker nuttige verwijzingen. Het gevaar bestaat immers dat een thesaurus overladen wordt met allerlei associatieve verwijzingen. Houd het aantal beperkt door je steeds af te vragen of de gebruiker het resultaat van de zoekactie kan verbeteren als er gewezen wordt op de associatieve relatie. Een associatieve relatie wordt weergegeven als *related term* (RT)- relatie en is altijd wederkerig.

associatieve
relaties

Twee voorbeelden van zinvolle associatieve relaties

Erosie en woestijnvorming hebben iets met elkaar te maken. Het is een zogenoemde causale relatie, want erosie kan woestijnvorming tot gevolg hebben, maar niet alle woestijnen zijn het gevolg van erosie. Een hiërarchische relatie mag dus niet gelegd worden. Op dat moment is de associatieve relatie zinvol.

Erosie

RT Woestijnvorming

Woestijnvorming

RT Erosie

Onderwijs en studenten hebben ook met elkaar te maken. Toch is dit geen geoorloofde hiërarchische relatie, want hier wordt een concrete (studenten) aan een abstracte term (onderwijs) gerelateerd. Bovendien wordt niet alle onderwijs aan studenten gegeven, want een kind in groep 3 is nu eenmaal geen student. Om deze redenen is een associatieve relatie hier meer op zijn plaats.

Onderwijs

RT studenten

Studenten

RT onderwijs

Stap 5: Alfabetisch rangschikken

Als alle verwijzingen, *scope notes*, kwalificaties, enzovoort zijn aangebracht, kunnen de termen alfabetisch worden geordend; de thesaurus is dan in ruwe vorm zichtbaar.

We constateerden al eerder dat het werken met invulkaarten praktisch is bij een kleine verzameling termen of bij het begin van een thesaurusproject. Is de verzameling groter dan een paar honderd termen en is een gedeelte van de structuur zichtbaar, dan is het gebruik van speciale thesaurussoftware aan te raden.

generieke lijst

Thesaurussoftware biedt naast hulp bij het bouwen en onderhouden van de thesaurus ook de mogelijkheid om lijsten te genereren. De bekendste lijsten zijn de alfabetische lijst en de hiërarchische of generieke lijst. De alfabetische lijst toont alle termen met de daarbij horende verwijzingen. De hiërarchische of generieke lijst toont alle thesaurustermen met hun onderliggende hiërarchie. Niet-voorkeurstermen staan daar niet in. Deze boomstructuur is vooral bruikbaar voor de thesaurusbouwer die zo snel na kan gaan of er ergens foutieve en/of onvolledige hiërarchische verwijzingen zijn aangebracht. Ook zijn deze lijsten digitaal of gedrukt

prettig voor de gebruiker om in te grasduinen. Meer visuele vormen zijn ook nuttig, maar de creatie hiervan is vaak een tijdrovend en lastig project.

Een voorbeeld van een alfabetische lijst en een systematische lijst:

Alfabetische lijst	Systematische lijst
Appartementen	Appartementen
BT Woningen	..Flats
NT Flats	..Studio's
NT Studio's	Bedrijfswoningen
NT Bedrijfswoningen	Beweegbare bruggen
BT Woningen	..Draalbruggen
	..Ophaalbruggen
Beweegbare bruggen	Boogbruggen
BT Bruggen	Bouwwerken
NT Draalbruggen	..Bruggen
	Beweegbare bruggen
Boogbruggen	Draalbruggen
BT Bruggen	Ophaalbruggen
	Boogbruggen
Bouwwerken	Hangbruggen
NT Bruggen	Iusbokken
NT Gebouwen	Pontonbruggen
	Voetgangersbruggen
Bruggen	
BT Bouwwerken	
NT Beweegbare bruggen	
NT Boogbruggen	
NT Hangbruggen	
NT Iusbokken	
NT Pontonbruggen	
NT Voetgangersbruggen	

Afb. 4.12 - Alfabetische en systematische lijst.

Het is belangrijk dat woordenlijsten zoals thesauri worden onderhouden. Een onderwerpsgebied blijft altijd in beweging en de woorden die men gebruikt bij het zoeken naar onderwerpen ook. Afhankelijk van de grootte en complexiteit van het onderwerp en de snelheid van de veranderingen in het vakgebied zullen er dus met enige regelmaat wijzigingen in de lijst moeten worden aangebracht. Wijzigingen in de thesaurus leiden vaak tot wijzigingen in het materiaal dat met de thesaurus toegankelijk gemaakt is. Wie een term toevoegt, wil materiaal daarmee indexeren. Wanneer een term uit de thesaurus verwijderd wordt, kan er vaak voor gekozen worden deze term ook te verwijderen uit het materiaal dat geïndexeerd is met die bepaalde term.

4.6 Samenvattend

Om de uniformiteit te bewaken bij het toekennen van termen maken we vaak gebruik van verschillende soorten woordenlijsten. De meest uitgebreide vorm van een woordenlijst is een thesaurus. De thesaurus is een postcoördinatieve informatietaal die zich vooral kenmerkt door het gebruik van verschillende soorten verwijzingen tussen termen. Deze relaties kunnen gelijkwaardig, associatief of hiërarchisch zijn. Bij hiërarchische relaties is het van belang dat deze zuiver zijn. De relaties

kunnen gebruikt worden bij het zoeken naar materiaal. Generiek zoeken en vervangend zoeken zijn de belangrijkste zoekfunctionaliteiten van een thesaurus. Bij de bouw van een thesaurus zijn vijf stappen te onderscheiden waarbij met name termredactie en het bepalen van de relaties van grote invloed zijn op de kwaliteit van een thesaurus.

Gebruikte literatuur

Becker, P., van der Linden, M., Magrijn, H. & Sieverts, E. (2010). *Organiseer je informatie; aan de slag met thesauri, taxonomieën, tags en topics*. Leidschendam: Biblion. ISBN 9789054839545

Nederlandse Norm - NEN-ISO 25964-1(en) Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval (ISO 25964-1:2011,IDT). September 2011

The ISO25964 Data Model for the structure of an Information Retrieval Thesaurus / Leonard Will from: Bulletin for the American Society for Information Science and Technology - April/May 2012. - Vol 38, nr 4 pag 48-51

KADER

ISO 25964

ISO, de internationale organisatie voor standaardisering is een wereldwijde federatie van nationale organisaties op het gebied van standaardisering. In Nederland is dat het NEN (Nederlandse Norm). Samen maken ze verschillende standaarden, waar de standaard op het gebied van thesaurusbouw er één van is.

ISO 2788 was lange tijd de enige standaard voor het maken van een thesaurus, maar deze dateerde oorspronkelijk uit 1974, met een revisie in 1986. Wel verscheen in 1985 de ISO 5964 met richtlijnen voor het maken en onderhouden van een meertalige thesaurus.

Door de opkomst van internet en intranet leek een thesaurus overbodig te worden. Taxonomieën waren het nieuwe toverwoord om informatie terug te vinden en bovendien was het mogelijk om met behulp van statistische methoden de meest relevante documenten bovenin het zoekresultaat te sorteren. Toch bleek dat niet voldoende en is het toegankelijk maken door middel van thesauri nog steeds van belang. De belangrijkste redenen zijn:

- Niet alle materiaal is als volledige tekst beschikbaar en daardoor kan de tekst niet dienen als bron voor het toegankelijk maken. Denk maar aan alle (niet-digitaal beschikbare) boekcollecties in de verschillende mediatheken en documentatiecentra.
- Beeld- en geluidmateriaal is niet goed toegankelijk te maken zonder aanvullende termen.
- Taal vormt soms een barrière waardoor het zoekresultaat veel minder is dan het zou kunnen zijn. Thesaurustermen kun je toekennen aan documenten in verschillende talen.
- In een meer gesloten omgeving als een intranet werkt *page-ranking* niet optimaal en zijn er aanvullende methoden nodig voor het toegankelijk maken van materiaal.

Daarom is er besloten een nieuwe ISO-norm te maken die in twee delen is gepubliceerd. Deze standaard vervangt de ISO 2788 en ISO 5964. Het eerste deel verscheen in 2011 (ISO 25964-1) en is bedoeld om een thesaurus te maken. Het tweede deel (ISO 25964-2) richt zich met name op de interoperabiliteit; de mogelijkheid om de thesaurus te gebruiken als middel om verschillende autonome systemen met elkaar te laten communiceren. Daarbij vormt de thesaurus, vaak achter de schermen en onzichtbaar voor de eindgebruiker, een waardevolle aanvulling op het zoekgedrag en de zoekmogelijkheden van een eindgebruiker.

Verder lezen

Nederlandse norm NEN_ISO 25964-1 (en) Information and documentation
- Thesauri and interoperability with other vocabularies - Part1: thesauri for
information retrieval (ISO 25965-1:2011,IDT)

KADER

Checklist opstellen randvoorwaarden thesaurus

Voordat daadwerkelijk begonnen kan worden met het ontwerpen van (autorisatie)lijsten of een thesaurus, moet men rekening houden met verschillende aspecten die van invloed kunnen zijn op de inhoud van de (autorisatie)lijsten of een taxonomie. Onderstaande lijst is een korte opsomming van deze aspecten. Door deze vragen/aandachtspunten voor de ontwikkeling van de thesaurus te beantwoorden, kunnen de kaders en randvoorwaarden van de thesaurus in kaart worden gebracht. Zo wordt gekeken wat er nodig is, voor wie en op welke manier.

De lijst is grofweg verdeeld over vier categorieën: gebruikers en (andere) stakeholders, systemen en technologie, de content die men toegankelijk wil maken en het materiaal dat al bestaat. Al deze punten kunnen tussen organisaties, maar vaak ook binnen een organisatie enorm verschillen.

1. Gebruikers en stakeholders:

- Wie zijn de gebruikers en hoe zoeken deze gebruikers? Bekijk zoekstatistieken, inventariseer gebruikerswensen, stel eventueel persona's op.
- Wie gaat de thesaurus beheren?
- Wie gaat trefwoorden toekennen?
- Hoeveel fte en budget is hiervoor nodig en welke expertise en is die binnen de organisatie beschikbaar
- Wie zijn de stakeholders en wat is hun belang en doel?

2. Systemen en technologie:

- Welk(e) systeem of systemen, nu en in de toekomst, moet(en) gebruik maken van de thesaurus?
- Welke functionele toepassingen zijn er voor de thesaurus voorzien?
- Beschikt het systeem over een taxonomie/thesaurus managementtool van voldoende kwaliteit?

3. Content:

- Welke content moet toegankelijk gemaakt worden?
- Welke globale thema's en onderwerpen betreffen het?

4. Bestaande woordenlijsten:

- Wat is er binnen de organisatie beschikbaar aan woordenlijsten, maar ook aan netwerkschijven, lijsten uit databases etc.?
- Zijn er buiten de organisatie woordenlijsten, taxonomieën, thesauri beschikbaar die wellicht herbruikbaar zijn?

Hoofdstuk 5

Eigennamen in thesauri en autorisatielijsten

In paragraaf 4.1 is al gemeld dat termen op basis van individuele begrippen meestal niet in een thesaurus worden opgenomen. Het opnemen van allerlei eigennamen leidt vaak tot een te grote thesaurus of termenlijst. Toch kan er behoefte zijn aan het opnemen van namen zoals die van personen, organisaties, landen, gebeurtenissen enzovoort, omdat gebruikers op deze namen willen kunnen zoeken. In dit hoofdstuk gaan we dieper in op de mogelijkheden en de richtlijnen voor het gebruik van (eigen)namen.

Voorbeelden van dit soort namen zijn:

- groepen van personen, zoals afdelingen of projecten (Personeelszaken, project verbouwing);
- instellingen/organisaties, waaronder klanten, concurrenten (Albert Heijn);
- producten en merken (Apple);
- afzonderlijke objecten, bouwwerken, vervoermiddelen, literaire werken (Abdij van Egmond, Eiffeltoren);
- namen van tijdperken, tijdvakken en namen van gebeurtenissen, zoals life events (20e eeuw, Tachtigjarige oorlog, verhuizing, inhuldiging Koning Willem-Alexander);
- geografische aanduidingen, zoals landen, gebieden, gemeenten, maar ook kantoorvestigingen (Amersfoort);
- persoonsnamen, waaronder medewerkers, externe relaties (Mark Rutte, Sanne Wevers).

Daarnaast zijn er namen die specifiek voor een sector zijn. Zo zal een onderwijsinstelling vaak lijsten hebben van opleidingen en vakken en heeft een zorgverzekeraar behoefte aan een lijst van behandelingen die al dan niet vergoed worden.

Dit hoofdstuk is als volgt opgebouwd. Eerst zullen we de algemene regels geven voor het opnemen van namen (paragraaf 5.1 - 5.4). Daarna bespreken we een aantal uitzonderingen en we besluiten met het beschrijven van persoonsnamen. Die hebben als verschil met de overige namen dat het daar gaat om mensen met een voornaam en een achternaam.

5.1 Toegankelijk maken van (eigen)namen

Zoals gezegd hebben gebruikers soms behoefte aan het zoeken op eigennamen. Soms gaat het om enkele namen die noodzakelijk zijn, zoals een overzicht van ministeries. Eventueel kun je een enkele eigenaam toelaten in een thesaurus, maar dat is alleen verstandig als het om een paar namen gaat. Als het om veel namen gaat, kan gekozen worden voor het opnemen van de soortnaam als term:

Autocoureurs in plaats van Max Verstappen
Winkelketens in plaats van IKEA

Daarmee maak je de precisie van de zoekactie natuurlijk lager. Om dat op te lossen kun je beter een zoekmogelijkheid aanbieden via bijvoorbeeld een apart veld voor die betreffende eigenaam, zoals een persoonsnamen- of geografisch veld. Dan kan er een speciale persoonsnamenthesaurus of geografische thesaurus gemaakt worden. Als zo'n extra thesaurus geen optie is, bijvoorbeeld omdat het gaat om een beperkt aantal namen of omdat het financieel niet haalbaar is, dan is het zinvol om enige controle (met name ten aanzien van de termkeuze en termredactie) op de presentatie van de termen uit te voeren. Deze controle wordt ook wel autorisatie genoemd (zie ook paragraaf 4.1). In dat laatste geval ontstaat ook een lijst van te gebruiken termen (die we aanduiden als autorisatielijst), maar ontbreekt daarbij meestal het bij de thesaurus kenmerkende verwijspatroon.

autorisatie

amsterdam

search

show on map

advanced search

all countries

1005 records found for "amsterdam"

Name	Country	Feature class	Latitude	Longitude
1 Amsterdam AMS, Amsterdam, Amesterdao, Amesterdão, Amistardam, Amistardam, Amistdam, Amsteldamum, Amsterdam, Amsterdamm...	Netherlands, North Holland Gemeente Amsterdam	capital of a political entity population 741,636	N 52° 22' 26"	E 4° 53' 22"
2 New Amsterdam Gord J N-J-Amstehdam, Nijaguis Amsterdams, New Amsterdam, Nieuva Amsterdam, Nueva Ámsterdam, QSK, nyamsa...	Guyana, East Berbice-Corentyne	seat of a first-order administrative division population 35,039	N 6° 14' 52"	W 57° 31' 1"
3 Nieuw Amsterdam Fort Amsterdam, Fort Nieu-Amsterdam, Fort Nieu Amsterdam, Nieuw Amsterdam, Nieuw-Amsterdam	Suriname, Commewijne	seat of a first-order administrative division population 4,935	N 5° 53' 8"	W 55° 5' 19"
4 Île Amsterdam Amsterdam, Amsterdam Island, Amsterdamland, Amsterdemon, Amsterdameya, Amsterdamin, Amsterdameya, Amster...	TF,FR	island	S 37° 50' 0"	E 77° 31' 0"
5 Îles Saint-Paul et Nouvelle-Amsterdam Saint-Paul-et-Amsterdam	French Southern Territories, Saint-Paul-et-Amsterdam	first-order administrative division population 24	S 37° 50' 0"	E 77° 31' 0"
6 Gemeente Amsterdam Amsterdam, Gemeente Amsterdam	Netherlands, North Holland Gemeente Amsterdam	second-order administrative division population 810,935	N 52° 22' 22"	E 4° 53' 54"
7 Amsterdam Airport Schiphol AMS, Aeroport d'Amsterdam-Schiphol, Aeroport «Schiphol», Aeropuerto de Amsterdam-Schiphol, Aeropuerto de...	Netherlands, North Holland Haarlemmermeer	airport	N 52° 18' 37"	E 4° 45' 37"
8 New Amsterdam Airport QSK, SYNA	Guyana, East Berbice-Corentyne	airport	N 6° 15' 0"	W 57° 31' 0"
9 Amsterdam-Zuidoost	Netherlands, North Holland Gemeente Amsterdam	section of populated place population 84,811	N 52° 18' 27"	E 4° 58' 20"
10 Amsterdam Arena Amsterdam Arena, Amsterdam Arena, Amsterdam arena, Amsterdam Arena, Amsterdam Arina, San van gong Amster...	Netherlands, North Holland Gemeente Amsterdam	stadium	N 52° 18' 51"	E 4° 56' 31"
11 Amsterdam Airport Schiphol Aerodrom Amsterdam, Aerodrom Schiphol, Aerodrom Schipkhol, Aeroport d'Amsterdam-Schiphol, Aeroport d'Amst...	Netherlands, North Holland Haarlemmermeer	populated place population 100	N 52° 18' 0"	E 4° 45' 0"
12 Science Park Amsterdam	Netherlands, North Holland	park	N 52° 21' 18"	E 4° 57' 21"
13 Rijksmuseum Amsterdam Arc-museum na kùstleite, Arc-museum na kùstleite, Derchavijn muzej, Drazavni muzej u Amsterdamu, Drazhav...	Netherlands, North Holland Gemeente Amsterdam	museum	N 52° 21' 36"	E 4° 53' 7"

Afb. 5.1 - <http://www.geonames.org/search.html?q=amsterdam&country=> (GeoNames, 2016).

Autorisatie begint met het maken van afspraken. Zo kun je bijvoorbeeld afspreken dat je niet de in Nederland gebruikelijke schrijfwijze hanteert, maar de internationale schrijfwijze voor steden gebruikt (Beijing in plaats

van Peking) en personen beschrijft met: achternaam, voornaam. Dit laatste kun je vastleggen in een afsprakenprofiel (Engels: *application profile*).

Voor geografische namen is het mogelijk om gebruik te maken van de GeoNames database. Hierin zitten inmiddels zo'n 11.000.000 unieke namen van steden, gebouwen en landen. Afbeelding 5.1 hierboven laat een voorbeeld zien van een zoekactie op Amsterdam.

5.2 Verwijzingen

Bij autorisatielijsten of thesauri die geconstrueerd zijn op basis van eigennamen kunnen in principe ook gelijkwaardigheids-, hiërarchische en/of associatieve relaties voorkomen (zie paragraaf 4.5, stap 4). Als de autorisatielijst daartoe mogelijkheden biedt, kunnen deze relaties op dezelfde wijze worden uitgewerkt zoals we in de betreffende paragraaf beschreven hebben. Verwijzingen kunnen worden gemaakt van niet-gebruikte namen en niet-gebruikte naamvormen naar de zogenoemde voorkeursnaam. Dit wordt alleen gedaan wanneer de verwachting bestaat dat gebruikers zoeken op die andere namen en naamvormen. Zo kan er bijvoorbeeld een verwijzing gemaakt worden van 'Boerenleenbank' naar 'Rabobank'.

In een aantal gevallen, maar zeker bij namen van corporaties en geografische gebieden, kan een hiërarchische relatie te herkennen zijn en kunnen NT en BT verwijzingen worden aangebracht. Zoals:

Europa

NT Duitsland
 Frankrijk

Nederland

NT Noord-Brabant
 Limburg
 Zuid-Holland

Ook kunnen verwijzingen op basis van associatieve relaties worden aangebracht. Denk bijvoorbeeld aan een verwijzing tussen oude en nieuwe namen (MKB Brandstof veranderde zijn naam in Tom). Ook een relatie tussen het bedrijf en de merken is mogelijk, zoals Unilever waar het merk Ben & Jerry's onder valt.

5.3 Algemene richtlijnen voor het opnemen van namen

Wie namen op wil nemen in een lijst, moet als eerste besluiten wat bij elkaar hoort. In het algemeen proberen we lijsten zuiver te houden, dat

wil zeggen dat we een lijst maken van één soort, bijvoorbeeld een lijst van organisaties, van afdelingen, van projecten en niet één grote lijst waar al deze zaken door elkaar staan.

In verschillende regelwerken (zie bijvoorbeeld NBBI, 1989) zijn binnen de formele ontsluiting veel afspraken gemaakt over de schrijfwijze van corporatienamen. Ten behoeve van de constructie van termen voor een autorisatielijst kan de reglementering eenvoudiger, dat wil zeggen zoveel mogelijk aansluitend bij de meest bekende naam en naamvorm en wel in natuurlijk zinsverband.

Om de schrijfwijze van een eigennaam als zoekterm in een meer of minder geautoriseerde lijst te bepalen, kunnen de regels van termredactie, termkeuze en termdefinitie worden gehanteerd (zie paragraaf 4.5). In de volgende paragrafen geven we van een aantal soorten eigennamen aan op welke wijze ze in een geautoriseerde lijst voor kunnen komen. Dat wil zeggen dat we hier niet alle mogelijke soorten auteurs en organisaties zullen behandelen, maar eerder een richtlijn geven hoe je tot de te gebruiken termen kunt komen.

De belangrijkste uitgangspunten voor de naam en de schrijfwijze van namen zijn:

1. Kies voor de naam en de schrijfwijze die in Nederland in het algemeen bekend is.

niet: TU/e
maar: Technische Universiteit Eindhoven

2. Lidwoorden aan het begin van de corporatie worden weggelaten, tenzij deze voor de duidelijkheid of grammaticaal nodig zijn.

De Nederlandse Bank
Het Klokhuis

3. Toevoegingen aan een naam om bijvoorbeeld de rechtsvorm aan te geven, worden zoveel mogelijk weggelaten.

niet: Philips Nederland B.V.
maar: Philips Nederland

4. In het geval van een naamsverandering wordt de nieuwe naam gebruikt. De documenten die in het systeem zitten met de oude naam blijven in principe onder de oude naam terug te vinden. In naamsthesauri kunnen de andere namen als RT worden aangegeven en in de *scope note* kan dat toegelicht worden.

Bussum

SN: Per 1 januari 2016 samen met Muiden en Naarden
samengevoegd tot gemeente Gooise Meren

RT: Muiden

RT: Naarden

RT: Gooise Meren

Gooise Meren

SN: Sedert 1 januari 2016, ontstaan uit samenvoeging van
Bussum, Muiden en Naarden

RT: Bussum

RT: Muiden

RT: Naarden

5.4 Vaker voorkomende namen

Als de naam vaker voorkomt, maak dan met behulp van een kwalificatie onderscheid tussen dezelfde namen. Dat kan middels het toevoegen van een geografische bepaling zoals het eerdergenoemde GeoNames, of door een andere verduidelijking.

Den Haag (gemeente)

Den Haag (vestiging)

Eva Janssen, hoorspelactrice

Eva Janssen, triatlete

Amsterdam Arena (hotel)

Amsterdam Arena (stadion)

5.5 Geografische namen

Geografische namen zijn namen van geografische eenheden, zoals (gedeelten van) werelddelen, landen, gemeenten, provincies, departementen, steden, dorpen, wijken, straten, pleinen, havens, streken, eilanden, zeeën, rivieren, bergen, woestijnen, et cetera. Daarnaast kunnen organisaties ook specifieke geografische aanduidingen hebben, bijvoorbeeld kantoorlocaties of afzetgebied.

De belangrijkste uitgangspunten voor de naam en de schrijfwijze van geografische namen zijn:

1. Maak gebruik van de officiële naam.
Voor buitenlandse geografische benamingen wordt de naam gekozen, waarmee zij in het Nederlands in het algemeen voorkomen.

Londen	(i.p.v. London)
Den Haag	(i.p.v. 's-Gravenhage)

2. Handhaaf natuurlijk zinsverband bij namen die opgebouwd zijn uit twee of meer woorden.

Zuidoost Azië
Golf van Biskaje

3. Een geografische naam begint niet met een lidwoord of ander voorvoegsel, tenzij dit met een daarop volgend(e) woord/woordcombinatie een onverbrekelijk geheel vormt.

Wel:	Den Helder	Niet:	Het Gaasterland
	Het Gooi		De Achterhoek
	's-Heerenberg		

4. Gebruik een kwalificatie als dezelfde benaming voor verschillende eenheden wordt gebruikt.

Jura (Frankrijk)
Jura (Zwitserland)
Utrecht (Gemeente)
Utrecht (Provincie)
Utrecht (Bisdom)

De toevoeging 'Nederland' wordt zoveel mogelijk weggelaten

5. Wanneer een geografische eenheid van naam verandert door politieke of andere oorzaken, wordt in de term de naam gebruikt, die relevant is voor het betreffende onderwerp en/of de tijd waarin het onderwerp is geplaatst.

Leningrad / Sint Petersburg
Ayers Rock / Uluru
Birma / Myanmar

In de praktijk wordt dikwijls het één door het ander vervangen. Een verwijzing (bij geografische thesauri) van de niet-voorkeursterm naar de voorkeursterm ligt dan voor de hand.

5.6 Namen van tijdvakken, tijdperken en gebeurtenissen

tijdvak

Een tijdvak is een exacte aanduiding in de tijd, zoals eeuwaanduidingen en jaartallen.

19e eeuw
1830-1839

Bij een tijdvak kun je kiezen voor een schrijfwijze in voornamelijk woorden of voornamelijk cijfers.

19e eeuw of 1800-1899
1830-1839
1997
2000

De keuze van de termen is hierbij natuurlijk heel sterk afhankelijk van de specificiteit van de termenlijst.

tijdperk

Een (geologisch) tijdperk is een periode met eigennaam. Vaak is het begin en/of einde niet exact aan te geven.

Middeleeuwen
Tachtigjarige oorlog
Cambrium

Het is in deze gevallen niet noodzakelijk en niet gewenst de tijdspanne er met behulp van een kwalificatie bij te plaatsen. De gekozen naam benoemt immers impliciet ook de periode. Wel kan een toelichting relevant zijn.

Middeleeuwen
Scope note: tijdsaanduiding tussen 500 en 1500

Onder gebeurtenissen worden verstaan: congressen, conferenties, concilies, festivals, sportmanifestaties, herdenkingen, tentoonstellingen, oorlogen, veldslagen, zeeslagen, vredes- en andere historische gebeurtenissen enzovoort.

Bij gebeurtenissen kiest men de naam waaronder de betreffende gebeurtenis in Nederland het meest bekend is. Indien nodig kan de plaats van handeling en/of jaartal als kwalificatie toegevoegd worden. Het natuurlijk zinsverband wordt vanzelfsprekend zoveel mogelijk gehandhaafd.

Olympische Zomerspelen (Rio, 2016)
Floriade (Amsterdam, 1972)

Vrede van Münster
 Noorderslag (2016)
 Dolle dinsdag
 Ronde van Spanje

5.7 Persoonsnamen

Een persoonsnaam (persoonlijke eigennaam) is een naam van een natuurlijk persoon. Een persoonsnaam kan zijn: een voornaam, een familienaam (achternaam), een pseudoniem, een bijnaam of een aangenomen naam.

Hadewych
 Max Verstappen
 Anton Jansen
 Rutger Kopland
 El Greco
 Johannes Paulus II
 Battus

Belangrijke uitgangspunten om te komen tot elke vorm van autorisatie zijn:

1. Inversie

De achternaam is het belangrijkste zoekargument.

Verstappen, Max in plaats van Max Verstappen
 Vondel, Joost van den in plaats van Joost van den Vondel

2. Meest bekende naamsvariant

Als een persoon onder verschillende namen of naamsvormen voorkomt kies je voor de schrijfwijze waaronder hij of zij het meest bekend is. Dat kan zich voordoen bij een minimaal verschil in schrijfwijze, zoals:

Richards, Keith (meest bekend-voorkeur)
 Richard, Keith

maar ook in gevallen waarbij bijvoorbeeld de echte naam nauwelijks bekend is:

Wyman, Bill (meest bekend-voorkeur)
 Perks, William

Drs. P (meest bekend-voorkeur) in plaats van Heinz Hermann Polzer

3. Kwalificatie bij identieke namen

Denk hierbij aan een kwalificatie zoals beroep, predikaten als junior en senior, eventueel geboorte- en sterfjaar. De keuze valt dan op een kwalificatie die in dat specifieke geval het meest significant is voor de gebruiker.

Delden, Lex van (componist)

Delden, Lex van (toneelspeler)

4. Bij buitenlandse namen: gangbare schrijfwijze

De laatste jaren is dat verschoven van het gebruik van de Nederlandse variant naar de schrijfwijze in de oorspronkelijke taal.

Ludwig II in plaats van Lodewijk II

In dit geval is er ook geen verwarring meer mogelijk met Lodewijk II van Frankrijk (Louis II).

Bij geografische namen ligt dat anders, daar wordt over het algemeen de Nederlandse schrijfwijze gehanteerd.

Londen in plaats van London

Milaan in plaats van Milano

Autorisatie in deze vorm is niet alleen belangrijk bij het inhoudelijk toegankelijk maken van informatie, maar juist ook op het gebied van de formele ontsluiting. Dat geldt met name voor personen en organisaties in de functie van auteur. Op dat aspect is een aantal bruikbare werken verschenen, waaronder de Richtlijnen Persoonsnamenthesaurus, opgesteld door PICA (OCLC, 2009). Deze richtlijnen lopen grotendeels in de pas met Regels voor de titelbeschrijving, in dit geval met name met deel 3 van deze Regels: Catalogusbouw (afgekort RT/cat) (FOBID, 1994). Op basis van deze Richtlijnen is voor Nederland een Persoonsnamenthesaurus opgesteld die gebruikt wordt binnen het GGC-systeem van OCLC (voorheen PICA). Deze bevatte in 2000 tussen de 1,5 en 2 miljoen naamvormen.

Inmiddels is er ook een Virtual International Authority File (VIAF) beschikbaar, met bijna 10 miljoen naamvormen van personen, waarin OCLC-autorisatielijsten uit ongeveer 15 landen (maar niet uit Nederland) zijn opgenomen (OCLC, 2010).

5.8 Samenvattend

Er zijn situaties waarin gebruikers willen zoeken op namen van personen, instanties, landen, tijdperken enzovoort. Deze worden over het algemeen echter niet in een thesaurus opgenomen omdat die daarmee te uitgebreid zou worden. We kunnen in dat geval aparte zoekingen creëren op basis van geautoriseerde namenlijsten. Het is belangrijk dat deze lijsten consequent zijn opgesteld, dat wil zeggen dat alle namen op dezelfde wijze worden behandeld. In dit hoofdstuk hebben we een aantal regels besproken die bij het opstellen van namenlijsten kunnen worden toegepast. De meeste van deze regels zijn gebaseerd op internationale standaarden. Het staat de bouwer van een systeem natuurlijk vrij om van deze regels af te wijken, maar het is wel van belang dat de uitzondering dan weer consequent wordt toegepast.

Gebruikte literatuur

FOBID (1994). *Regels voor de titelbeschrijving. Deel3: regels voor Catalogusbouw*. Den Haag: NBLC

GeoNames (2016). Geraadpleegd op: 04-07-2016, van <http://www.geonames.org>

NBBI (1989). *ABC-regels: handleiding voor het alfabetisch rangschikken*. Alphen aan den Rijn: Samsom

OCLC (2009). *Richtlijnen voor de aanlevering van gegevens; Richtlijnen voor het catalogiseren van alle materialen inclusief thesauri*. Geraadpleegd op: 22-02-2010, van: <http://www.oclc.org/?id=1042&ln=nl>

OCLC (2010). *VIAF: the virtual international authority file*. Geraadpleegd op: 22-02-2010, van: <http://www.oclc.org/research/activities/viaf/>

Hoofdstuk 6

Digitale en fysieke ordening

In hoofdstuk 3 is al beschreven hoe we als mensen de neiging hebben de wereld om ons heen te ordenen. Het gebruik van categorieën is eigen aan de menselijke geest. Het in een hiërarchie onderverdelen van die categorieën maakt daar vrijwel altijd deel van uit. Dat zien we al wanneer we boodschappen doen. In een supermarkt kun je de meeste producten gemakkelijk bij elkaar in de buurt vinden: de appels bij het fruit en de melk bij de andere zuivelproducten. In hoofdstuk 3 beschreven we ook hoe producten en documenten binnen een webomgeving aan de hand van vier stappen geordend konden worden. In dit hoofdstuk kijken we naar hiërarchische schema's die we hiervoor gebruiken.



Afb. 6.1 - De supermarkt.

Dit doen we door verder in te gaan op het begrip navigational taxonomy of directe toegankelijkheid. Het begrip directe toegankelijkheid is al aan de orde gekomen in hoofdstuk 3: de taxonomie is een op zichzelf staand systeem voor de ordening en directe toegankelijkheid van de documenten. Op basis van de gebruikte ordening kan de gebruiker gericht zoeken in een kast of in een bestand. Wat we in dit hoofdstuk toevoegen is de nadruk op fysieke plaatsing van producten of documenten. De informatie waarin we zoeken, is op een logische manier geordend, zodat we zonder hulp van bijvoorbeeld een zoekmachine snel bij de juiste informatie komen.

In het eerste deel van dit hoofdstuk kijken we naar het plaatsen van bestanden in mappen op een computer of digitale schijf. Bij het opbergen en zoeken van digitale bestanden binnen een mappenstructuur op een computer geldt dat het voor sommige bestanden helder is waar je ze plaatst, terwijl dat bij andere wat minder vanzelfsprekend is. Als we bijvoorbeeld op zoek zijn naar de notulen van een vergadering, zoeken we gericht in een map 'notulen'. Die map kan ergens in een gemeenschappelijk schijfgebied staan, maar ook ergens in de cloud zoals in Dropbox of Onedrive. We plaatsen deze bestanden doorgaans op basis van een gekozen ordening. Dat lijkt op de bovengenoemde supermarkt. In beide gevallen geldt ook dat sommige producten of bestanden gemakkelijk te vinden zijn, terwijl andere producten lastiger zijn. Want waar staan de hoestbonbons? Bij het snoep of bij de medicijnen?

Een Word-document kan over een bepaald onderwerp gaan, heeft een datum en is door iemand geschreven. Er zijn dus meer mogelijkheden om dat document op te bergen: op onderwerp, op datum of op persoon. En natuurlijk sla je een document maar één keer op, en het liefst daar waar het ook gezocht wordt. Dat maakt het noodzakelijk om een soort werkwijze of regels te maken waardoor documenten waarvan de plaatsing minder vanzelfsprekend is, toch ook op één plek terechtkomen. In hoofdstuk 3 hebben we gesproken over vier stappen om tot een ordening te komen. In dit hoofdstuk breiden we dat aantal stappen uit: in eerste instantie naar zeven en later naar tien. Dat is noodzakelijk om digitale of fysieke documenten terug te kunnen vinden. Het volgen van deze stappen van ordening wordt ook wel het bouwen van een classificatie genoemd. Een classificatie is een schema van voorgedefinieerde categorieën, waarin zaken die bij elkaar horen bij elkaar in de buurt staan en zaken die van elkaar verschillen, gescheiden zijn.

We gaan hierbij uit van klassieke inzichten die we vertalen naar en toepassen in een moderne, voornamelijk digitale omgeving. De ordening van documenten of bestanden laten we hierbij in zeven stappen verlopen.

6.1 Digitale ordening

Onder digitale ordening verstaan we in het kader van dit boek: het plaatsen van een specifiek document op één digitale plek. Die plaats is dan bijvoorbeeld de mappenstructuur op een gemeenschappelijk schijfgebied of een externe schijf of server. De documenten of bestanden waar het om gaat wil je op één plaats zetten omdat het tijd scheelt om ze in te delen zodat je er niet steeds naar moet zoeken. Dat geldt met name voor bestanden die verschillende mensen plaatsen en gebruiken. Zelf weet je vaak nog ongeveer welke woorden je gebruikt hebt, zodat je ze daarmee wel weer kunt terugvinden. Maar als een ander de notulen van een

projectvergadering maakt en opslaat, waar worden deze dan opgeslagen? Deze kunnen staan onder ‘verslagen’ of ‘notulen’, onder de naam van het project of ‘vergadering’ of zijn ze alleen op datum terug te vinden?

Digitale ordening begint met de al eerder genoemde ordening die gebruikt werd in hoofdstuk 3. De eerste stappen zullen we daarom hier nog maar summier beschrijven, omdat de nadruk ligt op de vervolgstappen. Die vervolgstappen moeten ervoor zorgen dat het document maar op één plek in een mappenstructuur wordt opgeborgen. De structuur moet als het ware de weg wijzen, zodat de documenten vanzelf hun plek krijgen.

Als voorbeeld nemen we een klein bedrijf, illustratief voor allerlei kleine bedrijfjes waarin één of meer zzp’ers samenwerken, vaak vanaf verschillende locaties. Zij delen hun bestanden op een gemeenschappelijke plaats. Het bedrijf van dit voorbeeld biedt allerlei ondersteunende diensten aan voor grotere bedrijven, variërend van informatiemanagement en contentmanagement-klassen tot administratief/financiële diensten.

6.1.1 Doel, begrippen, homogene groepen en collocatievolgorde

Stap1: Doel

In het geval van dit kleine bedrijf, voor de gelegenheid gedoopt tot ZZP Company, gaan de mappen een zodanige structuur krijgen, dat de medewerkers hun spullen makkelijk kunnen opslaan en ook weer terugvinden.

Stap 2: Verzamelen van begrippen

Alle zzp’ers maken gebruik van een gezamenlijke schijf voor verslaglegging, projectplannen, urendeclaraties, facturen en evaluaties. Het merendeel daarvan heeft betrekking op specifieke klanten.

Stap 3: Maak homogene groepen

Om het voorbeeld wat beheersbaar te houden noemen we een paar belangrijke groepen met daarin wat mogelijke begrippen:

- Producten (workshops, trainingen);
- Jaren (2015, 2016, 2017);
- Vorm (offertes, contracten, notulen, rapporten);
- Klanten (Kant A, Klant B).

Het in het kader besproken *cardsorten* kan ook hier behulpzaam zijn. Aan de hand van de daar gemaakte keuzes kun je ook zien welke onderwerpen “bij elkaar horen”.

Stap 4: Collocatievolgorde

In dit eenvoudige voorbeeld speelt de collocatievolgorde vooral een belangrijke rol bij *tijd*. Daar zullen we kiezen voor een chronologische volgorde. Voor de andere groepen is die volgorde minder belangrijk, maar we moeten er wel een kiezen. Bij *vorm* en *producten* kiezen we waarschijnlijk voor een pragmatische volgorde, want “waar ik het vaakst gebruik van maak of het meest van heb, wil ik bovenaan zien”. Bij *klanten* kiezen we waarschijnlijk voor het alfabet. Overigens realiseren we ons wel dat de verkenner van het besturingssysteem hierin niet altijd meewerkt: de mappenvolgorde kun je niet volledig aanpassen. Dat vereist soms het inventief benoemen of nummeren van mapnamen.

Tot dusverre is het voorbeeld nog gelijk aan de stappen en activiteiten zoals die in hoofdstuk 3 aan de orde zijn gekomen ten aanzien van structurering van websites. Nu is het van belang te bepalen in welke map je de materialen opbergt, zodat je ze later weer kunt vinden. Daarvoor zijn de volgende twee stappen nodig: het vastleggen van de combinatie- en de schemavolgorde.

6.1.2 Stap 5: Combinatievolgorde

Op grond van de voorgaande stappen zijn groepen mappen gemaakt, maar die groepen zijn nog niet geordend. En, belangrijker nog, ze zijn niet met elkaar in combinatie gebracht. Zo is het nog niet duidelijk waar de offerte van klantB uit 2016 terecht moet komen. In de map ‘offertes’, in de map van de klant of op het jaar?

Een belangrijk verschil met de ordening binnen een website is dat de weg naar de informatie daar altijd door middel van hyperlinks, vanuit elke willekeurige invalshoek getoond kan worden. Zodra je documenten in een digitale mappenstructuur gaat opbergen, loop je er tegenaan dat die maar op één plaats weggezet kunnen worden. Dat houdt in dat er een algemeen besluit moet komen welke plaats de beste is. Daarom is nu een andere strategie nodig. Daarbij maakt het wel degelijk uit welk pad je kiest; het ene pad leidt namelijk wel tot de gewenste informatie en het andere pad doet dat niet. Het juiste pad moet dan wel logisch en consequent gekozen zijn. In ons voorbeeld zagen we dat een onderwerp meer invalshoeken kan hebben. In deze stap bepalen we welke invalshoek de belangrijkste is, en vervolgens welke iets minder belangrijk is, tot de minst belangrijke aan toe.

combinatievolg- orde

Deze keuze noemen we de combinatievolgorde. Deze volgorde van elementen die loopt van bijzonder naar algemeen, leg je ergens vast om een eenheid van klasseren te bereiken. Want wanneer de combinatievolgorde niet steeds dezelfde is, dan zal bij het klasseren van informatie nu eens het ene onderwerpelement en dan weer het andere als primair gezien worden. Het gevolg daarvan is dat je er als zoeker niet meer van verzekerd bent dat in een bepaalde map alle informatie over een onderwerp verzameld is. Dat is het geval als er zowel onder de

**cross-
classification**

klantenmappen een onderverdeling naar jaren is (KlantA-2015) als onder de jarenmappen een onderverdeling naar klant (2015-KlantA) en op beide plaatsen een deel van de documenten terechtkomt. Wanneer gelijksoortig materiaal toch in verschillende klassen terechtkomt, noemen we dit *cross-classification*. Een ander gevolg van het niet consequent aanhouden van de combinatievolgorde zou bijvoorbeeld zijn dat de trainingen voor KlantA onder 'trainingen' gezocht moeten worden, maar de workshops voor KlantB onder 'KlantB'.

collocatie

Terug naar het voorbeeld: een offerte voor Klant B uit 2016 heeft de drie aspecten offerte, klant en jaar. Bij het aanmaken van de ordening zal bepaald moeten worden wat het meest specifieke aspect is. In dit geval is dat de klant. Dat wil zeggen dat je alle informatie over klanten bij elkaar ziet staan. Dit noemen we collocatie. Tegelijkertijd staat de informatie over offertes van klanten verspreid over de verschillende klanten. Dat noemen we allocatie. Dit is een onvermijdelijk gevolg van gemaakte keuzes.

allocatie

De regels van de combinatievolgorde zijn ooit opgesteld voor de plaatsing van boeken en ook in die omgeving vaak toegepast naar eigen behoefte. Dat geldt ook in de digitale wereld. Het is vooral van belang om erachter te komen wat in de eigen situatie als meest specifieke element wordt gezien. Zeker daar waar documenten met elkaar gedeeld worden op een gezamenlijk schijfgebied, is het zaak dat de volgorde helder en werkbaar is voor alle medewerkers die lees- en schrijfrechten hebben op een dergelijk schijfgebied. Pas als een indeling ook consequent wordt toegepast, biedt het zijn meerwaarde. Hieronder werken we een voorbeeld uit vanuit een praktische invalshoek: het theoretische schema staat in paragraaf 6.2.2.

Sommige aspecten worden door iedereen als algemeen ervaren. Zo worden de aspecten *vorm* (informatie aangeboden als encyclopedie, als artikel, in de vorm van een verslag of een offerte), *plaats* (Rotterdam, Nederland of India) en *tijd* (2015, 2016 of voor 2000) vaak zonder probleem als algemeen herkend. Andere aspecten als *processen* (reorganisatie), *onderdelen* en *producten* (workshop, training) zijn meer specifiek.

Als we een volgorde aanbrengen in onze groepen uit het voorbeeld stellen we dat het meest belangrijke, en daarmee meest specifieke, aspect de klanten zijn. Daarvoor werk je immers. Daarna zijn in dit voorbeeld in afnemende volgorde van belang:

- Producten (workshops, trainingen)
- Vorm (offertes, contracten, notulen, rapporten);
- Jaar (2015, 2016, 2017).



Afb. 6.2 - Start mappenstructuur.

In het geval dat het vaststellen van deze volgorde voor problemen zorgt, is het handig om het schema te gaan gebruiken dat staat toegelicht in paragraaf 6.2.2. Hierdoor is het mogelijk om deze volgorde vast te stellen op grond van reeds langer bestaande beginselen.

Nu is de basismappenstructuur vast te stellen. Er komen mappen voor elke klant en daarbinnen (voor zover van toepassing) mappen voor informatie over de verschillende producten. Indien noodzakelijk worden die mappen weer onderverdeeld naar de vorm en indien van toepassing, per jaar. Afhankelijk van de omgeving waarbinnen je ordent, zul je kunnen bepalen



wat daar als meest specifiek gezien moet worden. Het is vaak een intuïtief proces, waarbij binnen een bepaalde omgeving van een bedrijf of branche dezelfde volgorde wordt gekozen. In bovenstaand voorbeeld zijn de klanten het meest specifiek en komen de offertes verspreid te staan, namelijk steeds bij elke klant. Maar een bedrijf dat zich met het verwerken van betalingen bezighoudt, zal liever de offertes en facturen als meest specifiek kenmerk beschouwen, omdat dan alle facturen en offertes van de verschillende

Afb. 6.3 - De mappenstructuur.

klanten bij elkaar in de buurt komen te staan. Is er eenmaal een keuze gemaakt dan zullen we die ook consequent moeten aanhouden.

6.1.3 Stap 6: Bepaal de volgorde van de mappen (schemavolgorde)

Schemavolgorde is de volgorde waarin de onderwerpen in het schema worden aangeboden. In dit geval is dat in welke volgorde de mappen geordend moeten worden. In principe is dat het omgekeerde van de combinatievolgorde. Hiervoor is gesteld dat de combinatievolgorde van bijzonder naar algemeen loopt. De schemavolgorde loopt daarom van algemeen naar bijzonder. Hij is omgekeerd aan de combinatievolgorde omdat daardoor alle onderwerpen altijd een plaats in het schema kunnen krijgen, of ze nu algemeen of meer specifiek zijn.

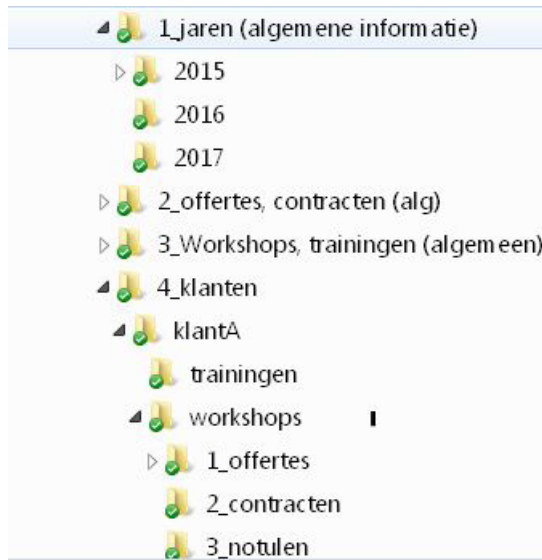
Combinatievolgorde van ZZPcompany	Schemavolgorde van ZZPcompany
Klant	Jaar
Product	Vorm
Vorm	Product
Jaar	Klant

Tabel 6.1 - Schema- en combinatievolgorde als elkaars inverse.

Hoe dat werkt leggen we in het volgende voorbeeld uit: Na enige tijd blijkt de gekozen structuur toch nog voor problemen te zorgen. Een zzp'er wil weten waar de notulen staan van de algemene overleggen die er incidenteel zijn. Hij zoekt een workshopidee dat ooit tijdens zo'n overleg besproken is. Op dat moment blijkt dat er geen plaats is voor het bewaren van algemene notulen. De beheerder geeft aan dat dergelijke ideeën vaak tijdelijk even neergezet worden bij een klant, maar er zou een plaats moeten komen waar meer algemene stukken zoals ideeën voor workshops, maar ook notulen een plaats kunnen krijgen.

Op dit moment is de basismappenstructuur in de verkenner van het bedrijf opgesteld volgens de vastgestelde combinatievolgorde klant - vorm - jaar. Wat nodig is, is een aanpassing van het schema zodat de schemavolgorde wordt gevolgd. En die loopt van algemeen naar specifiek. Dat zorgt er meteen voor dat onderwerpen altijd een plek hebben. Het schema in de vorige afbeelding moet daarvoor dus worden uitgebreid. Voor alle producten waarvan wat meer algemene bestanden bestaan, worden een niveau hoger mappen opgenomen, op het niveau van de

klanten. In de map 'workshops' die dan ontstaat, zitten nu de ideeën voor workshops, naast de workshops die bij meerdere klanten zijn gehouden. Uiteindelijk wil je een bruikbare structuur maken waarin alle informatie een plek heeft.



Afb. 6.4 - Aanpassing van de oorspronkelijke mappenstructuur.

In bovenstaande uitwerking (afbeelding 6.4) staat het schema (de mappen) al uitgeschreven, waarbij de schemavolgorde ook goed te herkennen is: jaar, vorm, product en klanten. Dit is voldoende in deze omgeving. De documenten zijn te vinden door op mappen en submappen te klikken.

Dit voorbeeld klinkt als een gemakkelijk, vanzelfsprekend proces, maar wie er zelf mee aan de slag gaat, merkt dat het toch onverwacht lastig is. Hoe zien de homogene groepen er precies uit? Wat is het meest specifieke element om de combinatievolgorde te bepalen? Alleen een consequente indeling heeft meerwaarde. Anders heb je toch weer heel vaak de zoekfunctie van je desktop nodig, of zelfs een specifiek zoekprogramma (wat overigens helemaal geen onhandig instrument is).

In het voorbeeld hierboven is sprake van een klein bedrijf, maar stel dat dit bedrijf groeit en verschillende afdelingen krijgt, die elk hun eigen informatievoorziening hebben. Zo kan een groter bedrijf een afdeling P&O hebben. Ook zijn er afdelingen die informatie met elkaar moeten delen, zoals bijvoorbeeld een afdeling R&D en de afdeling accountmanagement. Werken met mappen gaat dan niet meer zo vanzelf, doordat vaak ook elke afdeling eigen regels hanteert.

Voor wie meer wil weten over digitale ordening kan hoofdstuk 8 interessant zijn. Als instellingen en bedrijven complexere situaties tegenkomen, gaan ze vaak over tot het gebruik van SharePoint als middel om bestanden en documenten te beheren. Daarom is in dit boek een apart hoofdstuk hieraan gewijd. Dat bouwt voort op de hier gehanteerde principes maar laat ook een manier zien waarop een hiërarchische indeling wordt gebruikt als grofmazige structuur waarbinnen de verfijning plaatsvindt door middel van tags. In de rest van dit hoofdstuk richten we de blik op de fysieke plaatsing van documenten.

6.2 Fysieke plaatsing van documenten

In het begin van dit hoofdstuk stelden we dat er zeven stappen nodig zijn voor het uitwerken van een mappenstructuur en vier voor websites. Is dat niet genoeg? Voor veel hedendaagse ordening is dat inderdaad voldoende, maar niet als het gaat om fysieke plaatsing van documenten, veelal boeken, in een kast. Hoewel steeds meer materiaal digitaal beschikbaar komt, zijn er ook nog steeds fysieke collecties, van hele kleine tot vele kilometers aan materiaal. De gebruikelijke ordening van dergelijke collecties is onderwerpsgericht, want daardoor staat bij elkaar wat bij elkaar in de buurt hoort. Wie een bibliotheek of boekhandel binnenloopt, ziet dat meteen terug: fictie staat bij elkaar op alfabet, jeugdboeken staan bij elkaar en binnen de non-fictie staan ook materialen over hetzelfde onderwerp in elkaars buurt. Boeken over reizen, managementboeken, boeken over kunst en over treinen staan geordend op onderwerp en de onderwerpen die met elkaar verwant zijn, staan bij elkaar in de buurt. In de rest van het hoofdstuk beantwoorden we de volgende vragen: hoe zijn deze materialen geordend en welke stappen zijn daarvoor nodig?

classificatie

Alle hiervoor besproken theorie is in wezen gebaseerd op de klassieke ordeningsmethode van het bouwen van een classificatie. Een classificatie is een systeem waarin begrippen (onderwerpen) naar hun inhoudelijke samenhang op basis van een indelingscriterium geordend zijn in zo homogeen mogelijke groepen (klassen). Die klassen kunnen ook weer, op basis van een ander indelingscriterium, onderverdeeld worden in een aantal (sub)klassen. Zo ontstaat een hiërarchisch opgebouwd systeem dat, wanneer het is vastgelegd en voorzien van codes of notaties, een classificatieschema wordt genoemd.

classificatieschema

Eind 19e eeuw is men begonnen met het maken van de grote universele classificaties vanuit het geloof dat alle kennis van de wereld te ordenen is. Vanuit een heel ander vertrekpunt volgt Google in wezen een soortgelijke gedachte: met zijn zoekmachine wil het bedrijf het liefst alle informatie van de wereld vindbaar maken.

Voordat classificaties hun intrede deden, waren boeken vaak geordend op volgorde van binnenkomst of grootte en kon je als bezoeker niet vrijelijk rondsnoeien of grasduinen in de collectie. In die zin waren classificaties een belangrijke stap in de emancipatie van de gewone mensen in een tijd dat ook scholing steeds meer van belang werd en steeds meer mensen konden lezen en daardoor ook gebruik konden maken van collecties.

De Amerikaan Melville Dewey gaf als het ware het startschot voor het maken van dergelijke grote classificaties met zijn Dewey Decimal Classification (DDC). De DDC is vooral in de Angelsaksische wereld wijdverspreid geraakt. Otlet en La Fontaine volgen met hun Universele Decimale Classificatie (UDC) vanaf 1904. Ook deze wordt in tal van instituten en bibliotheken nog gebruikt. Naast het UDC wordt, met name in openbare bibliotheken, maar ook bijvoorbeeld bij een aantal hogeschoolbibliotheken het SISO gebruikt. SISO staat voor “Schema voor de Indeling van de Systematische catalogus in Openbare bibliotheken”. In hoofdstuk 13 worden deze classificaties uitgebreider besproken.

Behalve de hierboven genoemde universele classificaties bestaan ook specialistische classificaties, die een deel van de wetenschap in kaart brengen.

Ranganathan construeerde in 1933 zijn Colonclassificatie: een classificatie op basis van de PMEST-formule: *personality, matter, energy, space, time*. Uit de begripseenheden die bij deze facetten hoorden zou elk gewenst onderwerp samengesteld kunnen worden. Hij was daarmee de eerste die het analytisch-synthetische principe (zoals verder uitgelegd wordt in paragraaf 6.2.3) heeft beschreven. Deze classificatie was ook universeel en probeerde al het menselijk kennen en kunnen te analyseren in samenstellingen van fundamentele, logische categorieën (facetten). Hoewel deze categorieën in de praktijk te algemeen en vaak ook te theoretisch blijken te zijn, is het principe in veel classificatieschema's toegepast. (Ranganathan, 2010)

Om recht te doen aan bibliografische collecties en om de oorsprong aan te geven van de classificatietheorie volgen we nogmaals alle stappen van 1 t/m 10 met de nadruk op de stappen 7 en 8 omdat daarin de verschillen tussen digitale opslag en fysieke documenten het meest uitgewerkt worden. Tegelijk vormt deze paragraaf ook een samenvatting van hetgeen in “Organiseer je informatie” meer in detail wordt beschreven. Wie aan de slag wil met het bouwen van een classificatie, verwijzen we graag naar dit werk uit 2010.

6.2.1 Stap 1 t/m 4: Doel bepalen, verzamelen van begrippen, homogene groepen en collocatievolgorde

De eerste vier stappen verschillen niet met wat we daarover al eerder hebben geschreven.

Bij de collocatievolgorde hebben we nog wel een aanvulling. Tot nu toe waren vooral de pragmatische en de chronologische volgorde aan bod gekomen. De pragmatische volgorde is in een fysieke omgeving echter moeilijk vast te stellen. Daarom noemen we hier nog een aantal andere principes die als uitgangspunt gebruikt kunnen worden:

- Evolutionair: bijvoorbeeld psychologie: zuigeling - kleuter - adolescent - volwassene – bejaarde;
- Chronologisch: bijvoorbeeld de wijze van studeren: propedeuse - hoofdfase – minor – afstuderen;
- Geografisch: bijvoorbeeld Nederland van noord naar zuid: Groningen - Friesland - Drenthe – enzovoort;
- Alfabetisch: het is geen systematisch beginsel, maar als bovenstaande principes allemaal niet van toepassing zijn, is het een praktisch alternatief.

Wie goed kan bepalen wat een logische volgorde is, helpt daarmee de zoekers, want die doorziet een goed gekozen indeling meteen. Iedereen is ook bekend met de omgekeerde situatie: een slecht gekozen collocatievolgorde kost tijd om te begrijpen en levert soms verkeerde keuzes op.

6.2.2 Stap 5 en 6: combinatievolgorde en schemavolgorde

Bij de digitale ordening in paragraaf 6.1 is al het nodige gezegd over de combinatievolgorde. Hier gaan we daar nog wat preciezer op in, want bibliografische classificaties moeten samengestelde onderwerpen kunnen omvatten. Immers, in veel documenten worden diverse onderwerpen (concepten) juist in hun samenhang besproken. Bij het analyseren van een samengesteld onderwerp moet, rekening houdend met het doel van de classificatie, worden vastgesteld welk onderwerpselement het meest bijzondere, het meest specifieke is; welk onderwerpselement daarna in aanmerking komt; welk daarna et cetera, in afnemende volgorde van belang ten opzichte van het primaire onderwerp.

In het voorbeeld van de mappen zijn we uitgegaan van een wat intuïtieve werkwijze. Maar uit de uitgangspunten die we daar al noemden, is wel een min of meer algemeen geldende combinatievolgorde te halen, ook weer lopend van bijzonder naar algemeen. Die uitgangspunten zijn:

- onderwerp vóór vorm;
- geheel vóór deel;
- doeleinden en eindproducten vóór productiemiddelen;
- meer concrete zaken vóór minder concrete zaken;
- soorten vóór acties/handelingen.

Deze combinatievolgorde is weergegeven in tabel 6-2. Zoals deze tabel laat zien, is de combinatievolgorde in principe altijd van bijzonder naar algemeen.

Het meest bijzondere aspect wordt als het primaire onderwerp gezien. Daarmee volgen we een grondregel bij het toegankelijk maken van informatie: dat informatie geïndexeerd wordt bij het meest specifieke aspect, zodat het bij het zoeken ook op het meest specifieke aspect gevonden kan worden.

Als voorbeeld nemen we een document dat gaat over het assembleren van mobiele telefoons in China. Wie de kenmerken labelt met behulp van de tabel 6-2 krijgt de volgende uitkomst:

- Assemblage (proces)
- Mobiele telefoons (product)
- China (land)

De tabel laat zien dat de mobiele telefoons hier het meest specifiek zijn en dat komt ook wel met ons gevoel overeen. De assemblage geeft een wat algemenere context weer en het landaspect China is het meest algemene aspect. De combinatievolgorde wordt dan dus: mobiele telefoons - assemblage - China. Het document is dan terug te vinden bij het onderwerp 'mobiele telefoons', daar zorgt de volgende stap van de schemavolgorde voor.

Combinatievolgorde (van specifiek naar algemeen)
Producten / soorten
Delen / onderdelen
Materialen
Kenmerken
Processen / handelingen
Bewerkers
Plaats
Tijd
Vorm

Tabel 6.2 - De combinatievolgorde.

We zijn ons ervan bewust dat de aspecten van tabel 6-2 niet altijd duidelijk herkenbaar zijn in de informatie die we willen indelen. Soms is het doel van de classificatie ook zodanig dat van deze volgorde afgeweken moet worden. Dat is niet erg, als de consequentie daarvan maar doorgetrokken wordt in latere stappen van de structurering, ter voorkoming van de eerder genoemde *cross-classification*.

schemavolgorde

Ook stap 6, de schemavolgorde is hiervoor al besproken. Schemavolgorde is de volgorde waarin de onderwerpen in het schema worden aangeboden. In principe is dat het omgekeerde van de combinatievolgorde. We stelden al eerder dat het omzetten van de combinatievolgorde in schemavolgorde alleen interessant is als er ook daadwerkelijk klassen nodig zijn voor de plaatsing van de algemene kenmerken. Dat is zeker in een bibliografische omgeving het geval omdat het informatie over een onderwerp bevat en niet gaat over de plaatsing van een specifiek product zoals flessen wijn in een wijnkelder of boodschappen in een supermarkt. Een pak melk bevat melk en kan als zodanig weggezet worden bij de andere zuivelproducten. Maar een boek over melk kan makkelijk gaan over meer dan alleen het fysieke product melk. Het zou ook informatie over veeteelt in Nederland of de boerenbedrijven of de kwaliteit van grassoorten kunnen bevatten en dan staat het niet meer bij het onderwerp 'melk', maar bij een meer omvattende, grotere en meer algemene rubriek zoals 'landbouw'.

Combinatievolgorde (van specifiek naar algemeen)	Schemavolgorde (van algemeen naar specifiek)
Producten / soorten	Vorm
Delen / onderdelen	Tijd
Materialen	Plaats
Kenmerken	Bewerkers
Processen / handelingen	Processen / handelingen
Bewerkers	Kenmerken
Plaats	Materialen
Tijd	Delen / onderdelen
Vorm	Producten / soorten

Tabel 6.3 - De combinatievolgorde en schemavolgorde als inverse van elkaar.

In het algemeen kunnen we stellen dat als we op zoek gaan naar informatie die fysiek geordend is, bijvoorbeeld boeken in een kast, we ervan uitgaan dat vooraan meer algemene werken staan, terwijl meer naar beneden in diezelfde kast de specifiekere onderwerpen staan. Daarvoor zorgt de schemavolgorde.

We zien in tabel 6-3 dat de combinatievolgorde van het (voor het doel/de doelgroep) meest specifieke onderwerpselement afloopt naar het meest algemene. De schemavolgorde (afstelvolgorde, *filing order*) loopt van algemeen naar bijzonder.

6.2.3 Stap 7 en 8: werk de classificatie uit in een schema en ken notatie toe

Als de inhoud van de klassen (stap 2 en 3) en de volgorde zowel in de klasse (stap 4) als tussen de klassen onderling is vastgesteld (stap 5 en 6), dan kan het schema verder uitgewerkt worden. Met name deze en de volgende stap is voor bibliografisch materiaal belangrijk: het bepaalt hoe het schema er uiteindelijk uit zal zien en hoe het materiaal tenslotte in de kast komt te staan. De classificatie is immers een gecontroleerde taal die gebruik maakt van kunstmatig gemaakte termen (codes of notaties) die de onderwerpen een plaats geven in de systematiek. Deze termen en hun codes worden in stap 7 en 8 gemaakt.

specificiteit

Specificiteit of nauwkeurigheid is van belang bij het ontwerp van een schema. Specificiteit is de mate waarin de inhoud van een document exact wordt omschreven in termen van een informatietaal. Een schema waarin te weinig klassen zijn gedefinieerd, kan informatie alleen op een algemeen niveau plaatsen waardoor de werkelijke informatie verloren gaat.

Hoe specifieker de informatietaal (in dit geval de classificatie) hoe specifieker de informatie ingedeeld kan worden. Het vergroten van de specificiteit van een classificatie is mogelijk door:

- verder hiërarchisch onder te verdelen door meer niveaus toe te voegen, en/of
- combinatie met andere onderwerpselementen volgens de combinatievolgorde.

Een classificatie kiest voor de ordening nooit de alfabetische volgorde van de opgenomen begrippen, maar kent aan die begrippen een notatie toe. Een notatie is van belang omdat het de plaats van het onderwerp in het schema vastlegt en bijdraagt aan de overzichtelijkheid van de classificatie. Daarnaast is de notatie goed te gebruiken als onderdeel van een signatuur: het geeft het document een plaats in de kast.

signatuur

Een signatuur is in veel gevallen opgebouwd uit de notatie, gevolgd door de eerste letters van de achternaam van de auteur of, bij afwezigheid daarvan, de eerste letters van de titel.

Een notatie kun je opbouwen op verschillende wijzen:

- alleen letters of alleen cijfers (getallen of decimalen) en
- gemengd: alfanumeriek

Ook symbolen, zoals &, (), %, "", ; kunnen in de notatie voorkomen.

Het is van belang om de notatie pas toe te gaan kennen als het schema is uitgewerkt. Wie dat te vroeg doet, verslikt zich vaak in de regels voor de notatie. Toch behandelen we deze stappen tegelijkertijd, omdat notaties vaak gebonden zijn aan de keuze van het schema en het de voorbeelden verheldert om meteen met notaties te werken.

De drie manieren om een schema uit te werken zijn:

- analytisch-synthetisch: geen samengestelde onderwerpen in het schema, geen onderscheid tussen hoofd- en hulprubrieken;
- enumeratief: samengestelde onderwerpen zijn al in het schema uitgewerkt;
- enumeratief met analytisch-synthetische mogelijkheden: een gedeelte van de samengestelde onderwerpen is uitgewerkt en veel voorkomende onderverdelingen worden aangeboden in de vorm van bijzondere en/of algemene hulprubrieken.

Hieronder staan de drie vormen van de schema's naast elkaar.

Onmiddellijk valt op dat het enumeratief schema het langst is: dat komt doordat alle rubrieken uitgeschreven worden.

De verschillen tussen deze drie schema's werken we hierna uit. De overeenkomst ligt in het feit dat ze allen precoördinatief zijn. De klassen worden namelijk door de indexeerder en dus vóór de zoekfase vastgesteld.

Analytisch-synthetisch	Enumeratief met analytisch-synthetische mogelijkheden	Enumeratief
A1 voor 1900	640 Instrument bespelen	640 Instrument bespelen
A2 1900-heden	641 Instrumentreparatie	640.1 instrumenten voor 1900
B1 instrument bespelen	650 Piano	640.2 instrumenten 1900-heden
B2 Instrument reparatie	650.1 Piano bespelen	641 Instrumentreparatie
C1 Piano	650.2 Piano reparatie	641.1 instrumentreparatie voor 1900
C2 Viool	651 Viool	641.2 Instrumentreparatie; 1900-heden
	651.1 Viool bespelen	650 Piano
	651.2 Vioolreparatie	650.1 Piano's voor 1900
	Hulptabel van tijd	650.2 Piano's 1900-heden
	(1800) Voor 1900	651 Piano bespelen
	(1900) 1900-heden	651.1 Piano bespelen voor 1900
		651.2 Piano bespelen i1900-heden
		652 Piano reparatie
		652.1 Pianoreparatie voor 1900
		652.2 Pianoreparatie 1900-heden
		660 Viool
		660.1 Violen voor 1900
		650.2 Violen 1900-heden
		Etc...

Afb. 6.5 - Drie schema's.

analytisch-synthetisch schema

facetclassificatie

Analytisch-synthetisch schema

Een analytisch-synthetisch schema of facetclassificatie is voor de bouwer het gemakkelijkst te maken. In dit schema zijn de homogene groepen afzonderlijk nog uitstekend te herkennen. Deze homogene groepen worden in deze context ook wel facetten genoemd, een naam die Ranganathan introduceerde. En zijn facetten zijn welbekend: in paragraaf 3.2 hebben we al laten zien op welke manier deze facetten behulpzaam zijn bij het zoeken op het web. De COLON classificatie van Ranganathan wordt gezien als een extreem voorbeeld van een analytisch-synthetisch schema.

In schema's staat doorgaans de meest algemene groep (klasse) vooraan in het schema en de meest specifieke achteraan, zo ook in dit geval. In geval van een samengesteld onderwerp kan een aantal van die facetten gecombineerd worden tot een nieuwe klasse. Deze klasse staat niet in het schema, maar ontstaat pas als een document over dit samengestelde onderwerp geïndexeerd wordt. Daarbij is de volgorde van de facetten die van de aangegeven combinatievolgorde. Dit is van belang omdat alleen als die wordt aangehouden, het document op de meest specifieke plaats in de kast terecht komt.

Bij de notatie van een analytisch-synthetische classificatie wordt meestal een letter- of alfanumerieke notatie toegepast.

In een heel eenvoudig voorbeeld is dat weergegeven in onderstaand overzicht:

A1 Nederland	B1 Broedgedrag	C1 Eenden
A2 Frankrijk	B2 Voedselvergarig	C2 Ooievaars
A3 Italië	B3 Trekken	C3 Reigers

Een boek over het broedgedrag van ooievaars in Nederland krijgt van de indexeerder de notatie C2 : B1 : A1. De notatie weerspiegelt daarmee de combinatievolgorde. Die loopt altijd van specifiek naar algemeen en zorgt ervoor dat alle documenten over ooievaars bij elkaar komen te staan. Ook de persoon die in de kast zoekt naar een boek over dit onderwerp zal gaan zoeken bij de 'Ooievaars'.

enumeratief classificatie- schema

Enumeratief schema

In een enumeratief classificatieschema zijn alle begrippen die bij het klasseren gebruikt mogen worden, al opgesomd. Het zijn eenvoudige maar vooral ook samengestelde onderwerpen. Alle zinvolle combinaties van verschillende elementen zijn al van te voren door de bouwer in het schema vastgelegd in klassen (rubrieken). De indexeerder gebruikt vervolgens alleen de combinatie van begrippen die al in het schema zijn opgenomen.

Het enumeratief uitwerken van een classificatie betekent dus dat de potentiële complexiteit van alle onderwerpen al in de bouwfase wordt vastgelegd. Hiermee wordt in die bouwfase al besloten welke onderwerpen (ook de samengestelde begrippen) zo vaak zullen voorkomen dat het zinvol is ze in het schema op te nemen. Ook hierbij geldt dat de schemavolgorde van algemeen naar specifiek loopt en er dus bepaald moet worden welke algemene klassen er moeten worden opgenomen. In een schema over vogels moet een document over broedgedrag van ooievaars een plaats kunnen krijgen, maar ook documenten over ooievaars in het algemeen en broedgedrag (van vogels) in het algemeen.

Een enumeratief schema kan erg omvangrijk worden omdat bijvoorbeeld hele algemene aspecten als land of tijd bij praktisch alle onderwerpen een rol kunnen spelen. Vandaar dat zuiver enumeratieve schema's nauwelijks voorkomen.

enumeratief schema met analytisch- synthetische mogelijkheden

Enumeratief met analytisch-synthetische mogelijkheden

Bij schema's die gebouwd zijn volgens dit principe is het uitgewerkte, vaste karakter van het enumeratieve schema gecombineerd met het vrije analytisch-synthetische principe. Een enumeratief schema met analytisch-synthetische mogelijkheden heeft een enumeratief uitgewerkte

basis. Samengestelde onderwerpen worden in het schema opgenomen. Daarentegen zijn algemene aspecten als tijd en plaats niet in het schema vastgelegd, maar worden als algemeen toepasbare hulprubrieken (algemene hulpgetallen) toegevoegd aan het schema. Zodra er bij het indexeren behoefte is aan een onderverdeling naar aspecten uit deze hulptabellen (zoals plaats en tijd), kan de indexeerder deze aan de oorspronkelijke klasse toevoegen (natuurlijk rekening houdend met de geldende combinatievolgorde).

In het voorbeeld over de vogels kan de bouwer besluiten wel een klasse te maken over broedgegedrag van vogels, maar de landen in een aparte tabel te vermelden die gebruikt kan worden wanneer dat van toepassing is bij het klasseren van een specifiek document

In een enumeratief schema staat een code vaak voor een combinatie van begripseenheden. In die code is dan weer de combinatievolgorde te herkennen. Zo is in onderstaande voorbeeld het repareren van piano's weergegeven in de code 650.2.

Zoals in de vorige paragraaf aangegeven, komen zuivere enumeratieve classificaties nauwelijks voor en wordt voor algemene aspecten zoals tijd en plaats meestal gewerkt met hulptabellen. Dat geeft deze enumeratieve classificaties een analytisch-synthetisch karakter.

Een boek over pianospelen in de 19e eeuw zal de notatie 650.1 (1800) krijgen. Voor een gedeelte is de combinatievolgorde al in het schema vastgelegd (650.1 = piano bespelen), gedeeltelijk wordt hij door de indexeerder toegevoegd aan het onderwerp: (1800) wordt achteraan de notatie geplaatst. Tijd is namelijk een algemener kenmerk dan de activiteit: bespelen en het soort: het instrument piano.

Ook hier een eenvoudig voorbeeld:

- 640 Instrument bespelen
- 641 Instrumentreparatie
- 650 Piano
 - 650.1 Pianospelen
 - 650.2 Piano reparatie
- 651 Viool
 - 651.1 Vioolspelen
 - 651.2 Vioolreparatie

Hulptabel van tijd
 (1800) Voor 1900
 (1900) 1900-heden

6.2.4 Schema bouwen en notaties

Wanneer de bouwer voor een van de bovenstaande drie schema heeft gekozen, dan kan hij de homogene groepen zodanig uitwerken dat het schema ontstaat.

In alle gevallen is het zo dat een goed schema plaats geeft aan alle gebruikte onderwerpen. Dat betekent dat er in het schema ook plaats is voor onderwerpen die nog niet bekend zijn bij de bouw van het schema. Als er behoefte is aan een nieuwe rubriek, moet het schema aangepast kunnen worden en eigenlijk kan dat op drie manieren:

1. Plaatsen onder overige (er is dan altijd plaats voor nieuwe ontwikkelingen) bijvoorbeeld het maken van een rubriek 'overige vervoermiddelen' onder een rubriek 'vervoermiddelen'.
2. Een extra rubriek maken (als notatie dat toelaat: uitbreiden in de breedte). Het voorbeeld over het bespelen van instrumenten zou bijvoorbeeld uitgebreid kunnen worden met een extra rubriek over het bespelen en repareren van gitaren.
3. Rubriek verder onderverdelen (verder de diepte in). Deze vorm van uitbreiding is altijd mogelijk. Een voorbeeld hiervan is door een rubriek 'fietsen' onder te verdelen in de subrubrieken 'sportfietsen' en 'stadsfietsen'.

In enumeratieve schema's kunnen de notaties vrij lang worden. Vanuit die redenering is er een aantal eisen ontwikkeld, die deels met elkaar botsen, namelijk: gastvrij, expressief, gemakkelijk te herinneren en kort. Dit alles is nooit tegelijk te verwezenlijken. Hoofdstuk 7 van "Organiseer je Informatie" gaat hier verder op in.

Notaties hebben het grote nadeel dat ze drempelverhogend werken en daarmee niet bijdragen aan de gebruiksvriendelijkheid. Met name publieke instellingen als openbare bibliotheken, die nu nog een classificatie als ordeningsschema gebruiken, werken steeds meer met nieuwe mogelijkheden van plaatsing en ordening, bijvoorbeeld door gebruik te maken van pictogrammen en trefwoorden om materiaal op een meer gebruiksvriendelijke manier te presenteren.

6.2.5 Stap 9 en 10: Register samenstellen en handleiding schrijven

Alfabetische registers op classificaties hebben de functie om samenhangen die het schema heeft verbroken weer zichtbaar te maken. Bijvoorbeeld een onderwerp dat in verschillende contexten/sferen voorkomt, wordt in de vorm van meer dan een registreringang via het alfabet weer bij elkaar gebracht. Als voorbeeld hiervan kun je denken aan het onderwerp kinderen dat op verschillende plaatsen in het schema terugkomt: kindergeneeskunde bij de geneeskunde, het opvoeden van kinderen bij

de pedagogiek, kinderpsychologie bij de psychologie, de maatschappelijke zorg voor kinderen bij de sociologie, enzovoort.

cross-classification

Daarnaast vergroot het register de toegankelijkheid van een schema: het wijst de weg naar de plaats in het schema, het legt beslissingen/twijfelgevallen vast en het probeert daarmee *cross-classification*, het klasseren van gelijksoortig materiaal in verschillende klassen te voorkomen. Doordat classificatieschema's doorgaans digitaal beschikbaar en doorzoekbaar zijn, is het register minder van belang geworden. Voor wie zich echter wil verdiepen in de verschillende vormen van registers vormt het boek "Organiseer je informatie" een bruikbaar startpunt.

Over het maken van handleidingen is veel informatie beschikbaar. Toch noemen we twee zaken nog heel kort: bedenk voor wie de handleiding wordt geschreven. Vaak zijn er twee mogelijke groepen:

- de indexeerder die kan lezen hoe een classificatie in elkaar zit, wie hem beheert en hoe hij gebruikt moet worden bij het indexeren. Dit laatste betreft bijvoorbeeld de indexeerdiepte en eventuele synthetische mogelijkheden;
- de zoeker die kan lezen hoe de classificatie werkt en hoe hij deze het beste kan raadplegen.

Hoewel een classificatie zo ontworpen zou moeten worden dat de zoeker geen handleiding nodig heeft, lukt dat toch niet altijd. Geef in dat geval de zoeker aanwijzingen over de manier waarop de classificatie toegepast is bij het toegankelijk maken van informatie.

6.3 Samenvattend

In dit hoofdstuk hebben we het gehad over ordeningssystemen. In een fysieke omgeving noemen we ze ook wel classificaties. In het eerste deel van het hoofdstuk legden we de ordening van materiaal in digitale mappen uit aan de hand van zeven stappen van ordening. Belangrijk hierbij was de introductie van de combinatie- en schemavolgorde. Waar de combinatievolgorde ervoor zorgt dat materiaal bij het meest specifieke facet geplaatst wordt en daardoor terug te vinden is, zorgt de schemavolgorde ervoor dat de informatie als geheel gerangschikt wordt van algemeen naar bijzonder.

In het tweede deel verschuift de aandacht naar de ordening van fysieke collecties. Naast de hiervoor genoemde stappen is het bij dit materiaal ook nodig dat het een notatie krijgt toegekend, waardoor het materiaal in de kast kan worden geplaatst. Hiervoor zijn drie soorten schema's mogelijk:

- enumeratieve schema's waarin alle mogelijkheden worden uitgeschreven;
- enumeratieve classificaties met analytisch-synthetische mogelijkheden waarin algemene rubrieken kunnen worden gekoppeld aan het onderwerp;
- analytisch-synthetische schema's waar alle facetten benoemd worden en zowel de zoeker als de indexeerder zelf combinaties maakt.

Met de laatste stappen: het maken van notaties, register en handleiding maken we het 10-stappenplan voor het ontwerpen van een classificatie compleet. Omdat slechts weinigen classificaties vanaf de basis opbouwen, zijn we hier maar kort op ingegaan.

De rol van de indexeerder is al een aantal keer genoemd als degene die een informatietaal gebruikt om materiaal daadwerkelijk toegankelijk te maken. Deze rol staat centraal in het volgende hoofdstuk.

Gebruikte literatuur:

Ranganathan (2010). Shyah Ramamrita Ranganathan. In *Encyclopedia Britannica*. Geraadpleegd op: 21 juni 2016, van <https://www.britannica.com/biography/Shiyali-Ramamrita-Ranganathan>

Verder lezen:

Becker, P., van der Linden, M., Magrijn, H. & Sieverts, E. (2010). *Organiseer je informatie; aan de slag met thesauri, taxonomieën, tags en topics*. Leidschendam: Biblion. ISBN 9789054839545

Bedford, D. (2013). *Evaluating Classification Schema en Classification Decisions*. Geraadpleegd op 25 augustus 2016 van: <http://onlinelibrary.wiley.com/doi/10.1002/bult.2013.1720390206/epdf>

Kwasnick, Barbara H. (1999). *The role of classification in knowledge representation and discovery*. Geraadpleegd op 21 juni 2016, van: https://www.ideals.illinois.edu/bitstream/handle/2142/8263/librarytrendsv48i1d_opt.pdf

Hoofdstuk 7

Toegekend indexeren

informatie- verrijking

In de voorgaande hoofdstukken bespraken we een aantal informatietalen die we gebruiken om informatie toegankelijk te maken. Achtereenvolgens schonken we in hoofdstuk 3 aandacht aan de taxonomie, stond de thesaurus centraal in hoofdstuk 4, de autorisatielijsten in hoofdstuk 5 en de classificatie in hoofdstuk 6. Uiteindelijk moeten de termen uit een informatietaal als metadata worden toegekend aan de afzonderlijke documenten of pagina's in een informatiesysteem of website. We proberen op die manier zo goed mogelijk duidelijk te maken waar het document over gaat zodat het op inhoud is terug te vinden. Dit hoofdstuk gaat daarop in. We bespreken het (door mensen) indexeren van tekst en beeldmateriaal en de uitwerking daarvan in verschillende informatietalen. Dit indexeren is een vorm van informatieverrijking: het toevoegen van gegevens (metadata) aan het oorspronkelijk materiaal, met als doel het materiaal beter toegankelijk te maken in een informatiesysteem. Omdat de termen niet, zoals bij een zoekmachine, zonder meer uit de tekst worden gehaald, spreken we over toegekend indexeren (Engels: *assigned indexing*). Deze informatieverrijking vindt plaats door middel van het toekennen van termen, tags of een classificatiecode, maar ook door het schrijven van een samenvatting of het toekennen van een term uit een autorisatielijst zoals taal of land.

Hoe meer termen er aan een document zijn toegekend, des te beter is het vindbaar. Zeker als deze termen verspreid zijn over verschillende facetten. Op deze manier kan een document terug worden gevonden op bijvoorbeeld projectnummer, onderwerp-omschrijving en documenttype. Daarbij moeten afwegingen worden gemaakt: ken je termen toe over alle deelaspecten van een werk of beschrijf je het onderwerp op een globaler niveau? We gaan hier in paragraaf 7.2.2 dieper op in.

Zoals gezegd, gaat het bij toegekend indexeren meestal om handmatig werk. De mogelijkheden om door software automatisch de juiste termen te laten toekennen of te laten ontlelen uit een document (Engels: *derived indexing*) komen aan de orde in hoofdstuk 10.

Indexeren kost tijd en dus geld, maar dat geldt voor vergeefs zoeken ook. Het loont in de meeste gevallen de moeite omdat we een document slechts eenmaal hoeven te indexeren, terwijl het wellicht vele malen gezocht wordt. Wel is het belangrijk dat we de gebruiker steeds centraal stellen: wat heeft deze nodig om het gezochte te kunnen vinden, welke

kwaliteit moet het gevondene bieden, hoe belangrijk is het dat hij deze informatie vindt en hoe gemakkelijk wordt het hem daarbij gemaakt?

**document-
gerichte
benadering**

**gebruikers-
gerichte
benadering**

In de informatiewetenschap bestaat geen algemene theorie met betrekking tot indexeren. Wel zijn er twee benaderingen te onderscheiden: de documentgerichte benadering en de gebruikersgerichte benadering. Bij de eerste benadering is het document zelf het meest van belang en komt de indexeerder alleen op basis van de inhoud van de tekst tot een beschrijving van de inhoud. Het achterliggende idee hierbij is dat alleen dan de tekst goed weergegeven wordt en bovendien in de loop van de tijd zijn geldigheid behoudt (Soergel, 1985). Bij de gebruikersgerichte benadering staat het document natuurlijk ook nog steeds centraal, maar neemt de indexeerder de vrijheid bij de keuze van de termen rekening te houden met de informatiebehoefte van de gebruiker (Lancaster, 2003). In de rest van dit hoofdstuk zullen we uitgaan van de gebruikersgerichte benadering omdat die meer aansluit bij de praktijk van de informatiespecialist.

Voordat we termen of een samenvatting aan een document kunnen toevoegen, moeten we eerst een inhoudsanalyse uitvoeren. Je kunt pas zeggen dat een document het kweken van orchideeën als onderwerp heeft en dat het een beginnersniveau betreft als je kennis hebt genomen van de inhoud. Je kunt ook pas zeggen dat op een foto een vaas is afgebeeld als je de foto hebt bekeken en geïnterpreteerd. Inhoudsanalyse is daarmee de eerste stap van het proces van indexeren.

Samenvattend kunnen we zeggen dat het proces van indexeren grofweg in twee stappen verloopt:

- het voorwerk: de inhoudsanalyse, het karakteriseren van de inhoud;
- de keuze en toekenning van termen of andere inhoudelijke kenmerken (tags, rubrieken, classificatiecodes, enzovoort) aan een document.

7.1 Inhoudsanalyse

inhoudsanalyse

In de communicatiewetenschap is inhoudsanalyse een veelgebruikte term voor onderzoek naar media-inhoud. De meest klassieke en meest geciteerde definitie is die van Berelson (1952). Berelson was een gedragswetenschapper die bekend is door zijn werk op het gebied van communicatie en massamedia. Hij schreef: "De inhoudsanalyse is een onderzoekstechniek voor de objectieve, systematische en kwantitatieve beschrijving van de manifeste inhoud van de communicatie". Ook binnen de informatiewetenschap zijn objectiviteit en systematiek kenmerken van een goede inhoudsanalyse. Met objectiviteit bedoelen we hier dat verschillende mensen in dezelfde situatie tot een overeenkomstige inhoudsanalyse moeten komen. De systematiek houdt dan in dat de

analyse volgens een vast proces verloopt, gericht op het doel waarvoor de analyse wordt gemaakt.

conceptuele analyse

Wie de inhoud van een document gaat analyseren, zal systematisch moeten kijken naar de concepten, facetten, en/of invalshoeken die in een document samenkomen. We spreken dan ook wel over conceptuele analyse. Bij een sociaalwetenschappelijke publicatie gaat het dan over facetten als: het onderzochte probleem, de onderzochte populatie naar leeftijd, naar geslacht, naar sociale klasse en naar geografische herkomst, de gebruikte onderzoeksmethode en dergelijke. Hoewel er bij verschillende vakgebieden en verschillende soorten documenten heel verschillende facetten kunnen voorkomen, is het toch goed om te proberen een zo systematisch mogelijke aanpak te hanteren. Het nadenken over facetten helpt ook bij de inrichting van een systeem: voor welke facetten moeten metadatavelden worden gedefinieerd? Wie een goede zoekvraag wil formuleren, zal immers dezelfde methode hanteren. Op medisch terrein is dat bijvoorbeeld uitgekristalliseerd tot de PICO methode. Daarin staat de P voor *population* (de patiënt en zijn probleem), de I voor *intervention* (de behandeling of aanpak van dat probleem), de C voor *comparison* (de vergelijking met andere behandelmethode(n)) en de O voor *outcome* (de effecten of resultaten van die behandelingen). Op die manier zijn de vier belangrijkste aandachtsgebieden voor het formuleren van, op medische behandeling gerichte, zoekvragen samengebracht. Veel medische informatiesystemen zijn daar dan ook op ingericht. Voor de conceptuele analyse ten behoeve van het beschrijven van documenten kunnen we een voorbeeld nemen aan een dergelijke systematische aanpak. Daarbij moeten we onszelf wel steeds de vraag stellen of alle facetten werkelijk noodzakelijk zijn. Ieder facet moet immers beschreven worden door middel van één of meer termen. In systemen waarbij de gebruikers zelf termen moeten toekennen is het daarom belangrijk om goed na te denken over het aantal velden en termen dat indexeerders moeten invullen. Daarbij gaat het dus om de balans tussen wat gewenst is voor de vindbaarheid en wat werkbaar is voor de indexeerder.

inhouds- karakteristiek

Als je een inhoudsanalyse wilt maken, moet je natuurlijk kennis kunnen nemen van de inhoud: de taal en het schrift moeten bekend zijn en de inhoud moet begrepen kunnen worden. Enige kennis van het onderwerp en vakgebied is dan ook vaak noodzakelijk. Voor een gebruiker die zijn eigen document in een systeem plaatst, zal dat geen probleem zijn, maar een informatiespecialist die literatuur van anderen moet beschrijven zal zich daarvan bewust moeten zijn.

De uitkomst van de inhoudsanalyse noemen we een inhoudskarakteristiek. Dat is een kernachtige zin waarin de belangrijkste elementen van het document terug te vinden zijn en die de basis geeft voor de toekenning van termen. Wie de inhoudsanalyse alleen gebruikt als tussenstadium voor de vertaling naar termen, maakt de inhoudskarakteristiek meestal in

het hoofd of in de vorm van een paar genoteerde steekwoorden. Het kan echter ook voorkomen dat de inhoudskarakteristiek zelf onderdeel is van de documentbeschrijving, bijvoorbeeld in de vorm van een samenvatting. Hoewel het resultaat er vaak anders uitziet, is het doel van het maken van een inhoudsanalyse en het schrijven van een samenvatting vrijwel gelijk: in beide gevallen willen we aangeven waar het document over gaat.

Wie een inhoudsanalyse maakt, gaat daarbij altijd op zoek naar de handeling in het document, want dat geeft eigenlijk altijd antwoord op de vraag: “waar gaat dit document over?” Deze informatie is doorgaans te vinden in een of meer van de onderstaande elementen, ongeveer in deze volgorde van waarschijnlijkheid en belang:

- titel
- abstract of samenvatting
- inhoudsopgave
- voorwoord en verantwoording
- inleiding, beginalinea's, conclusies
- illustraties, tabellen en literatuuropgaven
- teksten die opvallen door een afwijkende lay-out
- auteursnaam of uitgeversnaam
- (al bestaande) metadata
- URL

Het maken van een inhoudskarakteristiek is soms lastig. Vaak nemen we de titel als uitgangspunt. Maar een titel is lang niet altijd bruikbaar omdat het niet altijd de beknopte inhoud van het document weergeeft. Sommige auteurs kiezen ervoor een pakkende titel te gebruiken die de lezer in spanning laat over de inhoud. De titel “Een koelkast vol familie” van Meininger en Davidse kan allerlei beelden oproepen, maar geeft niet duidelijk aan waar het boek precies over gaat. Ook andere elementen, zoals een uitgever of auteur kunnen een indexeerder verwarren; bijvoorbeeld als een auteur altijd over natuurkundige zaken schrijft en op een later moment gaat publiceren over zoekstrategieën.

Als we alleen een samenvatting als bron gebruiken voor de inhoudsanalyse dan moeten we ons goed realiseren dat het gaat om een weergave en dat het niet de tekst zelf is. Het hangt af van de vaardigheden van de maker van de samenvatting of het een accurate beschrijving van de inhoud van het document heeft opgeleverd. Bovendien kan de samenvatting een wervende functie hebben en daarmee niet helemaal neutraal de inhoud weergeven.

Met de inhoudsopgave en inleiding geeft de auteur als het ware een plattegrond van het document en in de inleiding worden ook keuzes ten aanzien van het onderwerp verantwoord. Het geeft inzicht in de globale structuur van het document. Vervolgens kan de indexeerder in de

beginlinea's, conclusies, illustraties en teksten met een afwijkende lay-out op zoek gaan naar meer specifieke informatie.

Ook metadata die al zijn toegekend aan een document, zoals een webpagina, kunnen helpen te bepalen waar een document over gaat. Tenslotte kan een URL een hulp zijn bij de inhoudsanalyse omdat daarin bijvoorbeeld een verwijzing naar de organisatie en de (sub)afdeling of categorie vermeld staat.

Inhoudsanalyse gaat niet alleen over de vraag waar het document over gaat, maar ook waarom dat voor jouw doelgroep interessant is. Daardoor kun je nooit spreken over "de beste termen" om iets toegankelijk te maken. Hetzelfde document kan in verschillende situaties heel anders geïndexeerd worden omdat de gebruikers om verschillende redenen in het document zijn geïnteresseerd. Dat houdt in dat een indexeerder bij de inhoudsanalyse niet alleen uitgaat van zijn materiaal maar ook van de doelgroep die dat materiaal gaat gebruiken. In paragraaf 7.3 geven we hiervan een voorbeeld.

Het lijkt wat omslachtig dat een indexeerder niet meteen aan de hand van een informatietaal de termen toekent aan een document, maar daarmee kunnen we wel voorkomen dat de inhoudsanalyse wordt gestuurd door de gecontroleerde termen. Het is verleidelijk om een term toe te kennen als we weten dat die voorkomt in de informatietaal: als de titel van het werk het woord 'molens' bevat en de indexeerder weet dat deze term voorkomt in de informatietaal, dan is het gemakkelijk om de term 'molens' toe te passen en niet verder te kijken. Maar als dit vrij vroeg in het stadium van analyse gebeurt, dan kan de analyse van de inhoud van het document hierdoor gekleurd worden.

Een slechte inhoudskarakteristiek levert verkeerde vertalingen op en een verkeerde vertaling betekent dat een zoeker dat werk niet vindt, of dat de zoeker resultaten krijgt die hij niet wenst (ruis). Er is dan sprake van tijd- en dus geldverspilling bij zowel de indexeerder als de zoeker. Tegelijkertijd vormt een goede inhoudsanalyse nog maar het halve werk: de vertaling naar de informatietaal is net zo belangrijk. Daarover meer in de volgende paragraaf.

7.2 Van inhoudsanalyse naar informatietaal

Als de fase van de inhoudsanalyse is doorlopen, heeft de indexeerder een beeld waar de auteur of maker het over heeft. De volgende fase is die van het vertalen van de inhoudskarakteristiek naar het vocabulaire van een informatietaal. Hierbij maakt het natuurlijk uit of we de termen toekennen op basis van een gecontroleerde informatietaal zoals een autorisatielijst, classificatie of een thesaurus (zie paragraaf 7.3) of dat ze vrij toegekend worden zoals bij tagging (zie paragraaf 7.4).

In een gegeven situatie is de keuze voor een bepaalde gecontroleerde informatietaal natuurlijk al gemaakt. Daarin speelden verschillende overwegingen een rol die betrekking hadden op zaken als indexeerdiepte en globaal- versus diepte-indexering. Deze begrippen bekijken we eerst voordat we in paragraaf 7.3 verder ingaan op de tweede fase van het indexerproces: het toekennen van termen.

7.2.1 Indexeerdiepte

indexeerdiepte

De indexeerdiepte is de mate waarin de inhoud van een werk wordt geanalyseerd en weergegeven in termen van de informatietaal. Als we afspreken dat de indexeerdiepte bij een bepaalde collectie om en nabij de vijf ligt, dan wil dat zeggen dat een document gemiddeld gesproken met ongeveer vijf termen toegankelijk is gemaakt. De indexeerdiepte levert daardoor vaak een getal op.

Een document is beter terugvindbaar als er verschillende termen aan zijn toegekend. Het is dus onhandig om twee documenten die ongeveer over hetzelfde onderwerp gaan van twee geheel verschillende aantallen termen te voorzien; in gelijke gevallen willen we namelijk beide documenten vinden. Als er twee documenten te vinden zijn over jongeren en internetgebruik en het ene document heeft tien termen en het andere drie, dan is het document met maar drie termen lastiger terug te vinden dan het document met tien termen. Heel in het algemeen kunnen we dus stellen dat het praktisch is als documenten (of documentsoorten) die voor gelijk gebruik bestemd zijn en gelijke complexiteit hebben, ook ongeveer dezelfde indexeerdiepte hebben.

Beleidsbeslissingen over het aantal te gebruiken termen zijn niet in harde limieten te geven, maar meer in een minimum-maximum range. Daarbij is het zinnig om een onderscheid te maken tussen systemen met een bibliotheekfunctie zoals een catalogus of databank met tijdschriftartikelen enerzijds en systemen voor documentmanagement anderzijds. In bibliotheeksystemen zullen de meeste items geïndexeerd worden met 8-15 termen. Bovendien kan het binnen een organisatie nog afhangen van het type document. We zouden daarvoor bijvoorbeeld een overzicht als onderstaand staatje kunnen maken:

Eigen technische rapporten	15-25 termen
Externe technische rapporten	10-15 termen
Patenten	15-20 termen
Tijdschriftartikelen	5-10 termen

Natuurlijk zal de individuele indexeerdiepte weleens verschillen van de gemiddelde indexeerdiepte. Een gemiddelde indexeerdiepte van vijf houdt in dat er waarschijnlijk ook documenten voorkomen met maar drie termen, of juist acht termen, afhankelijk van wat in een document aan de orde komt.

In documentmanagementsystemen liggen de aantallen lager. Het feit dat gebruikers zelf documenten plaatsen en termen moeten toekennen, speelt daarbij een belangrijke rol. Zij moeten er immers niet te veel tijd aan kwijt zijn. In het algemeen kunnen we ook stellen dat de documenten in deze context niet veel verschillende onderwerpen behandelen. In deze omgevingen kunnen we dan ook de 30-secondenregel hanteren: de indexeerder moet niet meer dan gemiddeld 30 seconden kwijt zijn aan het toekennen van de termen. De formele metadata helpen hierbij: formele metadata als klantnaam of projectnummer vormen vaak al waardevolle facetten om documenten vindbaar te maken.

Het aantal termen kan ook afhangen van het onderwerpsgebied. Ligt het onderwerp dicht bij de core business van de organisatie, kies er dan voor om er meer termen aan toe te kennen dan aan andere onderwerpsgebieden.

7.2.2 Globaalindexering en diepte-indexering

globaalindexering

diepte-indexering

Het laatste onderwerp waar we naar kijken alvorens we overstappen op het toekennen van termen gaat over de wijze waarop je naar een document kijkt. Bij het bepalen van de termen is het van belang om je te realiseren of de focus ligt op het beschrijven van het globale onderwerp van het document, of dat ook aparte delen van het document eigen termen krijgen. Als de focus ligt op het onderwerp van een heel document (het hele boek, de volledige website) dan spreken we over globaalindexering. Bij de tegenhanger, diepte-indexering, gaat het om het toekennen van genoeg termen om alle deelonderwerpen van een document compleet te beschrijven. Daarbij kijken we naar de verschillende onderwerpen die in delen van dat document zijn behandeld. Zo kan elk hoofdstuk uit een boek termen krijgen die alleen voor dat ene hoofdstuk van toepassing zijn. Het voordeel hiervan is dat een gebruiker dan ook boeken kan vinden waarin een gezocht specifiek onderwerp in één hoofdstuk aan de orde komt. Bij diepte-indexering kunnen de eenheden (paragraaf, hoofdstuk, filmfragment) heel klein of juist erg groot zijn. Het verschil met globaalindexering betreft de aandacht voor het geheel of juist voor onderdelen van de publicatie. Het onderscheid betreft niet noodzakelijkerwijs het aantal toegekende termen, al levert diepte-indexering door zijn aard doorgaans wel meer termen op dan globaalindexering.

Naast de keuze voor globaalindexering of diepte-indexering, is ook de specificiteit van de informatietaal van invloed op het uiteindelijke aantal termen dat een document kenmerkt: wie de (algemene) term 'auteursrecht' toepast, heeft minder termen nodig dan degene die een aantal specifieke aspecten die binnen het auteursrecht spelen apart opneemt, zoals bijvoorbeeld 'beeldrecht' en 'uitzendrecht'. Wie gebruik maakt van een informatietaal kan spreken van de specificiteit van de informatietaal, maar ook zonder een lijst waaruit we kunnen putten speelt

dit aspect. Gebruikers die zelf tags aanbrengen op een website voor het verzamelen en delen van documenten of URL's, zijn zich hier veelal van bewust en nemen bij de beschrijving vaak zowel de algemene termen ('auteursrecht' of zelfs 'recht') als de meer specifieke termen ('beeldrecht', 'digital rights management') op in de onderwerpsbeschrijving. Meer over tags staat in paragraaf 7.4.

We weten dat als er meer termen aan een document toegekend zijn, we het beter terug kunnen vinden. Daarom is het eigenlijk een merkwaardige situatie dat we bij boeken doorgaans kiezen voor globaalindexering. Een boek van gemiddeld een paar honderd pagina's waarin in afzonderlijke hoofdstukken tal van specialistische onderwerpen aan de orde komen, maken we met twee of drie termen toegankelijk, waardoor al die behandelde onderwerpen onvindbaar blijven. Anderzijds krijgen tijdschriftartikelen van enkele pagina's in bibliografische databases soms tientallen termen toebedeeld (Sieverts, 2009). Dat daar een keuze gemaakt wordt voor een grote indexeerdiepte, komt voort uit een andere achtergrond van de makers van deze beschrijvingen, waarbij het goed terug kunnen vinden van de informatie die hierin beschreven is, altijd voorop stond.

De keuze voor het toepassen van diepte-indexering is natuurlijk afhankelijk van een aantal factoren, zoals de kosten en/of de aanwezigheid van andere zoekmethoden.

Diepte-indexering leidt tot een hogere vangst, maar het is ook duurder. Bovendien heeft het ook vaak een wat lagere precisie tot gevolg. Want de opbrengt is groter, maar er worden ook items gevonden waarin:

- Verkeerde associaties gemaakt worden omdat twee termen gevonden worden die geen onderling verband of een onwenselijk verband hebben. Denk hierbij aan het voorbeeld in paragraaf 1.7 van mishandeling van of juist door hulpverleners. Voor wie op zoek is naar het ene aspect, is het andere niet relevant.
- Een deelaspect alleen in een klein gedeelte van het totale document wordt beschreven, waardoor het niet altijd relevant is voor de zoekers.

Uit bovenstaande blijkt dat een indexeerder voordat hij een keuze maakt voor een specifieke informatietaal, een aantal beslissingen zal moeten nemen over:

- de gewenste controle van termen;
- de specificiteit van de gekozen termen;
- de indexeerdiepte;
- de mate van globaalindexering of diepte-indexering;
- de middelen die beschikbaar zijn;
- de personen die gaan indexeren.

Overwegingen bij deze beslissingen zijn onder andere de beschikbare middelen en uiteraard het verwachte gebruik. Op basis hiervan kan beleid ten aanzien van de indexering worden gemaakt en een specifieke informatietaal gekozen worden; daarin zitten immers deze keuzes al verwerkt.

De genomen beslissingen hebben uiteindelijk invloed op het aantal termen dat aan een document zal worden toegekend, waarbij de 30 seconden regel bij een documentmanagementsysteem een mooie vuistregel is.

7.3 Regels bij het toekennen van termen

klasseren

We kennen termen toe vanuit de wens een document beter toegankelijk te maken. In deze paragraaf kijken we hoe we die termen toekennen met behulp van de informatietalen thesaurus en classificatie. Bij het gebruik van een classificatie noemen we deze activiteit ook wel klasseren: informatie onderbrengen in een bepaalde klasse van een in een schema vastgelegde structuur.

In een classificatie heeft de bouwer de combinatie van termen al gemaakt en kent de indexeerder de juiste termcombinatie toe. Dat geldt ook voor een mappenstructuur: de indexeerder kiest de juiste map om het document te plaatsen in een reeds bestaande structuur. Zelf een map creëren is wellicht mogelijk maar dan moet deze wel passen in de bestaande structuur.

Een thesaurus is een postcoördinatief systeem, wat betekent dat de zoekter in zijn zoekvraag een combinatie van termen maakt. De indexeerder maakt slechts op basis van de inhoudskarakteristiek een keuze voor een aantal termen. Hoeveel dit er zijn, is afhankelijk van de behoefte van de doelgroep en de specificiteit van de informatietaal.

In alle gevallen zijn er een of meer regels voor het toekennen van termen aan te geven:

1. Bekijk voor welke facetten termen moeten worden ingevuld.
2. Zorg dat zoveel mogelijk facetten ingevuld zijn.
3. Indexeer zo specifiek mogelijk als de informatietaal toelaat.
4. Indexeer zo specifiek als de gebruikersbehoefte vereist.
5. Benoem alle onderwerpen van het document die interessant zijn voor de gebruiker.

1. Bekijk voor welke facetten termen moeten worden ingevuld.

De bouwer heeft bepaald welke velden moeten of mogen worden ingevuld. Bij het indexereren is het belangrijk om te bekijken om welke

velden het gaat, of ze verplicht zijn en wat de gedachte achter de facetten is.

2. Zorg dat zoveel mogelijk facetten ingevuld zijn.

De bouwer heeft vooraf bepaald welke velden verplicht ingevuld moeten worden en welke niet. Voor de niet-verplichte velden geldt dat de indexeerder zorgvuldig moet afwegen of invullen een meerwaarde heeft voor andere gebruikers die het document willen vinden of beoordelen op relevantie.

3. Gebruik de specificiteit van de informatietaal.

De eerste regel van het indexeren is dat een indexeerder altijd zo specifiek mogelijk indexeert als de informatietaal toelaat. Bijvoorbeeld: een indexeerder moet een document over een programmeertaal indexeren. De inhoudsanalyse luidt dat het om een “introdutctie op de programmeertaal Java” gaat. De thesaurus heeft op dit punt de volgende termen:

Programmeertalen

NT C++

NT C#

NT Java

niveau-indexering

De indexeerder zal dan kiezen voor de term ‘Java’ en niet voor het meer algemene ‘programmeertalen’. Toch zien we in een aantal situaties niveau-indexering terug. We spreken over niveau-indexering (Engels: *redundant indexing*) wanneer zowel de overkoepelende term ‘programmeertalen’ als de onderliggende term ‘Java’ gebruikt wordt. Lancaster (2003) noemt dit: “onnodig en een voorbeeld van slechte indexering”. In het algemeen kennen we niet tweemaal een term van een ander niveau uit dezelfde boom toe. We zien het echter niet alleen terug bij onervaren gebruikers, maar ook in situaties waarin het retrievalsysteem de thesaurusfunctionaliteit niet goed kan ondersteunen. Het nadeel hiervan is, dat het dan niet meer mogelijk is om selectief te zoeken naar algemene documenten over programmeertalen, maar dat je dan altijd ook bijvoorbeeld documenten over alleen C++ als resultaat krijgt. Dat is te voorkomen door die bredere termen in een apart veld te zetten. In hoofdstuk 13 is hiervan een voorbeeld gegeven aan de hand van de CAB database.

4. Indexeer zo specifiek als de behoefte van de gebruikers vereist.

In paragraaf 7.1 gaven we aan dat het bij inhoudsanalyse niet alleen gaat over de vraag waarover het document gaat, maar ook waarom dat voor jouw doelgroep interessant is. Als gebruikers om verschillende redenen in het document zijn geïnteresseerd, kan hetzelfde document in verschillende situaties heel anders geïndexeerd worden. Dat houdt in dat een indexeerder

bij het indexeren niet alleen uitgaat van zijn materiaal maar ook van de doelgroep die dat materiaal gaat gebruiken. Dat is overigens (bij het gebruik van een informatietaal) ook vaak terug te zien in de specificiteit van de informatietaal zelf. Als illustratie vergelijken we de indexeerdiepte bij hetzelfde document in twee wezenlijk verschillende collecties.

De indexeerdiepte bij deze documenten verschilt aanzienlijk, wat waarschijnlijk samenhangt met de behoefte van de gebruiker. In afbeelding 7.1 heeft het boek “Cybersafety: an introduction” in de catalogus van de mediatheek van de Politieacademie veel meer termen gekregen dan in de catalogus van Bibliotheek.nl (afbeelding 7.2). Wie ertoe besluit verschillende collecties in één bestand doorzoekbaar te maken, dient zich ook van dit verschil bewust te zijn.

Volledige beschrijving		Pagina 72 / 563
Website mediatheek		
Nieuwe zoekactie	Materiaalsoort	
Geschiedenis		
Pagina >	Titel	Cyber safety : an introduction
Pagina <	Auteur	Rutger Leukfeldt ; Wouter Stol
Niveau *	Uitgave	The Hague : Eleven International Publishing, cop. 2012
Contact en leen informatie	Collatie	348 p. : graf., tab.
	Bibliografische annotatie	Met lit. opg.
	ISBN	9789490947750
Bewaar item	Rubriek	5200 digitale criminaliteit
Bewaarde items	Trefwoord	digitale criminaliteit ; internet ; informatiebeveiliging ; veiligheidszorg ; sociologie ; criminologie ; internationaal recht ; privacy ; hacking ; internetfraude ; slachtoffers ; daderprofiel ; kinderporno ; phishing ; spyware ; jongeren ; pornografie ; gaming ; cyberpesten ; criminaliteitsbestrijding ; politietaken ; digitale veiligheid ; botnet
Gebruikers acties		
Aanmelden		
Reserveren	Samenvatting	Introduction to the field of cyber safety. Presents an overview of current knowledge in that area. The concept of cyber safety refers to safety in cyber space. Cyber safety is thus wider than the commonly used concept 'cybercrime'. The book is divided into four parts. Part 1 covers the introductory concepts, looking at cyber safety from different perspectives (judicial, technological, sociological & criminological). In Part 2, various types of cybercrime are described individually, addressing issues such as definition, nature, scale and offender characteristics. The cybercrimes reviewed include hacking, e-fraud, child pornography, unauthorised file sharing and cyber warfare. Further more, criminal techniques like botnets, social engineering and phishing will be addressed. Part 3 deals especially with youth and cyber safety. Specific themes are: online sexual solicitations, sexually explicit internet material, cyber bullying and the perils and pleasures of playing video games. Part 4 is about organisations that play a part in safeguarding the internet. It addresses the role of the police, commercial organisations such as Internet Service Providers, and information security.
	Vindplaats	APELDOORN.CONCERNLOCATIE : 5200 LEUK APELDOORN.CONCERNLOCATIE : Magazijn A 11 18009 Alle exemplaren gereserveerd of uitgeleend Alle exemplaren gereserveerd of uitgeleend
	Details beschikbaarheid	Soortgelijke werken
Rate & Review(s)		
Rate	☆☆☆☆☆ (4.0 - totaal 1 stem)	
Geef uw waardering	Kies waardering Stem	
Review(s)		

Afb. 7.1 - Groot aantal trefwoorden bij de publicatie *Cyber safety* in de catalogus van de Politieacademie.

5. Benoem alle onderwerpen van het document die interessant zijn voor de gebruiker.

Voor alle informatietalen geldt dat een document een onderwerp kan behandelen dat gaat over iets dat in de informatietaal nog niet is benoemd. Het is goed om dat te constateren en dan in de gelegenheid te komen een term toe te voegen aan de informatietaal, in plaats van de term toe te kennen die het dichtst in de buurt ligt bij wat de indexeerder zou willen toekennen. De indexeerder heeft niet altijd de rechten om zelf termen toe te kennen. In dat geval zal hij zich moeten wenden tot

de redactie van de informatietaal met het verzoek om een term toe te voegen. Bij classificaties kan een subrubriek 'overig' in eerste instantie uitkomst bieden, maar deze mag daarvoor niet misbruikt worden. Als blijkt dat veel vergelijkbare documenten in deze rubriek terechtkomen, zal er een aparte rubriek voor moeten worden aangemaakt.

The screenshot shows the Bibliotheek.nl website interface. At the top, there's a navigation bar with links like HOME, E-BOOKS, JEUGD EN JONGEREN, LITERATUUR, WELKBOEK, THEMA, MUZIEK, and SERVICE. Below this is a search bar with the placeholder text 'Waar ben je naar op zoek?' and a 'Vind' button. The main content area features a book titled 'Cyber safety: an introduction' by Rutger Leukfeldt & Wouter Stol. The book cover is displayed on the left. To the right of the cover, the title and authors are listed. Below this, there are tabs for 'Gedrukt boek' and 'Digitaal boek'. Further down, there are fields for Genre (Non-Fictie), Onderwerpen (Technology, Nijverheid, Ambachten, Algemene sociale wetenschappen, Computercriminaliteit, internet, computerbeveiliging, Computerbeveiliging, Internet), and Taal (Engels). A button 'Vraag dit boek aan bij de bibliotheek' is present, along with a 'Aanvragen' button. At the bottom, there are social media sharing options and a footer with contact and service information.

Afb. 7.2 - Beperkt aantal termen bij dezelfde publicatie in de catalogus van Bibliotheek.nl.

7.4 Tagging en folksonomies

Toen rond 2005 onder de noemer Web 2.0 de mogelijkheden voor gebruikers groeiden om zelf eenvoudig content toe te voegen aan een website, ontstonden ook nieuwe technieken om de toegevoegde content snel en eenvoudig toegankelijk te maken. Vrijwel elke dienst waarbij gebruikers informatie plaatsen, biedt de gebruikers de mogelijkheid tags aan informatie of documenten toe te voegen. Dat kan bij eigen materiaal dat we op internet aanbieden, zoals afleveringen van de eigen weblog of foto's, maar ook bij boektitels die zijn verzameld op sites als Librarything. Deze tags hebben over het algemeen dezelfde functie als termen uit een informatietaal. Ze drukken meestal uit waar de informatie over gaat en ze vormen een hulpinstrument bij het terugvinden van de informatie of het document. Ook worden de toegekende tags gebruikt om suggesties

tags

te doen aan gebruikers: tags die veel samen worden toegekend worden door het systeem gekoppeld en wie zoekt op een tag krijgt suggesties over documenten met verwante tags. Bij het toekennen van tags zullen de meeste gebruikers (uiteraard) niet te werk gaan op de planmatige en systematische wijze die we eerder in dit hoofdstuk beschreven. Toch zal iemand ook hierbij, al is het onbewust, een inhoudsanalyse maken van het object dat hij met tags wil karakteriseren. Het persoonlijke, veelal abstracte beeld dat hij zich van dat object heeft gevormd, zal hij vervolgens ook vertalen naar een aantal toe te kennen tags, of je dat nu een informatietaal wilt noemen of niet. Daarbij zal hij vrijwel altijd afzonderlijke tags toekennen voor de verschillende elementen en invalshoeken van de beschreven objecten, dus postcoördinatief te werk gaan. Eén van de verschillen met officiële informatietalen, is dat tagging zich niet beperkt tot inhoudelijke elementen, maar dat formele kenmerken van objecten ook vaak in tags tot uiting worden gebracht. Dit komt onder andere omdat er vaak maar één veld voor het invullen van tags wordt aangeboden. Er wordt dus geen onderscheid gemaakt in verschillende typen tags (formeel, inhoudelijk), laat staan in de verschillende facetten die kunnen worden uitgedrukt.

Deze mogelijkheid tot het vermengen van verschillende typen tags is meteen ook een zwakke kant van deze methodiek. Die kan namelijk aanleiding geven tot zeer slechte precisie van zoekresultaten. Wie in Flickr foto's zoekt van berglandschappen met de zoekterm 'bergen' krijgt, weliswaar mooie, foto's van de Noorse stad Bergen te zien maar ook afbeeldingen die zijn gemaakt door fotografen die Bergen in hun achternaam hebben en natuurlijk ook foto's van bergen over de hele wereld.

**ongecontroleerde
trefwoorden**

folksonomies

Bovendien zijn tags in feite gewoon ongecontroleerde trefwoorden. Er zijn geen afspraken over enkel- of meervoud, over gebruik van zelfstandig naamwoorden of werkwoorden of over synoniemen. Ze hebben dan ook last van alle nadelen die aan ongecontroleerde vormen van toegankelijk maken kleven (Macgregor, 2006). Ze hebben echter het grote voordeel dat ze berusten op het woordgebruik van de informatiegebruikers zelf. Dit soort systemen wordt daarom ook wel aangeduid als folksonomies, een in 2004 door Thomas Vander Wal bedachte samentrekking van de woorden *folks* (mensen) en *taxonomy*, dus door *folks* zelf opgezette taxonomieën (Vander Wal, 2004). In het Engelse taalgebied spreekt men in dit verband ook wel van *grass-roots* taxonomieën. Daarbij wordt de gebruiker geacht zelf het best te weten waarop hij voor zijn onderwerp moet zoeken. Bovendien ziet men tagging dan als hulpmiddel bij het samenwerken van mensen, ook een belangrijk kenmerk van Web 2.0. Het taggen van materiaal kan dan helpen om virtuele gemeenschappen te laten ontstaan van mensen met gelijke interesse en daardoor hopelijk gelijksoortig woordgebruik. In dit kader kunnen we ook onderscheid maken tussen persoonlijk georiënteerde en sociale tags. De eerste soort gaat inderdaad

alleen uit van het persoonlijke gezichtspunt ('Oom Jan') of het persoonlijke oordeel ('interessant', 'nog lezen'), zodat het eigenlijk alleen betekenis heeft voor degene die de tag heeft toegekend. Met de tweede soort, hopen gebruikers juist anderen te bereiken en te interesseren voor het eigen (of andermans) product, of dat nu een blogpost, een gevonden wetenschappelijk artikel, een favoriet boek of een zelf op YouTube geplaatste video is.

Dit betekent dat tagging in feite voor een ander soort toegankelijkheid is bedoeld dan klassieke ontsluiting. Dat het tot slechte vangst en precisie kan leiden, speelt hierbij een veel minder belangrijke rol dan bij het zoeken in een professionele omgeving. Tags zijn meer bedoeld als ondersteuning bij het browsen. Gebruikers neigen ook meer af te gaan op suggesties die door collega's en vrienden door middel van tags zijn gedaan, dan dat ze op eigen initiatief diepgravend naar informatie zoeken. *Tag clouds* (woordenwolken) laten bij dergelijke systemen ook grafisch zien welke tags het meest frequent en/of het meest recent zijn gebruikt: hoe groter de letters hoe populairder de woorden. Zo vormen tags ook een hulpmiddel voor aanbevelingen. Analyse van toegekende tags kan daarbij helpen. Andere gebruikers die veelvuldig dezelfde tags toekennen als jijzelf, worden in dat verband wel aangeduid als *tag buddies* (Peters, 2009). Het is waarschijnlijk dat ze dezelfde interesse en dezelfde smaak hebben. De door hen getagde items kunnen dus ook voor jou interessant zijn. Je kunt zelfs zeggen dat ze met hun tags indirect hun eigen interesse inhoudelijk toegankelijk maken. Deze manier van analyseren van toegekende tags lijkt in feite sterk op wat onder meer Amazon doet op basis van koopgedrag van zijn klanten.

Voor kritische toepassingen in een bedrijfsomgeving waar men zich niet kan veroorloven om met een zoekactie relevante documenten te missen, zul je niet kunnen volstaan met tagging als enige hulpmiddel om die documenten vindbaar te maken (Alexander, 2009). In die omgeving ontbreekt ook meestal de tijd om de vele niet-relevante documenten te moeten doornemen die het resultaat zijn van de slechte precisie van een zoekresultaat. In gevallen waar geen menskracht beschikbaar is voor uitgebreide en gecontroleerde vormen van indexering, kan tagging echter een zinnig alternatief zijn. Dit geldt zeker ook voor niet-tekst materiaal, waar een gewone zoekmachine geen woorden aan de documenten zelf kan ontlenen. Zo zijn er steeds meer archieven die gedigitaliseerde foto's op internet zetten, met het verzoek aan bezoekers van de site om onbekende details toe te voegen. Ditzelfde geldt ook voor gescande tekstdocumenten waarvan de kwaliteit te slecht is om ze met optische karakterherkenning computerleesbaar te maken, zoals bij handschrift en oud drukwerk. Een voorbeeld hiervan is de website velehanden.nl

Aanbieders van Web 2.0 diensten blijken de nadelen van ongecontroleerde ontsluiting ook wel te onderkennen. Daarom ondersteunen veel systemen

hashtag

het tag-proces, door lijstjes te presenteren van eerder al toegekende tags; hetzij door dezelfde gebruiker voor ander materiaal, hetzij door andere gebruikers voor hetzelfde materiaal. Zo hoopt men gebruikers tot meer consistent tag-gedrag te verleiden. Anderzijds worden bij specifieke gebeurtenissen, of dat nu een specialistisch congres met slechts honderd deelnemers is, of een wereldwijd nieuwsfeit, wel vrijwillige afspraken gemaakt om een standaard *hashtag* te gebruiken. Daarmee kunnen deelnemers en belangstellenden dan al het materiaal taggen dat op die gebeurtenis betrekking heeft, op welke site dat zich ook bevindt.

Systemen met tags worden wel onderscheiden in nauwe en brede (Engels: *narrow / broad folksonomies* (Peters, 2009)). In een nauwe *folksonomy* kent de maker of eigenaar van een op internet geplaatst object daar alleen zelf tags aan toe. Een voorbeeld van een dergelijk systeem is Youtube. In dat geval is dus zeker geen sprake van “social” tagging. In sommige van deze gevallen is het mogelijk dat andere geregistreerde gebruikers die tags nog verder aanvullen. In dat geval spreekt men wel van een *extended narrow folksonomy*. Ook dan wordt een bepaalde tag toch maar één keer aan een object toegekend. Een voorbeeld van deze categorie is de fotosite Flickr. Bij beide nauwe varianten staat dus het object centraal.

Bij een brede folksonomy kan elke deelnemer vrij zijn eigen tags aan elk object toegekennen. Daar staat dus de gebruiker centraal. Het systeem registreert elke tag van elke gebruiker afzonderlijk, ook als vijftig verschillende gebruikers allemaal dezelfde tag aan een object toegekennen. Dat biedt dus de mogelijkheid om te zien welke termen het meest aan een bepaald object zijn toegekend, dus wat vermoedelijk het centrale onderwerp of de belangrijkste invalshoek is. Ook biedt dit meer vrijheid dan bij een nauwe folksonomy om ook andere invalshoeken gerepresenteerd te krijgen dan alleen de misschien wat eenzijdige blik van de oorspronkelijke eigenaar van een object. Voorbeelden van brede folksonomies zijn social bookmarking sites als Delicious en Connotea en de sociale catalogiseersite voor boeken LibraryThing.

Hoewel het verschil tussen deze twee manieren van taggen in de praktijk misschien van ondergeschikt belang lijkt, maakt het wel degelijk verschil voor de te bieden functionaliteit bij het zoeken en het presenteren van resultaten. Welk soort statistieken je van veel gebruikte en populaire tags kunt afleiden, is bijvoorbeeld nogal verschillend. Ook bij analyse van gebruikte tags, bijvoorbeeld voor het automatisch afleiden van semantische relaties daartussen, kan het nogal wat uitmaken of je met een brede of een nauwe folksonomy te maken hebt.

Onderzoek over tagging beperkte zich aanvankelijk vooral tot kwaliteitsanalyses: in hoeverre eenduidige termen worden gebruikt, in hoeverre er problemen zijn met synonymie, in hoeverre de tags vooral bedoeld zijn voor persoonlijk (her)gebruik en niet gericht op

publiekelijke bruikbaarheid, of voldoende tags zijn gebruikt om alle aspecten van een document, een foto of een video te beschrijven, dus wat de indexeerdiepte is, enzovoort (Macgregor, 2006; Golder, 2006). Vaak leidt dat tot de conclusie dat tagging meer een aanvulling is op klassieke ontsluiting, dan een vervanging daarvan (Guy, 2006). Ook wordt wel gekeken naar de verdeling van de frequentie van toekenning van tags. Die blijkt meestal een soort Zipf-verdeling te volgen: een beperkt aantal zeer veel gebruikte tags en een lange staart (Engels: *long tail*) van heel weinig of misschien zelfs maar eenmalig toegekende tags. Automatische mechanismen om de betekenis van tags en mogelijke semantische relaties tussen tags af te leiden, zijn vaak gebaseerd



Afb. 7.3 - Semantisch geordende tagcloud uit Web 2.0 boekcatalogus LibraryThing (<http://www.librarything.com/>; Wartena, 2010).

op statistische analyse van aan dezelfde objecten toegekende tags, vaak nog verfijnd door analyse van het daarbij samen voorkomen (Engels: *co-occurrence*) van verschillende tags. Afbeelding 7.3 geeft een voorbeeld van een op basis van dergelijke technieken geordende tagcloud van in LibraryThing aan boeken toegekende tags (Wartena, 2010). De in de woordenwolk getoonde termen zijn het resultaat van een statistische analyse van de tags die iets met het begrip 'russian literature' te maken hebben. Daarbinnen zijn de termen geclusterd op basis van een *co-occurrence* analyse. De termen op dezelfde regel hebben onderling meer met elkaar te maken dan met termen op andere regels. Toepassing van dergelijke technieken kan gebruikers bij hun zoekproces ondersteunen. Het inperken of uitbreiden van een zoekactie met enkele van de gepresenteerde tags, kan de precisie en (vooral) de vangst van zoekacties in zo ontsloten materiaal sterk verbeteren. Op die manier kunnen systemen proberen om, met behoud van de vrijheid van degene die tags toekent, toch een deel van de nadelen van een folksonomy te compenseren.

7.5 Samenvattend

We hebben in dit hoofdstuk gezien dat het goed is om je bij het indexeren bewust te zijn van een aantal stappen en keuzes, zoals de inhoudsanalyse, de keuze voor de indexeerdiepte en de specificiteit van de informatietaal. Ook hebben we gezien dat tagging een mooie aanvulling is op de verschillende instrumenten voor het toegankelijk maken van documenten, maar dat juist daarbij deze aspecten minder gestructureerd aan de orde komen. Ook als er wel gestructureerd gewerkt wordt, kunnen er uiteraard nog fouten ontstaan, bijvoorbeeld doordat de indexeerder een onderwerp niet herkent of verkeerd interpreteert. Ook kan een indexeerder andere termen toekennen dan die waarop een gebruiker zou zoeken. Hoewel “de enig goede indexering” niet bestaat, is er wel een aantal checks te geven. Iedere indexeerder kan deze voor zichzelf toepassen en op grond daarvan bepalen of zijn indexering voldoende is:

- Voldoet de indexeerdiepte (met name de uitputtendheid) aan de algemeen geformuleerde normen van de organisatie en/of collectie: zijn niet veel minder of veel meer termen toegekend dan afgesproken?
- Zijn de gekozen termen specifiek genoeg gekozen: is de volledige diepte van de informatietaal gebruikt (indien aanwezig)?
- Is de juiste term of combinatie van termen gebruikt om het onderwerp te beschrijven? In een organisatie maakt men voor deelgebieden vaak afspraken voor vaak-voorkomende of langlopende onderwerpen.
- Zijn alle onderwerpsaspecten van het onderwerp beschreven?

Wie bovenstaande regels bij het analyseren van het materiaal en het toekennen van termen gebruikt, verrijkt het materiaal waardoor het beter toegankelijk wordt en voor de zoekers beter terug te vinden.

Gebruikte literatuur:

Alexander, F. (april 2009). Folksonomies: business use. In *FUMSI*. Geraadpleegd op: 30 april 2010, van: <http://web.fumsi.com/go/article/manage/3791>

Berelson, Bernard (1952). *Content analysis in communication research*. New York: Free Press

Enser, P. (2008). *The evolution of visual information retrieval*. *Journal of Information Science*; 34(4), 531-546. DOI: 10.1177/0165551508091013

Golder, S., & Huberman, B.A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 198-208. Geraadpleegd op: 12 april 2010, van: <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>

Guy, M., & Tonkin, E. (2006) Folksonomies: tidying up tags? *D-Lib Magazine*, 12(1). Geraadpleegd op: 12 april 2010, van: <http://dlib.org/dlib/january06/guy/01guy.html>

IPTC (2010). *International Press Communication Council*. Geraadpleegd op 25 april 2010, van: <http://www.iptc.org>

Lancaster, F.W. (2003). *Indexing and abstracting in theory and practice*. Londen: Facet

Macgregor, G., & McCulluch, E. (2006). Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review*, 55(5), 291-300. Geraadpleegd op 13 april 2010, van: <http://strathprints.strath.ac.uk/2335/1/strathprints002335.pdf>

Panofsky, E. (1962). *Studies in iconology : humanistic themes in the art of the Renaissance*. New York [etc.]: Harper & Row

Peters, I. (2009). *Folksonomies: indexing and retrieval in web 2.0*. Berlin: De Gruyter/Saur. ISBN 978-3-598-25179-5.

RKD (2010). Iconclass. Geraadpleegd op: 1 april 2010, van: <http://www.iconclass.nl>

Sieverts, E.G. (29 juni 2009) **De mythe van de catalogus**. Geraadpleegd op 14 april 2010, van: <http://www.informatieprofessional.nl/bijdragen/7320-de-mythe-van-de-catalogus.html>

Soergel, D (1985). *Organizing information: principles of data base and retrieval systems*. Orlando: Academic Press

Vander Wal, T. (2004). Feed on this. In: *Weblog "Off the top"*, 3 October 2004. Geraadpleegd op: 30 maart 2010, van: <http://www.vanderwal.net/random/entrysel.php?blog=1562>

Wartena, C. (2010). Zoeken met tags; tagvarianten semantisch getemd. *InformatieProfessional*, 14(3), 20-23

KADER

Beeldmateriaal

Terwijl veel erfgoedinstellingen een inhaalslag maken door bestaande collecties met beeldmateriaal te digitaliseren, worden er iedere dag honderden miljoenen nieuwe foto's en vele uren video gemaakt en op sites geplaatst. Evenals tekst vraagt beeldmateriaal om een goede toegankelijkheid; niet alleen door te kunnen zoeken op fotograaf of datum van creatie, maar ook door inhoudelijke aspecten te gebruiken in het proces van zoeken en selecteren. Ten opzichte van tekstmateriaal is er sprake van een extra uitdaging: terwijl zoekmachines er steeds beter in slagen om tekstdocumenten te analyseren en indexeren, kunnen ze afbeeldingen of bewegend beeld niet zo gemakkelijk herkennen. De afbeeldingzoekers van de bekende zoeksoftware richten zich in eerste instantie dan ook op de tekst bij een foto: de documentnaam en de begeleidende tekst kunnen goede indicaties zijn van hetgeen op de foto te zien is. Ook Youtube baseert zich op de titels en omschrijvingen van de geplaatste video's waarbij het tevens op een slimme manier de historie van de zoeker en gelijkgestemden betreft in het doen van aanbevelingen. Deze manier van werken kan ons aardig helpen. We vinden makkelijk plaatjes op het web en zijn snel tevreden met bijvoorbeeld een keus uit 100 plaatjes van een aap. Maar net als bij het zoeken in tekstdocumenten zijn er ook veel voorbeelden te geven van mismatches. Zo geeft een zoekactie op 'Eric Sieverts' bij Google afbeeldingen niet alleen resultaten met afbeeldingen van één van de auteurs, maar ook foto's van collega's uit het informatievak en zelfs van Donald Trump! De reden hiervoor ligt in het feit dat Eric Sieverts over verschillende personen en onderwerpen schrijft en collega's over hem. De zoekmachine maakt dat onderscheid lang niet altijd. Om gerichter te kunnen zoeken, ontkomen we er dus niet aan om inhoudelijke metadata toe te kennen aan de afbeeldingen.

Erfgoedinstellingen zijn van oudsher gewend om handmatige beschrijvingen te maken van hun objecten. Bij fotomateriaal gingen archieven daarbij vaak grofmazig te werk en beschreven ze het materiaal over het algemeen niet op documentniveau maar op collectieniveau (bijvoorbeeld: archief fotograaf X, periode 1910-1927). Sinds de opkomst van internet hebben veel instellingen grote delen van hun collectie gedigitaliseerd om deze via het web voor een groot publiek beschikbaar te stellen. Daarbij zagen zij zich ook voor de taak gesteld om al deze objecten van uitgebreidere beschrijvingen te voorzien. De meeste erfgoedinstellingen hanteren hierbij een metadata-standaard die is gebaseerd op de internationale Dublin Core. Bij de meeste velden in deze standaard gaat het om formele metadata als fotograaf, titel, auteursrechthouder, formaat, en dergelijke. Bij digitale foto's hoeft dit lang niet allemaal meer handmatig te worden ingevuld. Veel foto's die nu worden

gemaakt krijgen van de camera automatisch veel (technische) metadata mee in de vorm van Exif files. Deze metadata verschaffen fototechnische gegevens zoals gebruikte camera, sluitertijd, kleurbewerking enzovoort. Deze gegevens passen niet allemaal in de Dublin Core standaard, maar kunnen in bepaalde gevallen zeker interessant zijn om te presenteren en doorzoekbaar te maken.

Dublin Core komt voort uit het erfgoed en wordt daar het meeste gebruikt. De International Press Telecommunications Council en de Newspaper Association of America (NAA) hebben hun eigen standaard, de IPTC standaard. Ook deze bevat een aantal velden waarin formele en inhoudelijke metadata kunnen worden ingevuld, bijvoorbeeld met behulp van programma's als Irfanview. Het mag duidelijk zijn dat deze standaard vooral gericht is op dagelijks gebruik in een journalistieke omgeving.

Zowel Dublin Core als IPTC bevat naast een aantal formele velden ook velden die over de inhoud van een afbeelding gaan. In hoofdstuk 7 is te lezen dat we kunnen indexeren op verschillende dieptes en abstractieniveaus. Dat geldt voor beeldmateriaal in sterke mate. Wie bijgaande afbeelding van een spelend kind beschrijft, kan bijvoorbeeld kiezen voor de termen 'kind', 'speeltunnel'. Concreter wordt het als de naam van het kind en de speeltuin worden genoemd: 'Anouk', 'Het Kinderparadijs'. Maar we kunnen er ook voor kiezen om met name de symboliek die de foto wellicht wil uitdrukken te beschrijven in termen als: 'jeugd', 'opening', 'weg naar vrijheid'.



Afb. 1 - CC BY: Pedro Ribeiro Simões

Deze verschillende niveaus van beschrijven zijn door de kunsthistoricus Edwin Panofsky (1962) als volgt beschreven:

1. Pre-iconografisch niveau: identificeren van vormen, objecten en hun onderlinge relatie: wat zie ik? In ons voorbeeld zouden dat termen kunnen zijn als 'kind', 'tunnel',
2. Iconografisch niveau: interpretatie en identificatie van afgebeelde scènes: wat weet ik? Hier noemen we dan de namen en context van het afgebeelde: 'Anouk in de speeltunnel in speeltuin Het Kinderparadijs tijdens het kinderfeestje voor haar 6e verjaardag'
3. Iconologisch niveau: de afbeelding is een symbool van iets en dient op de toeschouwer een beeld over te brengen: wat is er mee bedoeld? Hier passen dus termen als 'jeugd' en 'vrijheid'.

Overigens wordt dit laatste niveau ook wel in tweeën gesplitst waarbij onderscheid wordt gemaakt tussen de bedoeling van de maker en de persoonlijke interpretatie van degene die het bekijkt. Het derde hierboven beschreven niveau heet dan de iconografische interpretatie: hierin is de bedoeling van de maker vervat. Een vierde niveau is dan het iconologisch niveau. Hierin stellen we de vraag naar de diepere, cultuurhistorische betekenis die niet expliciet door de maker bedoeld hoeft te zijn. Bij veel hedendaags beeldmateriaal zoals foto's zien we meestal beschrijvingen terug op pre-iconografisch en iconografisch niveau. De indexeerder zal zich meestal houden aan interpretaties op deze eerste twee niveaus.

Het derde (en het eventuele vierde) niveau zien we terug in de kunstwereld waar de symboliek een belangrijke rol speelt in de totstandkoming van een kunstwerk. Dit moet dan in de beschrijving ervan terugkomen. Sommigen mensen die hun foto's op Flickr plaatsen kennen ook tags op dit niveau toe. Tenslotte zien we deze toepassing vaak terug in vlotte presentaties waarbij afbeeldingen metaforisch worden gebruikt.

In hoofdstuk 1 maakten we het onderscheid tussen ofness en aboutness van documenten. Voormalig onderzoeker James M. Turner paste deze concepten toe op beeldmateriaal. Hij noemde de fysiek inhoudelijke beschrijving (wat staat er op het plaatje?) de ofness van de afbeelding. De aboutness geeft aan waar de afbeelding over gaat. Vertalend naar Panofsky is de ofness ongeveer gelijk te stellen aan de pre-iconografische betekenis, terwijl de aboutness meer overeenkomsten heeft met het iconografisch niveau.

Naast de indeling van Panofsky zien we ook Enser's concepten om afbeeldingen te lezen regelmatig terugkomen (Enser, 2008). Peter Enser, voorheen werkzaam aan de University of Brighton onderscheidt twee soorten termen: unieke en niet-unieke. Unieke termen zijn dan termen die verwijzen naar één specifiek object, bijvoorbeeld 'The Medway Queen' (een stoomradarboot). De niet-unieke variant hiervan is bijvoorbeeld 'stoomradarboot'.

De verschillende interpretatieniveau's zoals hierboven beschreven worden in onderstaande tabel naast elkaar gezet. Hieruit blijkt ook dat alleen de indeling van Panofsky ingaat op de meer abstracte beschrijving van beelden.

Panofsky	Enser	Turner
Pre-iconografisch niveau	niet uniek	ofness
Iconografisch niveau	uniek	aboutness
Iconologisch niveau		

De analysemethode voor tekst en beeld verschilt, maar in beide gevallen geldt dat de analysefase een inhoudskarakteristiek moet opleveren: een omschrijving van het onderwerp, waaruit blijkt waar het document (tekst, beeld, video, geluid) over gaat.

Voor het kiezen van de juiste term, op welk niveau ook, kan gebruikt worden gemaakt van een informatietaal zoals de Arts and Architecture Thesaurus (AAT) die ook in het Nederlands beschikbaar is (zie hoofdstuk 13).

Naast de inhoudelijke beschrijving kan ook de GPS-locatie inhoudelijke informatie verschaffen. De GEO-data die camera's automatisch aan een foto meegeven, maakt het bijvoorbeeld makkelijk om het beeld te identificeren en via een kaart toegankelijk te maken. Wie op de kaart een locatie aanklikt, krijgt direct alle foto's te zien die daar ooit gemaakt zijn. In combinatie met het datumveld kan er gefilterd worden op foto's van een bepaalde locatie op een bepaald tijdstip, een door erfgoedinstellingen veelgebruikte methode om een locatie door de jaren heen te presenteren.

Commerciële aanbieders van afbeeldingen werken over het algemeen met grove doel waarvoor deze collecties gebruikt worden: foto's vinden die iets uitdrukken op het pre-iconografisch of iconologisch niveau. Het gaat er dan niet om om bijvoorbeeld een specifiek paard, gebouw of persoon te vinden, als het maar een afbeelding van een paard, gebouw in het algemeen of een specifieke situatie is.

De toepassing van Dublin Core en de indeling van Panofsky zijn niet alleen toe te passen op afbeeldingen, ook bewegend beeld kan volgens deze afspraken worden vastgelegd. Het probleem daarbij is vaak dat er zoveel beelden voorbijkomen, dat het ondoenlijk is om deze allemaal te beschrijven. Over het algemeen worden video's dan ook beschouwd als één document waaraan een beperkt aantal termen wordt toegevoegd. Alleen als het echt nodig is, worden de verschillende onderdelen door middel

van tijdcodes apart ontsloten. Om een video uitgebreid te beschrijven, kan ook de eventueel aanwezige ondertiteling als hulpmiddel worden ingezet. Dit tekstbestand kan door een zoekmachine worden geïndexeerd waardoor al het besprokene terugvindbaar is. Bij Nederlandse producties kan daarvoor de ondertiteling van pagina 888 van Teletekst worden gebruikt. Als de ondertiteling ontbreekt, kan nog spraakherkenning worden ingezet, zoals Youtube dat doet (waarbij de ondertiteling van Youtube overigens niet doorzoekbaar is). Uiteraard gelden hier de beperkingen van een zoekmachine zoals het gebrek aan context en het ontbreken van doorverwijzingen bij synoniemen. Voor het gericht toekennen van metadata heeft Het Nederlands Instituut voor Beeld en Geluid samen met enkele audiovisuele archieven de Gemeenschappelijke Thesaurus voor Audiovisuele Archieven (<http://gtaa.beeldengeluid.nl/>) ontwikkeld. Deze lijst bevat de voorgeschreven termen, ingedeeld in categorieën. Met name bij nieuwsitems moet er snel en gericht gezocht kunnen worden en dan moet de informatietaal uitsluitend geven of bijvoorbeeld de term 'ongeluk' of 'ongeval' wordt gehanteerd.



Afb. 2 - Afbeeldingen en andere erfgoed informatie van de binnenstad van Amersfoort zichtbaar gemaakt op www.Amersfoortopdekaart.nl

Sommige instellingen nemen hun toevlucht tot *crowdsourcing*: bezoekers van een website kunnen daarbij zelf informatie toevoegen aan gepresenteerde foto's. Zo heeft de Library of Congress veel foto's op de foto-site Flickr geplaatst om ze te laten verrijken met metadata van bezoekers. Op deze manier worden niet alleen kosten bespaard, maar wordt er vooral ook gebruikt gemaakt van externe kennis. Gebruikers kunnen bijvoorbeeld een oud gebouw of persoon herkennen en aanvullende informatie verschaffen.

In hoofdstuk 10 bespreken we vormen van geautomatiseerde ontsluiting van documenten.

Voor beeldmateriaal bestaan dat soort toepassingen ook; een beschrijving van de huidige stand van zaken vind je in het kader "Automatische beeldherkenning".

Gebruikte literatuur

Enser, P. (2008). The evolution of visual information retrieval. *Journal of Information science*; 34(4), p. 531-546. DOI: 101177/0165551508091013.

Panofsky, E. (1962). *Studies in iconology : humanistic themes in the art of the renaissance*. New York etc.: Harper&Row.

KADER

Wordpress

Wordpress is een veel gebruikte tool om blogs en websites te bouwen. Oorspronkelijk was het uitsluitend gericht op weblogs, maar de laatste jaren zijn de mogelijkheden voor het bouwen en onderhouden van websites enorm uitgebreid. Omdat de basisfunctionaliteiten gratis zijn en de tool in de cloud kan worden gebruikt, is het uitgegroeid tot één van de grootste contentmanagementsystemen van de wereld.

Wie een account heeft, kan zowel pagina's voor een website als *blogposts* aanmaken. Om de *blogposts* overzichtelijk te houden, biedt Wordpress de mogelijkheid om deze te groeperen in categorieën en tags. De eerste optie is bedoeld om de *posts* te groeperen op algemene onderwerpen. Bijzonder is dat *posts* in meer dan één categorie kunnen worden geplaatst, al is het natuurlijk aan te raden om dit aantal beperkt te houden omdat de categorieën anders weinig onderscheidend meer zijn. Tags zijn bedoeld om de *posts* van meer specifieke termen te voorzien. Ook hierbij kan de auteur van de post zelf bepalen hoever hij daarin wil gaan. Bij het kiezen van de categorieën is er sprake van enige controle op de invoer: de schrijver krijgt de beschikbare categorieën gepresenteerd om daar één of meer keuzes uit te maken. Als de gewenste categorie nog niet bestaat, kan eventueel een nieuwe worden toegevoegd.

Bij het toekennen van de tags is er sprake van vrije invoer. Dat maakt het geheel rommeliger: de gebruiker kan op verschillende momenten verschillende tags invullen voor hetzelfde onderwerp. Spelfouten worden gedeeltelijk voorkomen doordat de gebruiker suggesties krijgt voorgeschoteld voor tags zodra hij begint te tikken. Deze suggesties zijn gebaseerd op de al gebruikte tags in Wordpress. Via verschillende *widgets* kunnen de toegekende tags van alle *posts* als een *tagcloud* worden gepresenteerd aan de lezer. Deze kan vervolgens een tag aanklikken om alle *posts* met de betreffende tag te vinden. De tags hebben bij Wordpress nog een tweede functie: zoekmachines gebruiken de tags om de inhoud van de *posts* te identificeren. Wie slim gebruik maakt van de tags kan deze dus inzetten voor *search engine optimization* (zie kader "SEO" en SEA").

In de standaardfunctionaliteit van Wordpress is het niet mogelijk om categorieën en tags toe te kennen aan de webpagina's. Deze pagina's zijn als een statische website in een hiërarchische structuur in te delen. Via verschillende *plugins* is het mogelijk om de inhoud meer dynamisch te maken zodat pagina's op verschillende wijzen benaderd kunnen worden. Ook tags horen dan tot de mogelijkheden.

Hoofdstuk 8

SharePoint

SharePoint is bij veel organisaties al bijna net zo gewoon geworden als de andere Office-producten van Microsoft. In 2011 deed zo'n 80% van de Amerikaanse Fortune 500 organisaties "iets met SharePoint" (Low, 2011). Het zijn echter niet alleen grote organisaties die SharePoint gebruiken. Harde cijfers zijn er niet, maar naar alle waarschijnlijkheid heeft meer dan de helft van de Nederlandse organisaties SharePoint in een of andere vorm in gebruik, en dat aantal neemt naar verwachting de komende jaren verder toe.

SharePoint is een veelzijdige tool, die onder andere ingezet kan worden voor het samen werken aan en het delen van documenten. Omdat SharePoint flexibel is, met een keuze uit diverse functionaliteiten, is het breed inzetbaar. Maar daarmee is het ook een tool waarbij goed nagedacht moet worden over de functionele inrichting (welke functionaliteiten zijn er nodig en hoe moeten die ingezet worden). Dat nadenken over de functionele inrichting geldt ook voor het organiseren en vindbaar maken van documenten die binnen SharePoint opgeslagen zijn. SharePoint biedt diverse functionaliteiten zoals een fulltext zoekmachine, bibliotheken, sites, kolommen, inhoudstypen en (beheerde) metadata die helpen informatie goed vindbaar te maken. Het is echter de kunst om deze functionaliteiten zodanig in te zetten dat ze ook optimaal kunnen bijdragen aan die vindbaarheid. Voordat het daadwerkelijk functioneel inrichten binnen SharePoint begint, raden we aan goed na te denken over de informatie-architectuur. Die behelst voor SharePoint drie aspecten:

1. Site-architectuur; het bepalen van site-onderdelen per site en plaatsing van sites binnen een site verzameling.
2. Metadata-architectuur; het bepalen en inzetten van kolommen, termensets, inhoudstypen en mappen.
3. Pagina-architectuur; de inrichting van een pagina met tekst, afbeeldingen en webonderdelen.

Deze drie architectuuraspecten hangen nauw met elkaar samen. Zo bepaal je in de metadata-architectuur welke sites (zoals vastgelegd in je site-architectuur) je metadata gebruiken. En als je gebruik wil maken van metadata op een pagina moet je de pagina-architectuur en metadata-architectuur op elkaar afstemmen. Deze drie aspecten kunnen dus niet los van elkaar ontworpen worden, juist omdat ze zo afhankelijk zijn van elkaar. Idealiter worden ze in een SharePoint-traject dan ook gelijktijdig ontwikkeld.

In dit hoofdstuk gaan we achtereenvolgens op ieder van deze drie aspecten in.

8.1 Site-architectuur

De site-architectuur in SharePoint bestaat uit twee aspecten:

1. de aanwezige sites binnen een siteverzameling en de wijze waarop de verschillende sites zich tot elkaar verhouden;
2. de site onderdelen per site.

We behandelen achtereenvolgens deze twee aspecten.

8.1.1 Site-architectuur: sites binnen een siteverzameling

Vaak krijgt iedere afdeling en/of team een eigen SharePoint-site.

SharePoint site

Afhankelijk van waar een organisatie SharePoint voor wil gaan gebruiken, kunnen ook additionele sites aangemaakt worden voor een record center, document center en/of search center. Technisch gezien zijn er nauwelijks beperkingen en kun je zoveel sites maken als je wilt. Vaak zijn onderscheid in permissies of verschillen in site-opbouw (zie paragraaf 8.1.2) redenen om verschillende sites te maken.

siteverzameling

Eén of meer sites vormen samen een siteverzameling. Afhankelijk van hoe je sites binnen de siteverzameling plaatst, ontstaat een hiërarchische sitestructuur. Die hiërarchie heeft onder andere invloed op de rechtenstructuur. De hiërarchie van sites heeft echter ook invloed op de navigatiestructuur. In het kader van het vindbaar maken van informatie is het goed te weten dat je met verschillende sites ook verschillende zogenaamde contentsilo's creëert. De zoekmachine doorzoekt weliswaar alle documenten fulltext, maar er zijn ook gebruikers die liever navigeren naar de content die specifiek voor hen bedoeld is (zij prefereren een vorm van directe toegankelijkheid, zie paragraaf 1.3).

In hoofdstuk 3 hebben we gelezen dat er geen duidelijk maximum is aan het aantal niveaus van een site of siteverzameling. Wel is het belangrijk dat de gebruiker duidelijk overzicht houdt waar hij zich bevindt in de siteverzameling. De structuur moet dus helder zijn. Snellinkjes kunnen gebruikers snel naar onderliggende subsites brengen en ook daarbij is het belangrijk dat de gebruiker zich realiseert waar hij terechtkomt. Het is dus op zich geen probleem om verschillende contentsilo's te maken, maar deze moeten bijdragen aan een heldere site-architectuur. Naast sites, vormen ook onderdelen als bibliotheken, lijsten en mappen contentsilo's. Deze bespreken we in de volgende paragraaf.



Afb. 8.1 - Sites binnen een siteverzameling.

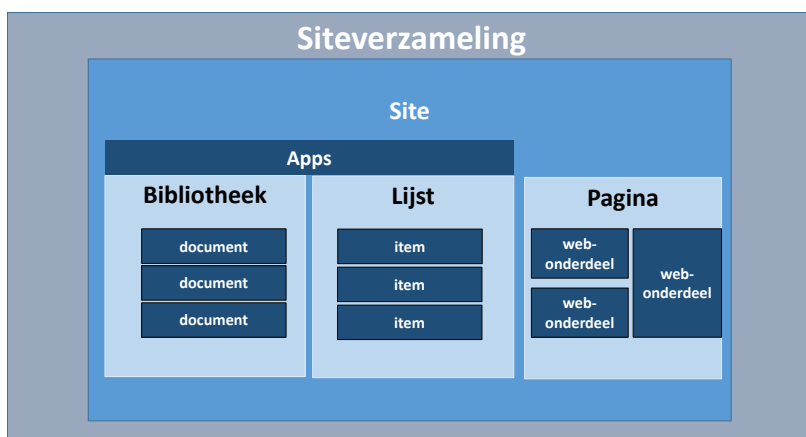
8.1.2 Site-architectuur: site-onderdelen per site

site-onderdelen Iedere afzonderlijke SharePoint-site bestaat op zijn beurt weer uit verschillende site-onderdelen. Een site-onderdeel kan zijn:

- bibliotheek**
 - een bibliotheek: bedoeld voor het opslaan van documenten, afbeeldingen en pagina's;
- lijst**
 - een lijst: bedoeld voor content waar niet noodzakelijk een document aan hoeft te hangen, bijvoorbeeld een takenlijst of agenda;
- pagina**
 - een pagina: bedoeld voor het presenteren van webcontent.

Lijsten en bibliotheken worden sinds SharePoint 2013 overigens ook aangeduid als apps.

sitetemplate Afhankelijk van het sitetemplate dat je kiest (of het nu een standaard SharePoint-template is, zoals een teamsite, een projectsite of een template dat je zelf hebt gemaakt) geeft SharePoint een aantal suggesties voor site-onderdelen. Vervolgens kun je weer andere site-onderdelen toevoegen of verwijderen, waarbij er geen limiet zit op het aantal site-onderdelen. Maar ook hierbij geldt weer: met iedere bibliotheek of lijst die je maakt, creëer je een silo. En die maken het lastiger voor de gebruiker die wil navigeren naar zijn documenten. Onze richtlijn is dan ook dat je één documentbibliotheek per site maakt. Uitzondering hierop zijn eventuele vertrouwelijke documenten. Hiervoor maak je een aparte bibliotheek met andere rechten. Wat in de praktijk soms gebeurt, maar wij sterk afraden, is het maken van verschillende bibliotheken voor verschillende onderwerpen, om zo binnen één site documenten over hetzelfde onderwerp bij elkaar te zetten. Hiermee creëer je namelijk weer een silo. SharePoint heeft daar bovendien betere tooling voor: (managed) metadata en/of inhoudstypen. Hier gaan we in de volgende paragraaf verder op in.



Afb. 8.2 - Site-onderdelen per site.

8.2 Metadata-architectuur

metadata-architectuur

Onder metadata-architectuur verstaan we in SharePoint het ontwerp van de verschillende soorten metadata. Daarbij gaat het om:

- welke (soorten) metadata moeten worden gebruikt;
- op welke manier metadata moeten worden ingezet;
- welke gecontroleerde trefwoorden er (eventueel) moeten worden toegepast.

Wie gebruik maakt van gecontroleerde of ongecontroleerde metadata, faciliteert indirecte toegankelijkheid. In deze paragraaf beschrijven we de verschillende aspecten van deze indirecte toegankelijkheid binnen SharePoint: hoe werkt het gebruik van metadata binnen SharePoint en wat zijn de mogelijkheden van gecontroleerde en ongecontroleerde metadata.

8.2.1 Metadata: eigenschappen in kolommen

kolommen

eigenschappen

SharePoint gebruikt zelf niet het begrip metadata, maar spreekt van kolommen en eigenschappen. Een kolomnaam is bijvoorbeeld 'titel'. Als 'jaarverslag 2016' de titel is (de waarde van de kolom 'titel' noemt SharePoint dat), dan is dit één van de eigenschappen (in onze terminologie: formele metadata) van dat document.

Sommige kolommen worden standaard door SharePoint aangemaakt en een aantal daarvan wordt automatisch door SharePoint ingevuld. Voorbeelden zijn aanmaakdatum, wijzigingsdatum, auteur, bestandstype, bestandsnaam, bestandsgrootte etc. De bouwer van een site of siteverzameling kan er daarnaast voor kiezen om zelf kolommen aan te maken. Een gebruiker wordt daarmee gevraagd (of verplicht, die keuze is aan de beheerder) om extra metadata bij een document in te vullen. Bij het aanmaken van kolommen heeft de beheerder de keuze

uit verschillende typen kolommen, bijvoorbeeld numeriek, datum/tijd, keuzelijsten of een veld met vrije tekst. Afhankelijk van het type kolom, gedraagt de kolom zich anders. Zo kun je in een numeriek veld alleen cijfers opgeven en geen letters.

Een bouwer kan (afhankelijk van de rechten die hij heeft) kolommen op drie verschillende niveaus aanmaken: op bibliotheekniveau, op siteniveau of op siteverzameling-niveau. Als je een kolom binnen een bibliotheek aanmaakt, is die ook alleen te gebruiken binnen die specifieke bibliotheek. Wanneer je kolommen over verschillende bibliotheken binnen een site of over verschillende sites heen wilt kunnen hergebruiken, zul je de kolommen op siteniveau respectievelijk siteverzameling-niveau aan moeten maken. Vervolgens geef je per bibliotheek aan welke sitekolommen er voor die bibliotheek van toepassing zijn. SharePoint heeft al een aantal sitekolommen voorgedefinieerd, zoals auteur, aanmaakdatum en beschrijving. Ook Dublin Core kolommen, zoals dekking, taal en uitgever biedt SharePoint je aan als sitekolom. Daarnaast kun je dus zelf oneindig veel eigen kolommen maken.

8.2.2 Beheerde metadata

Hiervoor stelden we dat een beheerder bij het aanmaken van een kolom keuze heeft uit verschillende typen kolommen, waaronder keuzelijsten. Sinds SharePoint 2010 (en daarmee ook beschikbaar in latere versies) is daar een type kolom bijgekomen: de beheerde metadata. Bij het aanmaken van een kolom van dit type geef je aan uit welke beheerde metadata (termenset) de kolom zijn waarden moet halen. Een termenset is volgens onze definitie een informatietaal (zie paragraaf 1.5). Een termenset kan op verschillende manieren ingericht worden en afhankelijk van hoe je dat doet, krijg je een autorisatielijst (zie hoofdstuk 4), *descriptive taxonomy* (zie paragraaf 3.2), of thesonomie (zie hieronder).

beheerde
metadata

termenset

thesonomie

Het is dus afhankelijk van hoe je een termenset inricht, of je een *descriptive taxonomy*, autorisatielijst, of thesonomie krijgt. Als je de termen uit de termenset alfabetisch (dus zonder hiërarchie) ordent, ontstaat een autorisatielijst. Breng je hiërarchische relaties aan tussen termen, dan ontstaat een *descriptive taxonomy*. Indien gewenst kun je de termen van deze *descriptive taxonomy* voorzien van synoniemen (met een onderscheid in voorkeurs- en verwijstermen, ook geschikt voor afkortingen, acroniemen, quasi-synoniemen) en omschrijvingen bij termen (vergelijkbaar met *scope notes* uit een thesaurus). Het is echter niet mogelijk om associatieve relaties aan te geven. Een termenset die op deze manier (dus met hiërarchie, synoniemen en *scope notes*) is opgebouwd, houdt het midden tussen een taxonomie en een thesaurus (immers, associatieve relaties ontbreken) en daarom noemen we deze ook wel een thesonomie.

managed metadata store

Termensets worden beheerd in de managed metadata store, een apart gedeelte binnen SharePoint waarin de termsets gemaakt en beheerd worden. Termsets worden vanaf deze centrale plek beschikbaar voor alle kolommen in de SharePoint-omgeving. De managed metadata store kan zijn eigen beheerders hebben. In de praktijk zien we dat dit soms de functioneel en/of technisch beheerders van de SharePoint-omgeving zijn. Het is echter ten eerste aan te raden om het beheer van de managed metadata store over te dragen aan informatieprofessionals, omdat zij al beschikken over de kennis die nodig is voor opbouw en beheer van informatietalen.

De beheerder van de managed metadata store kan één, maar ook meerdere termsets maken. Zo kan hij ervoor kiezen om niet alleen een termset voor onderwerp, maar bijvoorbeeld ook voor afdelingen, producten, processen, geografische plaatsen etc. te maken. Je kunt dus een *descriptive taxonomy*, thesonomie en/of autorisatielijsten naast elkaar gebruiken (zie ook hoofdstuk 5). De verschillende termsets kunnen vervolgens ingezet worden als verschillende facetten.

Als een organisatie in SharePoint 2007 (of in een eerdere versie) de gebruiker wilde laten kiezen uit een lijst met vaste waarden, dan had hij alleen de mogelijkheid om een (platte) keuzelijst (zie paragraaf 4.1) te maken. Die keuzelijsten zijn er in de latere versies van SharePoint nog steeds, maar er is dus de mogelijkheid van managed metadata bij gekomen. Keuzelijsten blijven nog steeds prima inzetbaar bij een beperkt aantal waarden (stelregel: minder dan 10 waarden), wanneer je geen gebruik wilt of kunt maken van hiërarchie tussen termen en wanneer je geen behoefte hebt aan het opnemen van synoniemen. Heb je meer dan 10 waarden en wil je wel gebruik maken van hiërarchische relaties tussen termen en/of synoniemen, zet dan beheerde metadata in.

Hoewel de managed metadata store mogelijkheden biedt voor het verplaatsen, hergebruiken, afschaffen en importeren van termen, mag het geen thesaurussoftware genoemd worden. De belangrijkste reden daarvoor is dat de uniciteit van termen niet gegarandeerd wordt en we beschreven al in hoofdstuk 4 dat dat één van de foutcontroles is die goede thesaurussoftware uit moet kunnen voeren. Verder zit er geen controle op foute verwijzingen, zijn associatieve relaties niet mogelijk en kent de managed metadata store geen standaard exportfunctionaliteit. Wanneer een termset is opgebouwd of geïmporteerd, dan is het de bedoeling dat deze onderhouden wordt in SharePoint. De termset in verschillende systemen gebruiken is daarmee dus niet mogelijk. Importeren van termsets kan wel. Daarvoor is een aantal kant-en-klare termsets op de markt, waaronder de WAND General Business Taxonomy. Deze is gratis beschikbaar en kan na import nog naar wens aangepast worden.

ondernemings- trefwoorden

8.2.3 Ondernemingstrefwoorden

Ondernemingstrefwoorden is een enigszins verwarrend begrip binnen SharePoint. Het zijn vrije trefwoorden, woorden die een gebruiker naar eigen inzicht kan invullen. Daarmee lijkt het op de tags die we kennen van Web 2.0 sites (zie hoofdstuk 7). Indexeren gaat via een *autocomplete* functionaliteit; terwijl de gebruiker tikt worden suggesties gegeven uit alle (gecontroleerde) termensets en alle eerder door gebruikers toegekende (ongecontroleerde) ondernemingstrefwoorden.

Ondernemingstrefwoorden kunnen op zichzelf of in combinatie met andere termensets gebruikt worden. Je kunt het toekennen van termen aan documenten dus zowel gecontroleerd als ongecontroleerd laten doen. Het één sluit het ander niet uit. Sterker nog, het kan elkaar aanvullen en versterken. In de managed metadata store worden namelijk alle ondernemingstrefwoorden in een aparte termenset verzameld. Hier bestaat ook de mogelijkheid om ondernemingstrefwoorden te verplaatsen naar termensets. Dit is een interessante mogelijkheid, want je zou dus gebruikers vrije trefwoorden kunnen laten toekennen en op basis hiervan je termensets (door)ontwikkelen.

inhoudstypen

8.2.4 Inhoudstypen

Voor het aangeven van de isness van documenten (wat voor soort document het betreft, zie hoofdstuk 1) zijn er binnen SharePoint twee opties. De eerste is het opstellen van een termenset en de tweede is het toepassen van inhoudstypen. Bij de eerste optie wordt een termenset met termen die de documentsoort aangeven aangemaakt, zoals 'rapport' of 'memo'. Net als termen uit andere termensets kunnen deze verrijkt worden met synoniemen, hiërarchische relaties en/of omschrijvingen.

Maar de isness van documenten is typisch iets dat je binnen SharePoint ook goed naar inhoudstypen kunt vertalen. Want naast het aangeven wat voor documentsoort het betreft, kun je met inhoudstypen het "gedrag" van documenten sturen. Zo kun je aan inhoudstypen bewaaren vernietigingstermijnen koppelen, document(werk)stromen laten starten en/of documentsjablonen gebruiken. Maar ook kun je, en die mogelijkheid is in de context van dit boek interessant, met inhoudstypen onderscheid maken in metadata. Zo kun je voor een memo andere metadata vereisen/vragen dan voor een rapport. Voor een rapport is bijvoorbeeld de publicatiedatum als metadata wel relevant, maar voor een memo niet. Wanneer je gebruik maakt van inhoudstypen, kun je die zo inzetten dat een gebruiker de kolom 'publicatiedatum' niet ziet wanneer hij voor het inhoudstype 'memo' heeft gekozen. Groot voordeel hiervan is dat je gebruikers niet belast met het vragen naar metadata die niet relevant zijn.

8.2.5 Toekennen van gecontroleerde termen

SharePoint ondersteunt het toekennen van termen uit termensets op de volgende wijze:

- Gebruikers voegen termen toe via een *autocomplete* functionaliteit en/of via de hiërarchische indeling van de termenset (de hiërarchische boom wordt gepresenteerd).
- Bij de *autocomplete* optie worden voorkeurstermen en verwijstermen getoond, maar verwijstermen kunnen niet geselecteerd worden om toe te kennen (de voorkeursterm wordt altijd toegekend). Zie ook hoofdstuk 4 (onder autorisatielijsten).

Voor het toekennen van termen uit termensets aan documenten gelden alle regels voor het toekennen van termen uit een thesaurus zoals die in paragraaf 7.3 beschreven zijn. In hetzelfde hoofdstuk stelden we ook dat “niveau-indexering wordt toegepast in situaties waarin het retrievalsysteem de thesaurusfunctionaliteit niet goed kan ondersteunen”. Het is te kort door de bocht om te stellen dat het retrievalsysteem van SharePoint de thesaurusfunctionaliteit niet goed ondersteunt. Dat is namelijk sterk afhankelijk van de gemaakte keuzes bij de functionele inrichting van SharePoint. Die keuzemogelijkheden beschrijven we in paragraaf 8.2.6.

Automatisch toekennen van termen op basis van inhoudsanalyse (zoals beschreven in hoofdstuk 10) is niet out of the box mogelijk. Daarvoor heb je een add-on (specifiek nog extra aan te schaffen product) nodig. Maar er zijn wel trucjes te bedenken, bijvoorbeeld door op basis van de plek waar het document wordt opgeslagen metadata toe te kennen. Je geeft dan een standaardwaarde aan een kolom in een bibliotheek mee, zodat alle documenten in die bibliotheek de standaardwaarde (metadata) mee krijgen. Dit zou ook op basis van mappen kunnen (zie paragraaf 8.2.7).

8.2.6 Zoeken met termensets

Termensets kunnen gebruikt worden voor generiek zoeken, vervangend zoeken, filtermogelijkheden en facetten. SharePoint maakt net als sommige andere systemen bij de toepassing van een informatietaal het onderscheid in de toepassingsmogelijkheden bij fulltext zoeken en zoeken binnen bibliotheken en lijsten.

Voor fulltext zoeken door de gehele SharePoint omgeving, een site of een bibliotheek/lijst geldt dat termensets ingezet kunnen worden bij het facetzoeken (zie paragraaf 2.5). Afhankelijk van de gestelde zoekvraag kan een gebruiker zijn zoekresultaten inperken op basis van termen uit de verschillende termensets (de verschillende facetten). SharePoint-search kan zo ingericht worden dat de zoekvraag van een gebruiker wordt uitgebreid met synoniemen, maar daarbij worden niet de verwijstermen uit termensets gebruikt, maar termen uit een apart synoniemenbestand.

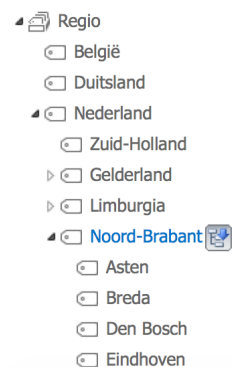
Dat betekent dus dat synoniemen op twee verschillende plekken bijgehouden moeten worden.

metagegevens navigatie- instellingen

Naast fulltext zoeken (waarbij Booleaanse operatoren ingezet kunnen worden), heeft een gebruiker ook de mogelijkheid om documenten binnen één bibliotheek te filteren door middel van de metagegevens navigatie-instellingen. Bij het activeren van deze functionaliteit krijg je twee opties voor filtering: hiërarchievelden of sleutelfiltervelden. Deze kunnen naast elkaar gebruikt worden en hebben ieder hun eigen mogelijkheden (en daarmee ook beperkingen). We leggen hieronder beide uit.

hiërarchieveld

In een hiërarchieveld is het mogelijk om generiek te zoeken. De hiërarchie van de volledige termenset(s) wordt getoond en wanneer een gebruiker kiest voor 'Noord-Brabant', worden de documenten waaraan de termen 'Noord-Brabant', maar ook 'Den Bosch' en 'Eindhoven' zijn toegekend, getoond. Een gebruiker kan er ook voor kiezen om alleen op de term 'Noord-Brabant' te filteren. Een gebruiker kan maar één term selecteren om te filteren, dus voor het tonen van alle documenten met de termen 'Noord-Brabant' en 'Limburg' moet de gebruiker twee afzonderlijke filteracties doen. Een hiërarchieveld kent geen mogelijkheid tot vervangend zoeken.

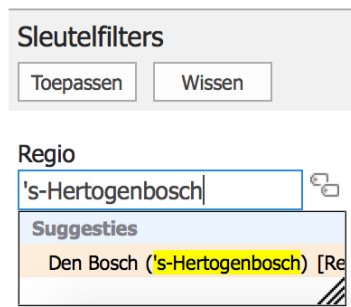


Afb. 8.3 - Hiërarchieveld.

Na activering van het sleutelfilterveld kan een gebruiker op twee manieren filteren. De eerste manier is via de hiërarchie van de volledige termenset, waarin termen geselecteerd kunnen worden. Bij de tweede manier kan een gebruiker zelf termen intikken, waarbij hij ondersteund wordt door een *autocomplete* functionaliteit. Hierbij wordt ook vervangend zoeken toegepast: verwijstermen worden wel getoond, maar ze kunnen niet geselecteerd worden om op te zoeken (de voorkeursterm vervangt automatisch de verwijstern).

sleutelfilterveld

Bij een sleutelfilterveld kan een gebruiker op meerdere termen (zowel 'Noord-Brabant' als 'Limburg') filteren. Maar generiek zoeken is in het sleutelfilterveld niet mogelijk: indien een gebruiker 'Noord-Brabant' kiest, worden ook alleen de documenten getoond met 'Noord-Brabant' als term (en niet 's-Hertogenbosch' of 'Eindhoven'). Bij keuze voor uitsluitend een sleutelfilterveld kan dus overwogen



Afb. 8.4 - Sleutelfilterveld.

SharePoint
mappenstructuur

worden om niveau-indexering toe te passen (zowel ‘Noord-Brabant’ als ‘Eindhoven’ toekennen) of om het hiërarchieveld aan te bieden.

De keuze voor een hiërarchie- en/of sleutelfilterveld is afhankelijk van de beslissingen die genomen worden bij de functionele inrichting. Het is echter belangrijk om te realiseren dat deze keuze bepaalt of je al dan niet moet niveau-indexeren om generiek zoeken mogelijk te maken.

8.2.7 Mappenstructuur

SharePoint kent, naast de mogelijkheden van (beheerde) metadata, de mogelijkheid om binnen bibliotheken documenten in mappen op te slaan. Dit werkt volgens hetzelfde idee als mappenstructuren van een netwerkschijf. De uitgangspunten van het opzetten van een hiërarchische indeling en mappenstructuur, zoals beschreven in hoofdstukken 3 en 6 gelden dus ook voor de SharePoint mappenstructuur. Het heeft daarmee ook dezelfde nadelen, zoals het feit dat documenten maar vanuit één dimensie benaderd kunnen worden. Naast de mappenstructuur kent SharePoint verschillende bibliotheken en sites, waarmee versnippering van informatie en een slechtere vindbaarheid op de loer ligt. Bovendien besluiten sommige organisaties om de bestaande netwerkstructuur een op een naar SharePoint over te zetten, inclusief verouderde en irrelevante documenten. De “veredelde netwerkschijf” is dan een begrip dat gebruikers noemen wanneer ze zien dat de SharePoint-omgeving niet méér is dan hun oude netwerkschijf in een nieuw jasje.

✓	📄	Naam		Gewijzigd	Gewijzigd door	Taal	Regio
▶ Regio : Breda (1)							
▶ Regio : Den Bosch (2)							
▶ Regio : Eindhoven (2)							
📄	Document 1	...	2 juni	☐	docent1	Engels	Eindhoven
📄	Document	...	2 juni	☐	docent1	Engels	Eindhoven
Sleep bestanden hiernaartoe om te uploaden							

Afb. 8.5 - Groepering van documenten op metadata.

De mappenstructuur weggooien, gaat sommige organisaties te ver. Gebruikers zijn van nature gewend aan het structureren en ordenen van documenten in mapjes. Er zijn echter binnen SharePoint mogelijkheden om gebruikers geleidelijk te laten wennen om meer (en uiteindelijk misschien wel uitsluitend) met metadata te werken. Zo bestaat er binnen bibliotheken de mogelijkheid om documenten te groeperen op kolommen.

Hiermee lijkt het voor een gebruiker of documenten in mapjes staan, maar eigenlijk zijn de documenten gesorteerd op metadata. Daarnaast raden we aan om te onderzoeken of op basis van mappen automatisch metadata toegevoegd kunnen worden. Dit zou zelfs helemaal onder water kunnen gebeuren, zodat een gebruiker geen extra kolommen te zien krijgt. Voordeel is wel dat deze metadata gebruikt kunnen worden in facetten van de zoekmachine.

8.3 Pagina-architectuur

pagina-architectuur

Na de site-architectuur en de metadata-architectuur richten we ons als laatste kort op de pagina-architectuur; de inrichting van webpagina's met onder andere tekst, afbeeldingen en webonderdelen.

Iedere site binnen SharePoint heeft minimaal één pagina: de introductiepagina van die site. Door middel van teksten en afbeeldingen maak je de pagina visueel aantrekkelijk. Webonderdelen fungeren op een pagina als een soort *widgets*: het zijn kleine blokjes die veelgebruikte informatie tonen. Die informatie kan van buiten SharePoint komen, zoals een buienradar of een RSS feed van een website. Maar webonderdelen kunnen ook informatie uit de SharePoint omgeving halen, bijvoorbeeld uit bibliotheken en lijsten van de onderliggende site. De introductiepagina fungeert dan als een soort etalage: het geeft doorkijkjes naar wat er allemaal binnen die betreffende site te vinden is. Daarbij kan een webonderdeel een deelselectie tonen van de volledige bibliotheek of lijst. Zo kun je bijvoorbeeld in een webonderdeel de laatste tien toegevoegde documenten van een bibliotheek laten zien, terwijl de complete bibliotheek alle documenten laat zien. Of een webonderdeel laat alleen de openstaande taken van de ingelogde gebruiker zien, terwijl de volledige takenlijst de taken van alle projectleden laat zien.

Bij de pagina-architectuur is het van belang om te zorgen voor consistentie in de pagina-lay-out. Daarnaast moet de belangrijkste informatie linksboven van de pagina staan (omdat een gebruiker daar meestal begint te lezen).

8.4 Samenvattend

Het goed inrichten van de informatie-architectuur in SharePoint, waar de metadata-architectuur een belangrijk deel van uitmaakt, is geen sinecure. Het is een proces waarover goed nagedacht moet worden, bij voorkeur voordat met de daadwerkelijk SharePoint-implementatie of doorontwikkeling gestart wordt. Hierbij is het van groot belang de gebruikers te betrekken. Door middel van workshops en interviews krijg je inzicht in hun wensen en behoeften.

Voor de metadata-architectuur organiseer je enkele workshops, eventueel aangevuld met interviews, waarin je een representatieve afvaardiging van de verschillende gebruikers vraagt hoe zij documenten willen opslaan en vinden. Uitkomst hiervan is een SharePoint-metadatamodel met bijbehorende termensets. Het metadatamodel geeft een overzicht van alle kolommen, met daarbij de relatie tot inhoudstypen, soort kolom (bijvoorbeeld tekst, numeriek, persoonsnaam of beheerde metadata), verplichtheid en in welke bibliotheken de kolom gebruikt wordt. Daarnaast stel je voor alle metadatavelden van het type managed metadata een termenset op. De werkwijze zoals we die beschreven hebben in paragraaf 3.3 (ontwikkelen van een taxonomie) geldt ook voor een SharePoint-taxonomie. *Cardsorting* (zie kader) is hierbij een techniek die zijn nut bewezen heeft.

Gebruikte literatuur

Low, L. (2011) *SharePoint 2010 : The First 10 Years*. Geraadpleegd op 13 mei 2016 van: <https://technet.microsoft.com/en-us/magazine/gg981684.aspx>

Verder lezen

Jamison, S. & Hanley, S & Cardarelli, M. (2010). *Essential SharePoint 2010: Overview, Governance, and Planning*. Upper Saddle River: Addison Wesley. ISBN 978-0-321-70075-9. Te downloaden van: <http://ptgmedia.pearsoncmg.com/images/9780321700759/samplepages/0321700759.pdf>

Jamison, S; Hanley, S; Bortlik. *Essential SharePoint 2013: Practical Guidance for Meaningful Business Results*. Upper Saddle River: Addison Wesley. ISBN 978-0321884114.
Te downloaden van: <http://ptgmedia.pearsoncmg.com/images/9780321884114/samplepages/0321884116.pdf>

Aalten, J. & de Vette, D. Lego voor organisaties. In: *Informatieprofessional*, 07, 2015

Hoofdstuk 9

Vormen van kennisrepresentatie

knowledge
organisation
systems

In de voorgaande hoofdstukken hebben we diverse vormen van inhoudelijk toegankelijk maken de revue laten passeren: thesauri, taxonomieën en classificaties. Naast deze op min of meer klassieke principes gebaseerde technieken, spelen ook andere begrippen een rol. Deels buiten de directe kring van bibliothecarissen en informatiespecialisten zijn de laatste jaren namelijk ook andere technieken ontwikkeld voor wat in het Engels wel *knowledge organisation systems* heten. Zowel in het Engels als in het Nederlands (kennisorganisatiesystemen) kun je dat tot KOS afkorten. Zoals deze benaming al aangeeft, hoeven die systemen niet altijd rechtstreeks bedoeld te zijn om informatie te ontsluiten, maar kunnen ze een ruimer doel hebben: het beschrijven of representeren van kennis. Hill e.a. (2002) onderscheiden de volgende soorten KOS'en:

- Systemen voor het rubriceren in categorieën, waaronder classificaties, *subject headings* en taxonomieën (zoals hier in de hoofdstukken 2 en 6 aan de orde komen);
- Metadata-achtige modellen (bijvoorbeeld naam-adres databases of structuren voor het registreren van formele gegevens over objecten; zie paragraaf 1.4);
- Relationale modellen, waaronder thesauri (hier in hoofdstuk 4), semantische netwerken en ontologieën (waarop we in dit hoofdstuk verder ingaan);
- Termlijsten, waaronder autorisatielijsten (zoals in hoofdstuk 4 en 5 genoemd) en woordenboeken.

interoperabiliteit

Een KOS kan verschillende functies vervullen. Zo kan een KOS een beschrijvende functie hebben, doordat het mogelijk maakt gecontroleerde labels aan objecten toe te kennen. Dat is bijvoorbeeld wat we met een thesaurus doen. In de tweede plaats kan een KOS definiërend zijn: het geeft betekenis aan die labels. Bij een thesaurus kan dat via *scope notes* gebeuren. In andere soorten KOS, zoals de in dit hoofdstuk besproken ontologieën, kan dat op nog veel gedetailleerder niveau. Ten derde kan een KOS via concordanties zorgen voor vertaling tussen verschillende kennisrepresentaties of ontsluitingssystemen. Dat betekent dat interoperabiliteit tussen verschillende ontsluitingssystemen, in het bijzonder het tegelijkertijd kunnen zoeken in de daarmee toegankelijk gemaakte collecties, eveneens tot het werkterrein van de KOS behoort. Tot slot kan een KOS ook een navigatiefunctie hebben via de systematisch gestructureerde wijze waarop de kennis daarin georganiseerd is. Die functie is belangrijk bij onder meer taxonomieën en classificaties, zoals

we eerder in dit boek al zagen. In het kader van KOS-onderzoek besteedt men ook aandacht aan methoden om gebruikers eenvoudiger toegang te bieden tot de vaak wat kunstmatige vocabulaires die voor gecontroleerde ontsluiting van informatie worden gebruikt. Een systeem kan zoektermen van gebruikers bijvoorbeeld vertalen naar gebruikte thesaurustermen of verwijzen naar de juiste categorieën in een taxonomie of classificatie. Zowel een vertaal- als een navigatiefunctie kunnen daar dus een rol spelen.

KOS'en die er uitsluitend op gericht zijn om kennis vast te leggen, dienen alleen als indirect hulpmiddel bij het toegankelijk maken van informatie. Zij helpen bij intelligent zoeken of bij het aan elkaar relateren van verschillende ontsluitingssystemen ten behoeve van de al genoemde interoperabiliteit van die systemen. Enkele van deze niet eerder besproken soorten kennisrepresentaties, semantische netwerken en ontologieën, komen in de volgende paragrafen kort aan de orde. Omdat het voor de diverse toepassingen essentieel is dat computers deze kennisrepresentaties kunnen lezen en interpreteren, zullen we in dit hoofdstuk ook ingaan op enkele gestandaardiseerde technieken om dat mogelijk te maken. Daarbij komen ook de kennisgrafen (Engels: *knowledge graphs*) aan de orde, die vooral in webzoekmachines een steeds belangrijker rol spelen.

9.1 Semantische netwerken

semantisch netwerk

betekenisrelaties

Een semantisch netwerk kan worden beschouwd als een digitale versie van een woordenboek, waarin niet alleen betekenissen van woorden, maar ook betekenisrelaties tussen die woorden zijn vastgelegd. Door te zeggen dat het een woordenboek is, geven we al aan dat in principe alle woorden van een taal er in opgenomen (kunnen) zijn. Dat betekent dat het geen beperkt en kunstmatig vocabulaire is zoals de ontsluitingssystemen die we tot nu toe in dit boek zijn tegengekomen. Anders dan in een thesaurus, hebben we dus niet te maken met voorkeurstermen.

semantische afstand

Door het leggen van relaties vormen al die woorden samen een netwerk, waarin elk woord met elk ander woord verbonden kan zijn, ook al is dat misschien via de tussenstappen van relaties met een heleboel tussenliggende woorden. Hoeveel tussenstappen daarvoor nodig zijn voor een willekeurig tweetal woorden, noemt men wel de semantische afstand van die twee woorden. Zo zal de semantische afstand tussen de woorden 'voetbal' en 'elftal' waarschijnlijk kleiner zijn dan die tussen 'voetbal' en 'viool'. Daarbij is het overigens een aardig spelletje te bedenken via welke tussenstappen ook die twee laatste woorden verbonden zouden kunnen zijn.

Dat een semantisch netwerk een woordenboek is, betekent ook dat het onderscheid maakt tussen verschillende betekenissen van woorden en die afzonderlijk beschrijft. De ene betekenis van een woord zal daarbij relaties hebben met andere woorden dan een andere betekenis van dat woord. Zo zal het woord 'voetbal' in de betekenis van het voorwerp waarmee gespeeld wordt, relaties kunnen hebben met onder meer de woorden 'bal' en 'ventiel':

voetbal{isEen}bal, voetbal{heeftOnderdeel}ventiel,
ventiel{isOnderdeelVan}voetbal.

Voor 'voetbal' in de betekenis van de sport kan dat bijvoorbeeld zijn 'balsport':

voetbal{isEen}balsport.

In een semantisch netwerk zullen deels andere soorten relaties gedefinieerd zijn dan in een thesaurus, die alleen de gelijkwaardigheidsrelatie (synoniem), de hiërarchische relatie (NT/BT) en associatieve relaties (RT) kent. In een semantisch netwerk beperken de relaties zich meestal tot die van taalkundige of semantische aard, zoals je die ook in een woordenboek kunt tegenkomen (Trim, 2015). Daarbij kun je onder meer denken aan:

- synoniemrelaties (fiets{isZelfdeAls}rijwiel, maar ook wat losser monowiel{isNetzoietsAls}fiets);
- en antoniemrelaties (gezond{isTegengesteldeVan}ongezond).

Het eerder genoemde {isEen} is in feite een hiërarchische relatie, zoals de NT/BT-relatie in een thesaurus. In semantische netwerken is die relatie transitief. Dat wil zeggen dat de relaties voetbal{isEen}balsport en balsport{isEen}sport automatisch impliceren dat ook voetbal{isEen}sport. Voor de deel-geheel relatie geldt dat ook. Zo zullen de relaties ventiel{isOnderdeelVan}fietsband, fietsband{isOnderdeelVan}fietswiel en fietswiel{isOnderdeelVan}fiets impliceren dat ventiel{isOnderdeelVan}fiets. Dit in tegenstelling tot de NT/BT relaties bij thesauri, die niet transitief zijn, vanwege de daar geldende regels voor modulatie/gradatie (paragraaf 3.3.3-stap 3). Daar mogen geen niveaus worden overgeslagen.

Semantische netwerken worden niet rechtstreeks gebruikt voor het ontsluiten van informatie. Ze kunnen wel een hulpmiddel zijn waarmee een computerprogramma de betekenis van woorden kan afleiden, zowel in te ontsluiten teksten, als in gestelde zoekvragen. Ze kunnen daarbij helpen mogelijke betekenissen van woorden te onderscheiden (zie het voorbeeld in paragraaf 11.3). Bij een gegeven woord kunnen ze ook



Wordnet

9.2 Ontologieën

semantisch web

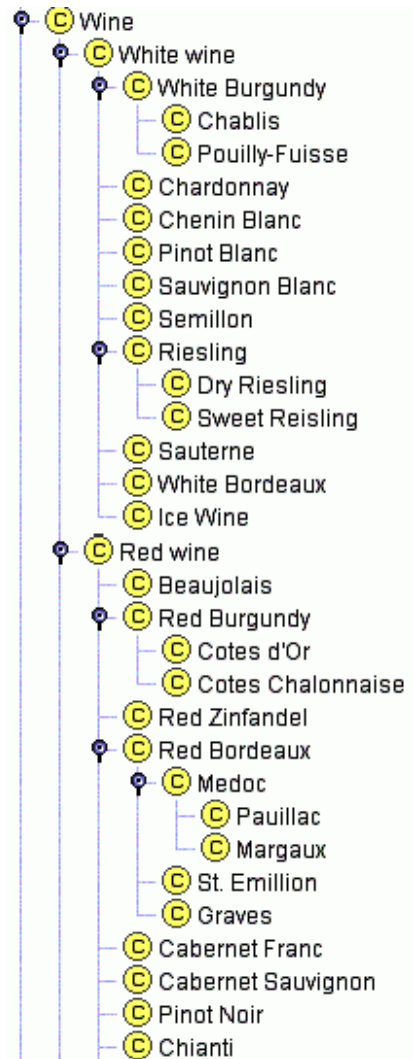
ontologie

In het kader van het semantisch web heeft het begrip ontologie wijdere bekendheid gekregen. Dit oorspronkelijk uit de filosofie afkomstige begrip was op het terrein van de kunstmatige intelligentie geïntroduceerd als aanduiding voor computerleesbare kennisrepresentaties. Dit is ook de betekenis die in het semantisch web aan het woord ontologie wordt gegeven. Een ontologie beoogt op een geformaliseerde en door computers interpreteerbare manier een gedetailleerde beschrijving te geven van (een stukje van) de werkelijkheid. Een ontologie wordt wel gedefinieerd als:

“Een strikt en uitputtend schema voor een bepaald onderwerpsdomein, meestal in een hiërarchische structuur, die alle relevante grootheden en hun relaties bevat, alsmede de regels waaraan die grootheden en relaties binnen dat domein voldoen”.

Deze omschrijving geeft al aan dat een ontologie nastreeft een veel vollediger representatie van de werkelijkheid te geven dan een thesaurus en zelfs vollediger dan een semantisch

netwerk. In een toepassing voor het semantisch web moet een computer echt betekenis van (tekst in) documenten en metadata kunnen afleiden en op basis daarvan kunnen redeneren en gevolgtrekkingen maken. Daartoe moet een ontologie ook in een formele computertaal kunnen worden beschreven. De *Web Ontology Language* (OWL) is een voorbeeld van zo'n taal (Isaac, 2006a). Door een zoekstelsel gebruik te laten maken van de rollen, eigenschappen en relaties die in een ontologie gedefinieerd zijn,



Afb. 9.2 - Voorbeeld van een hiërarchie van concepten in een wijnontologie (ontleend aan Noy, 2005).

OWL

zou dat vooral tot hogere precisie moeten leiden, al wordt ook wel van grotere vangst gesproken. Hard bewijs voor verbeterde zoekresultaten bestaat echter nog niet.

In de zojuist gegeven definitie van een ontologie was sprake van een hiërarchische structuur. Daarin lijkt een ontologie ook wel wat op een thesaurus. Zo kan in een wijnontologie bijvoorbeeld zijn vastgelegd dat Chardonnay, Riesling en Chablis soorten witte wijn zijn (zie afbeelding 9.2), en dat Rioja en Valdepeñas wijngebieden in Spanje zijn.

relaties

In de soorten relaties die tussen begrippen kunnen worden gedefinieerd, biedt een ontologie een veel grotere vrijheid dan een thesaurus. Daarin zijn maar drie soorten relaties mogelijk: gelijkwaardigheidsrelaties, hiërarchische relaties en associatieve relaties (zie hoofdstuk 4). De hiërarchische relaties NT en BT die in tegengestelde richting tussen twee termen verwijzen, mogen in een thesaurus worden gebruikt voor zowel generieke (geslacht-soort) als partitieve (deel-geheel) relaties. In een ontologie kun je eenvoudig besluiten onderscheid te maken tussen deze twee soorten hiërarchische relaties. Bijvoorbeeld:

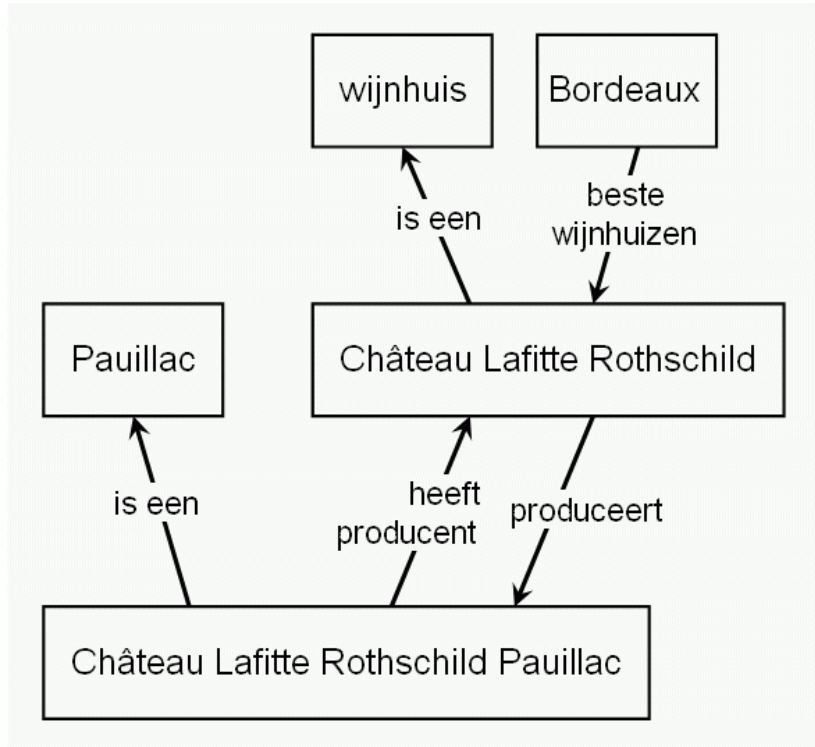
- `popconcert{isSoortVan}concert, concert{kanZijn}popconcert;`
- `arm{heeftOnderdeel}hand, hand{isOnderdeelVan}arm.`

Alle andere verwantschapsrelaties worden in een thesaurus in beide richtingen met de associatieve relatie RT aangeduid. In een ontologie is het mogelijk de aard van dergelijke associatieve relaties veel verfijnder te definiëren. Bovendien kun je tussen begrippen heel andere relaties definiëren dan ooit in een thesaurus zijn toegestaan. Ook zijn veel inhoudelijker relaties mogelijk dan in een semantisch netwerk. In al die gevallen mag je de beide richtingen van de relaties tussen twee termen ook verschillend benoemen wanneer daar behoefte aan bestaat. Bijvoorbeeld:

- `vogelgriep{isGevolgVan}H5N1_virus, H5N1_virus{isOorzaakVan}vogelgriep;`
- `Leiden{isGeboorteplaatsVan}Rembrandt, Rembrandt{isGeborenIn}Leiden;`
- `slagroom{isIngredientVoor}moorkop, moorkop{heeftIngredient}slagroom.`

Welke relaties in een bepaalde ontologie voorkomen is afhankelijk van het onderwerpsdomein. Bij het bouwen van een ontologie legt men tevoren vast welke dat zijn. Zo zal in het eerdere voorbeeld van een wijnontologie vastliggen dat specifieke wijnen een bepaalde kleur hebben, dat ze uit een bepaalde wijnstreek komen, dat ze van een bepaald soort druif en door een bepaald wijnhuis worden gemaakt, enzovoort. Ook de soorten

grootheden die binnen dit domein kunnen voorkomen, zoals 'kleur', 'druivensoort', 'wijnstreek', 'wijnsoort', 'wijnhuis' legt men daarbij vast. Afbeelding 9.3 geeft hiervan een eenvoudig voorbeeld.



Afb. 9.3 - Voorbeeld van relaties tussen verschillende soorten concepten in een wijnontologie (naar Noy, 2005).

Bij de in de eerdere definitie genoemde regels kun je denken aan de aanbeveling om bij een vismaaltijd witte wijn te schenken, of bij wildgerechten specifieke soorten Bourgogne. Op basis van dergelijke regels moet een computer ook gevolgtrekkingen kunnen maken.

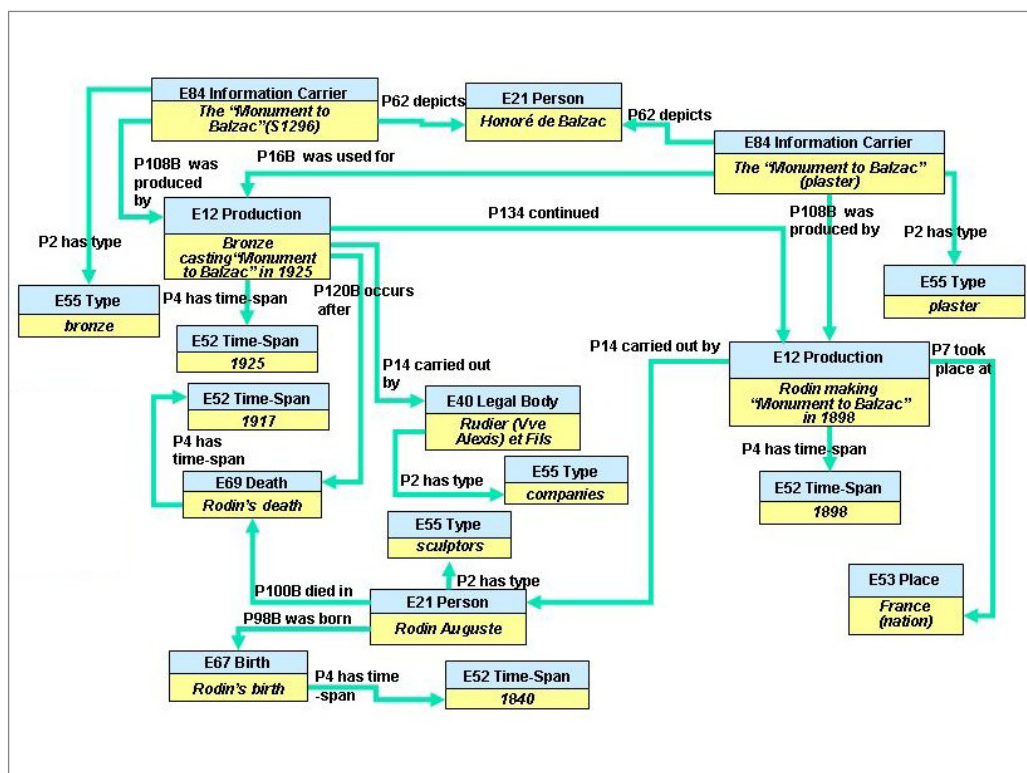
gevolgtrekkingen

upper ontologies

De ontologieën waarop de eerder gegeven definitie betrekking heeft, worden meestal opgezet voor vrij beperkte onderwerpsdomeinen. De wijnontologie die we als illustratie gebruikten, is daar een voorbeeld van. Daarin werden de specifieke eigenschappen van allerlei wijnen vastgelegd, alsmede de woorden die met wijnen te maken hebben en welke relaties die met elkaar hebben. Onder de naam ontologieën vallen echter ook wel algemener kennisrepresentaties. Daartoe behoren zogenaamde *upper ontologies*. Een voorbeeld is het Cyc-project dat onder meer tot doel heeft een ontologie op te zetten voor alledaagse kennis uit het dagelijks leven (Ontologist's, 2016). Daarin is voor een groot aantal deelgebieden domein-

inference engine

specifieke kennis vastgelegd. Bijvoorbeeld dat Nederland een land is, dat landen geopolitieke eenheden zijn en dat er ook kleinere geopolitieke eenheden bestaan, zoals provincies en gemeenten. Op basis daarvan kan de bijbehorende programmatuur (de *inference engine*) redeneren en conclusies trekken. Sommigen beschouwen ook het Wordnet semantisch netwerk als een voorbeeld van zo'n *upper ontology*. De kennis die in een semantisch netwerk als Wordnet is opgenomen, is echter erg beperkt. In paragraaf 9.1 zagen we al dat het in feite niet veel meer is dan een woordenboek waarin een beperkt aantal betekenisrelaties tussen de woorden is gelegd. Die bieden nauwelijks een basis voor redeneringen of het trekken van conclusies.



Afb. 9.4 - Voorbeeld van de gegevens over een door Auguste Rodin gemaakt beeld van Honoré de Balzac, volgens het CIDOC CRM framework. (ontleend aan http://cidoc.ics.forth.gr/crm_core/core_examples/balzac.html)

core ontologies

Een wat meer gespecialiseerde tussenvorm zijn de *core ontologies* die een specifiek, maar breed toepassingsgebied proberen te beschrijven.

CIDOC CRM

Een voorbeeld is het *Conceptual Reference Model* (CIDOC CRM), op het terrein van het cultureel erfgoed (Doerr, 2003). Dat bevat definities en een formele structuur voor het beschrijven van erfgoedconcepten, hun

relaties en hun eigenschappen. Afbeelding 9.4 geeft een voorbeeld van de toepassing van CIDOC CRM. De daarin met E-nummers aangegeven concepten zijn gestandaardiseerde soorten entiteiten. De met P-nummers aangegeven relaties zijn gestandaardiseerde soorten eigenschappen (Engels: *properties*) die de concepten verbinden. Dit soort ontologieën wordt vooral opgezet om het mogelijk te maken op heterogene wijze ontsloten materiaal uit verschillende systemen gezamenlijk doorzoekbaar te maken. In principe kunnen computers daarmee namelijk afleiden welke betekenis de in die systemen gebruikte velden en termen hebben. Zo kunnen ze een rol spelen om interoperabiliteit van verschillend opgezette informatiesystemen te bewerkstelligen.

In principe is het mogelijk een bestaande thesaurus om te werken tot een ontologie. Dat is bijvoorbeeld gedaan met de Agrovoc thesaurus van de Food & Agriculture Organization van de Verenigde Naties, de FAO (Soergel, 2004). Dit werd vooral nodig geacht omdat de “paar rudimentaire categorieën termen en relaties van een traditionele thesaurus” onvoldoende mogelijkheden boden voor realisering van het zeer intelligente systeem dat men voor ogen had. Een ontologie, door Soergel een “*semantically rich knowledge organization system*” genoemd, zou wel moeten leiden tot de gewenste:

- betere gebruikersinteractie die de gebruiker ondersteunt om tot een betere zoekvraag te komen;
- automatische intelligente zoekvraag-expansie;
- intelligente ondersteuning van menselijke indexeerders en verbetering van automatische indexering;
- ondersteuning van semantisch web toepassingen. Dit punt heeft ook weer betrekking op de rol van ontologieën bij de interoperabiliteit van informatiesystemen.

In de praktijk worden ontologieën meestal op andere wijze ingezet voor het toegankelijk maken van informatie dan meer klassieke technieken. Het zijn inderdaad *knowledge organisation systems* die kennis beschrijven. Die kennis kan vervolgens worden ingezet bij de interpretatie van de betekenis van documentteksten, inclusief daar eventueel al aan toegekende thesaurusterminen, en van zoekvragen. Ze zijn veel minder bedoeld om documenten zelf te ontsluiten door daar rechtstreeks termen uit een ontologie aan toe te kennen, al is dat ook niet verboden of per definitie onmogelijk.

Voor experimenten in het kader van het semantisch web zijn nog niet altijd “echte” ontologieën beschikbaar. Daarom wordt vaak nog gebruik gemaakt van al beschikbare thesauri, taxonomieën, autorisatielijsten en dergelijke. In het semantisch web jargon worden die met enige bravoure dan toch ook al ontologieën genoemd, hoewel de kennis daarin op lang niet zo gedetailleerde wijze is vastgelegd als we hierboven beschreven.

Zo worden zelfs metadataschema's als Dublin Core of Learning Object Metadata (LOM) soms al ontologieën genoemd. Daarmee is het woord ontologie een soort containerbegrip geworden voor allerlei soorten kennisorganisatiesystemen.

9.3 Vergelijking van kennisrepresentaties

In de beschrijving van de hier behandelde soorten kennisrepresentaties werden hier en daar al onderlinge verschillen aangegeven. Hier maken we op wat formelere wijze nog eens een vergelijking. Het overzicht in tabel 9-1 is geïnspireerd op de vergelijking die Garshol (2004) maakt tussen *topic maps*, thesauri, taxonomieën en ontologieën. *Topic maps* komen in dit boek verder niet aan de orde, omdat ze in Nederland nauwelijks worden toegepast. Wie daar meer over wil weten, verwijzen we naar het boek "Organiseer je informatie" (Becker, 2010).

	Aantal mogelijk te beschrijven soorten dingen	Aantal mogelijke eigenschappen daarvan	Aantal mogelijke soorten relaties daartussen	Primair doel is informatie toegankelijk te maken
Taxonomie	1	-	1	ja
Thesaurus	1	1	3	ja
Semantisch netwerk	1	1	∞	nee
Ontologie	∞	∞	∞	nee

Tabel 9.1 - Vergelijking van een aantal eigenschappen van taxonomieën, thesauri, semantisch netwerken en ontologieën.

In een taxonomie is maar sprake van één soort "dingen": de categorieën of rubrieken die gedefinieerd worden. Daarvan worden verder geen eigenschappen vastgelegd. De enige soort relatie die in een taxonomie kan bestaan is de hiërarchische relatie tussen algemenere en meer specifieke categorieën.

In een thesaurus worden uitsluitend termen gedefinieerd, dus ook maar één soort dingen. De enige eigenschap die je daarvan kunt vastleggen, is een beschrijving van hun betekenis door middel van een *scope note*. Wel kent de thesaurus drie standaard vastgelegde soorten relaties tussen de termen, namelijk de hiërarchische (NT/BT), de associatieve (RT) en de verwijzende (Use/Used for) relaties.

Ook een semantisch netwerk bevat beschrijvingen van maar één soort dingen, namelijk woorden. Als enige eigenschap van elk woord, bevat het de betekenis ervan, zoals in een verklarend woordenboek, vooral om onderscheid tussen verschillende betekenissen duidelijk te maken. In principe is er geen beperking in aard en aantal van de relaties die tussen de woorden gedefinieerd kunnen worden, al blijft dat aantal in de praktijk meestal tamelijk beperkt.

Bij ontologieën zijn er in het geheel geen beperkingen. Bij het bouwen daarvan kun je willekeurig veel afzonderlijke soorten dingen benoemen die je wilt kunnen beschrijven: personen, landen, steden, wijnen, componisten, plantensoorten, enzovoort. Voor elk daarvan afzonderlijk kan de bouwer een willekeurig aantal specifieke eigenschappen definiëren, waarmee hij elk van die soorten dingen wil kunnen beschrijven. En tenslotte is er ook volstrekte vrijheid in het zelf definiëren van de soorten relaties, die tussen de te beschrijven soorten dingen mogen bestaan. Men zegt daarom dat ontologieën (en dat geldt ook voor *topic maps*) open systemen zijn. Door niettemin bepaalde formele beperkingen aan het systeem op te leggen, waartoe je als systeembouwer ook de vrijheid hebt, kun je in een ontologie toch ook een taxonomie of een thesaurus representeren.

9.4 RDF en tripels

Voor de interoperabiliteit van kennis- en ontsluitingssystemen in een netwerkgeving als internet, is het van belang dat kennisrepresentaties als semantische netwerken, ontologieën en thesauri, computerinterpreteerbaar zijn. Dat wil zeggen dat een computer definities van begrippen en van relaties daartussen moet kunnen vinden en begrijpen. Daarbij kun je onder meer denken aan het automatisch leggen van relaties tussen zoekwoorden en termen uit een bestaande thesaurus, of aan het bepalen van equivalentie tussen termen uit verschillende systemen, zodat bijvoorbeeld bekend is dat met de term 'diabetes mellitus' in het ene systeem hetzelfde wordt bedoeld als met 'suikerziekte' in het andere systeem. In het bijzonder in het kader van het semantisch web probeert men interoperabiliteit van informatiesystemen te realiseren. Dat vereist interoperabiliteit van zowel gebruikte metadata (de daarin gekarakteriseerde eigenschappen) als van de inhoud; de termen van toegepaste ontsluitingssystemen, zoals classificaties en thesauri. Met dat doel is vanuit het World Wide Web Consortium (W3C) een aantal (technische) standaarden ontwikkeld, die dit gezamenlijk mogelijk moeten maken. Tot die standaarden behoren:

- RDF (*Resource Description Framework*), een algemene methode om de relatie tussen te karakteriseren objecten, alsmede hun eigenschappen en de daarvoor te gebruiken metadatastandaarden, in XML vast te leggen;

- RDF-S (RDF Schema), een methode om ontologieën (en meer algemeen kennisorganisatiesystemen) gestructureerd en computerinterpreteerbaar in RDF vast te leggen;
- OWL (*Web Ontology Language*), een formele taal voor het computerinterpreteerbaar beschrijven van ontologieën die veel meer mogelijkheden biedt dan RDF-S;
- SKOS (*Simple Knowledge Organisation System*), een conceptueel model waarin, met RDF als beschrijvingstaal, wordt aangegeven uit welk kennisorganisatiesysteem termen afkomstig zijn die als ontsluiting zijn gebruikt, en hoe die zich tot andere verhouden.

Op SKOS als praktische toepassing om (vooral) thesauri te beschrijven gaan we in paragraaf 9.6.1 nader in. RDF komt hier al gedetailleerder aan de orde, omdat het steeds vaker de basis vormt voor de wijze waarop kennisrepresentaties gestructureerd worden. Ook wie geen kennistechnoloog is of zich anderszins heeft gespecialiseerd op het terrein van formele kennisrepresentaties, kan steeds vaker met RDF te maken krijgen. Op RDF-S en OWL wordt hier niet verder ingegaan, omdat die standaarden wat al te ver buiten het kader van dit boek vallen.

RDF tripels

subject

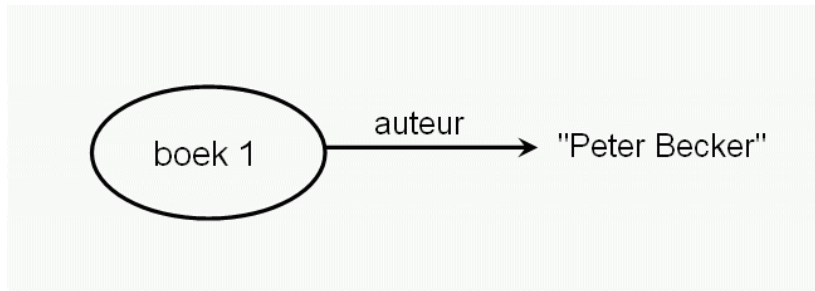
predicaat

object

RDF is een standaard beschrijvingstaal om in XML metadata aan objecten te koppelen (Schie, 2006). RDF werkt met zogenaamde tripels, drietallen: subject - predicaat - object. Deze benamingen zijn een nogal letterlijke vertaling uit het Engels van de grammaticale begrippen: onderwerp - gezegde - lijdend voorwerp. In het Nederlands kan dat wellicht wat verwarrend overkomen: het subject is in dit jargon het te beschrijven ding, in dit verband ook wel de *resource* genoemd, al kan het ook om een fysiek object gaan. Het predicaat is een eigenschap die dat subject heeft; enigszins vergelijkbaar met wat we in databasetermen een attribuut of een veld zouden noemen. Voor de inhoud of waarde van die eigenschap is de officiële benaming object. Grafisch kun je zo'n triplet weergeven als in afbeelding 9.5.

literal

Hier is het subject (de *resource*) een document (bijvoorbeeld dit boek) met als eigenschap (predicaat) dat het een auteur heeft. Het object is hier wie die auteur is, in dit geval met de waarde 'Peter Becker'. Als je dit triplet schrijft als "dit boek is geschreven door Peter", is de grammaticale achtergrond van de gebruikte Engelse benamingen wat beter herkenbaar. Bij zo'n triplet moet het subject - hier het document - een digitaal adres hebben, een URI (*uniform resource identifier*). De op internet gebruikte URL's (*uniform resource locators*) zijn de meest bekende vorm van URI's. In deze tripels kan het object direct een echte waarde zijn, zoals hier de auteursnaam. Dat wordt een *literal* genoemd. Een auteur kan echter meer eigenschappen hebben dan alleen maar die naam. Daarom kan een object op zijn beurt ook weer het subject zijn van volgende tripels waarbij het dus ook weer met een digitaal



Afb. 9.5 - Grafische weergave van een RDF triple.

adres (URI) geïdentificeerd moet kunnen worden. In die tripels heeft het zelf weer eigenschappen (predicaten), zodat een heel netwerk kan ontstaan. In het geval van de auteur kunnen die eigenschappen bijvoorbeeld de naam, het e-mailadres en de werkgever van de auteur zijn, zoals geïllustreerd in afbeelding 9.6.

Een beschrijving in RDF luidt dan:

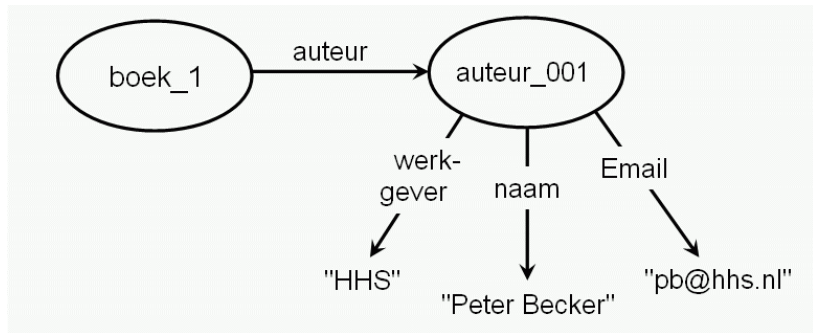
```

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  <rdf:Description rdf:about="http://www.xxx.com/yyy/boek1.xml">
    <auteur>
      <rdf:Description rdf:about="http://www.xxx.com/autId/7801">
        <naam>Peter Becker</naam>
        <werkgever>HHS</werkgever>
        <email>pb@hhs.nl</email>
      </rdf:Description>
    </auteur>
  </rdf:Description>
</rdf:RDF>
  
```

namespaces

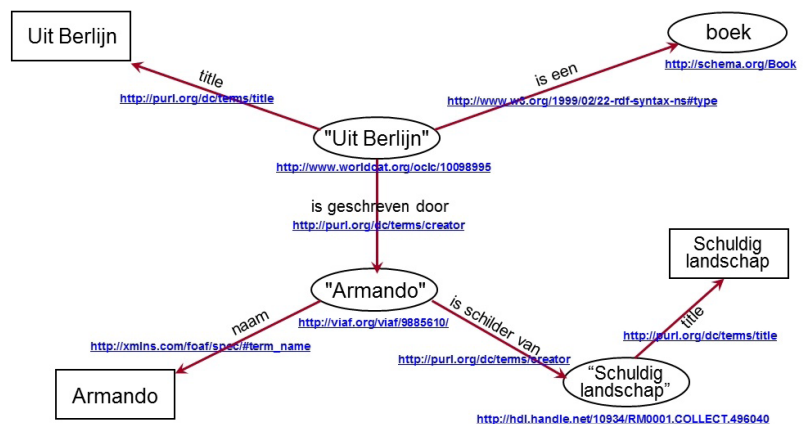
In dit voorbeeld zien we ook het mechanisme van *namespaces*, in de eerste regel van het RDF-voorbeeld aangeduid met "xmlns:" (*XML-NameSpace*). Daarmee verwijst RDF naar de computerleesbare beschrijvingen van de gebruikte standaarden. Dit maakt het ook mogelijk om aan hetzelfde object metadata te koppelen die uit verschillende standaarden afkomstig zijn en om naar al die standaarden te verwijzen.

Voor de predicaten, de eigenschappen van een subject, streeft men er naar die ook zo veel mogelijk een URI te laten hebben. Daarmee wordt dus in feite verwezen naar een computer leesbare definitie van dat predicaat.



Afb. 9.6 - Grafische weergave van een netwerkje van vier RDF-tripels.

Een visueel voorbeeld hiervan vind je in afbeelding 9.7. Daarin zijn rechthoekjes gebruikt om *literals* aan te geven; de ovalen hebben een eigen URI en kunnen dus zelf weer subject in andere tripels zijn. Zo staat het ovaal 'Armando' voor de persoon als concept, waarvan ergens op een computer allerlei eigenschappen bekend kunnen zijn. De rechthoek 'Armando' is domweg de reeks letters die zijn naam weergeeft.



Afb. 9.7 - Voorbeeld van netwerkje van tripels met URL's voor objecten, subjecten en predicaten, naar een idee van Lucas Koster.

In RDF kan iets dergelijks eruit komen te zien als het wat vereenvoudigde voorbeeld hieronder, waarbij de URI's voor de predicaten ook als *namespaces* worden meegegeven. 'DC' staat daarbij voor de Dublin Core metadata-standaard, 'vCard' is een veelgebruikte standaard om persoonsgegevens te beschrijven en 'MESH' zijn de *Medical Subject Headings* (zie hoofdstuk 13) waarmee het document

inhoudelijk wordt beschreven met zowel thesaurustermen als daarmee corresponderende classificatiecodes:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc=http://purl.org/dc/elements/1.1/
  xmlns:vcard="http://imc.org/vCard/3.0#"
  xmlns:mesh="http://nlm.nih.gov/mesh/MBrowser.htm">
  <rdf:Description rdf:about="http://www.xxx.nl/yyy/doc1.xml">
    <dc:title>Indexing of medical articles</dc:title>
    <dc:creator>
      <rdf:Description about="http://www.yyy.nl/authors/smithj">
        <vcard:FN>John Smith</vcard:FN>
        <vcard:EMAIL>smithj@nih.gov</vcard:EMAIL>
      </rdf:Description>
    </dc:creator>
    <dc:subject>indexing</dc:subject>
    <mesh:subject>Abstracting and Indexing as Topic</mesh:subject>
    <mesh:subject>International Classification of Diseases
    </mesh:subject>
    <mesh:classification>I1.453.245.100.</mesh:classification>
    <mesh:classification>I1.453.245.945.400.</mesh:classification>
  </rdf:Description>
</rdf:RDF>
```

triple stores

Intussen zijn al veel in RDF gecodeerde dataverzamelingen op internet beschikbaar gekomen. Omdat computers deze gegevens op standaardwijze kunnen lezen en dus kunnen gebruiken om allerlei systemen aan elkaar te koppelen, spreekt men sinds enige tijd over *linked data* om deze dataverzamelingen te omschrijven. Beschikbare bronnen voor *linked data* worden wel *RDF triple stores* genoemd. De wolk van dergelijke datacollecties, zoals afgebeeld in afbeelding 9.12, is in tamelijk korte tijd sterk in omvang gegroeid. Vooral Tim Berners Lee, de bedenker van het *world wide web* en een belangrijk pleitbezorger voor het semantisch web, heeft hierbij een stimulerende rol gespeeld. Deze computerleesbare datacollecties ziet hij namelijk als de ruggengraat voor dat semantisch web.

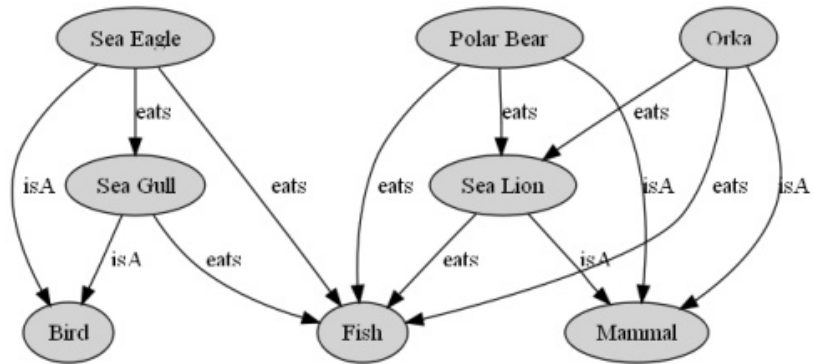
Tripels en RDF vormen ook de basis waarop de eerder in dit hoofdstuk besproken relaties in semantische netwerken en ontologieën kunnen worden beschreven. Die tripels rijgen zich in een grafische weergave aaneen tot enorme netwerken.

9.5 Graphs

grafen

De netwerken die uit de koppeling van tripels ontstaan, worden wiskundig ook wel grafen genoemd (Engels: *graphs*). Een klassieke tak van de wiskunde, de grafentheorie houdt zich met de formele kanten van

dergelijke netwerken bezig. Maar ook zonder kennis van die wiskundige achtergrond, kunnen we er wel een idee van krijgen. In afbeelding 9.8 zien we de *graph*-representatie van een klein stukje van een fictieve ontologie. De bolletjes daarin heten knooppunten (Engels: *nodes*), de verbindingen daartussen takken of pijlen (Engels: *edges*, *arcs* of *lines*).

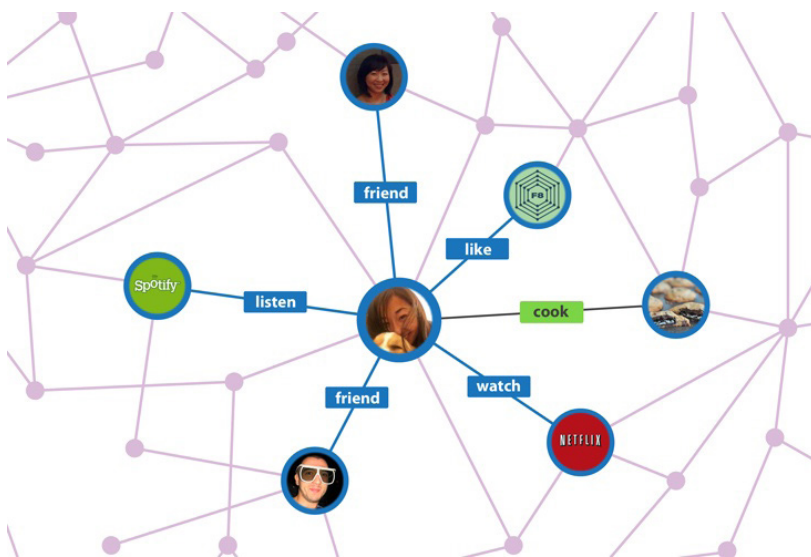


Afb. 9.8 - Graph-representatie van een klein stukje van een ontologie.

In een semantisch netwerk zijn het de woorden uit het woordenboek die een *graph* vormen. In een ontologie is het de kennis van een stukje van de wereld dat deels, waar het de relaties tussen de daarin gedefinieerde begrippen betreft, in een *graph* kan worden weergegeven.

Door grote zoekmachines als Google en Bing worden *graph*-technieken toegepast om de kennis vast te leggen die het mogelijk moet maken om gebruikers concrete antwoorden op zoekvragen te geven, in plaats van alleen een lijstje van tien blauwe links. Google tracht daartoe de kennis van de hele wereld in zijn *knowledge graph* weer te geven (zie paragraaf 9.6.3).

Ook Facebook maakt gebruik van *graphs*. Daar worden ze toegepast voor het vastleggen en beheren van de meer dan een miljard personen en de vele honderden miljarden eigenschappen en objecten die gebruikers hebben ingevoerd of geüpload, alsmede de relaties daartussen. De mogelijkheid die Facebook enige tijd geboden heeft om gestructureerd in deze enorme hoeveelheid gegevens te zoeken, heette ook "*Graph Search*". Die relaties tussen gebruikers onderling, tussen gebruikers en foto's, boeken, films, woonplaatsen, werkgevers enzovoort worden heel gestructureerd getypeerd. Daartoe definieert Facebook de tripels iets beperkter als: *user – action – object*. Bij die *actions* (overeenkomend met de predicaten in RDF) kun je denken aan: 'like', 'listen', 'watch', 'play', 'cook', 'friend'. Bij de objecten aan: 'photo', 'song', 'movie', 'game', 'recipe', 'car', 'person'. In afbeelding 9.9 is hier een schematisch voorbeeld van gegeven.



Afb. 9.9 - Schematisch voorbeeld van een stukje van de Facebook Graph.

Triple stores en *graphs* worden tegenwoordig vaak als alternatief gezien voor klassieke databases om het soort gegevens waar we het hier over hebben op een efficiënte manier te kunnen opslaan en opvragen.

9.6 Toepassingen

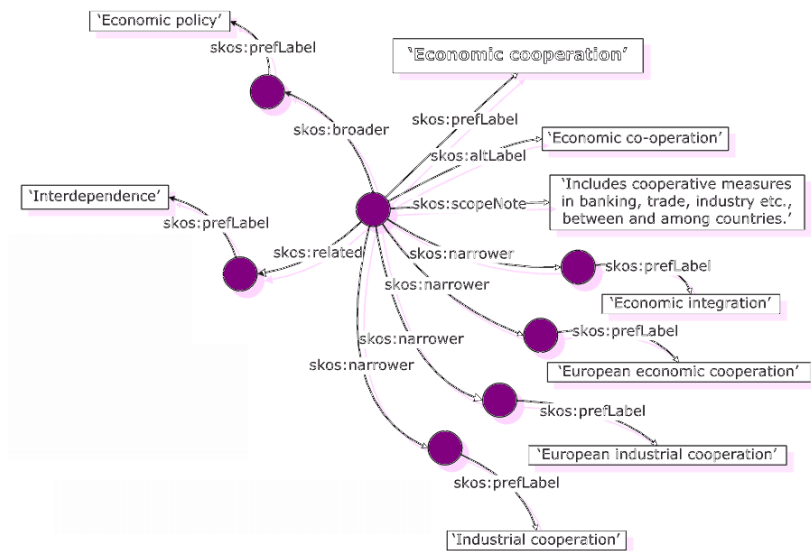
9.6.1 SKOS

SKOS

SKOS is een standaard die bedoeld is om kennisrepresentaties met behulp van RDF te beschrijven, of anders gezegd waarmee je schema's voor kennisorganisatiesystemen kunt opzetten (Isaac, 2006b, 2009a). Bij de in de volgende paragraaf besproken *linked data* speelt SKOS ook vaak een rol. Als voorbeeld bekijken we het gebruik van SKOS voor (een stukje van) een klassieke thesaurus (Miles, 2005). In de oorspronkelijke thesaurus is vastgelegd:

Term:	Economic cooperation
Used For:	Economic co-operation
Broader terms:	Economic policy
Narrower terms:	Economic integration, European economic cooperation, European industrial cooperation, Industrial cooperation
Related terms:	Interdependence
Scope Note:	Includes cooperative measures in banking, trade, industry etc., between and among countries.

SKOS gaat uit van concepten en niet van termen. In de grafische representatie in afbeelding 9.10 zijn die concepten door bolletjes gerepresenteerd. Elk concept kan door één of meer termen gekarakteriseerd worden. Anders zouden we ook niet kunnen weten om wat voor concept het gaat. In het daar gegeven voorbeeld van het beschrijven van gewone thesaurustermen en hun relaties, koppelt de eigenschap `<skos:prefLabel>` de voorkeursterm op RDF-wijze aan het concept. Ook de reguliere thesaurusrelaties zie je daar als eigenschappen (`<skos:narrower>`, `<skos:broader>` etc.) die naar andere concepten verwijzen. Daaraan zullen op hun beurt ook weer termen hangen en ook kunnen die weer eigenschappen hebben in de vorm van thesaurusrelaties.



Afb. 9.10 - Grafische weergave van thesaurusrelaties in een SKOS-representatie; ontleend aan <http://www.w3.org/2004/02/skos/core/guide/2005-10-06/>.

De computerleesbare beschrijving van deze SKOS-representatie in RDF luidt:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<skos:Concept>
  <skos:prefLabel>Economic cooperation</skos:prefLabel>
  <skos:altLabel>Economic co-operation</skos:altLabel>
  <skos:scopeNote>Includes cooperative measures in banking, trade, industry
etc., between and among countries. </skos:scopeNote>
  <skos:broader>
    <skos:Concept>
      <skos:prefLabel>Economic policy</skos:prefLabel>
    </skos:Concept>
  </skos:broader>
  <skos:related>
    <skos:Concept>
      <skos:prefLabel>Interdependence</skos:prefLabel>
    </skos:Concept>
  </skos:related>
  <skos:narrower>
    <skos:Concept>
      <skos:prefLabel>Economic integration</skos:prefLabel>
    </skos:Concept>
  </skos:narrower>
  <!-- ...more narrower terms omitted ... -->
</skos:Concept>
</rdf:RDF>
```

In eerste instantie beperkt SKOS zich nog tot de grootste gemene deler van de in diverse soorten kennisorganisatiesystemen gebruikte eigenschappen en relaties, de zogenaamde SKOS-core (Miles, 2005). Dat de eerste S van SKOS de S van *simple* is, geeft evenwel aan dat het ook expliciet de bedoeling is dat het een eenvoudige subset blijft van alle meer complexe zaken die met RDF-S en OWL mogelijk zijn. Daarbij richt het zich voornamelijk ook vooral op “eenvoudige” ontologieën, dat wil zeggen op thesauri en classificaties. Behalve relaties binnen een bepaalde ontologie, kan SKOS ook relaties leggen met andere ontsluitingsystemen of met andere soorten metadata. Zo kent het onder meer de eigenschappen <skos:exactMatch>, <skos:closeMatch> en <skos:broadMatch> om aan te geven hoe precies de overeenkomst in betekenis is van twee concepten uit verschillende vocabulaires. Dat is al een aanzet om via SKOS werkelijk tot interoperabiliteit van systemen te komen (Sanchez-Alonso, 2006). Verder kan SKOS ook worden gebruikt om documenten direct te koppelen aan concepten uit verschillende kennisorganisatiesystemen. Vooral in de wereld van het cultureel erfgoed wordt deze aanzet tot verbetering van de interoperabiliteit

als een belangrijke ontwikkeling gezien, omdat juist daar van origine weinig onderlinge standaardisatie bestond. Een voorbeeld van een toepassing op dit gebied is het Nederlandse CultuurLINK project (<http://cultuurlink.beeldengeluid.nl/>), waarin Beeld & Geluid een belangrijke rol speelt (Streefkerk 2015, Hildebrand 2015, Priem 2016).

Intussen is al een aantal systemen uit de klassieke ontsluitingswereld in SKOS gecodeerd, zoals de Dewey Decimale Classificatie (DDC) en de Library of Congress Subject Headings (LCSH) (Harper, 2006; Miles, 2009). Voor een Nederlands proefproject zijn ook een bij de Koninklijke Bibliotheek gebruikt ontsluitingssysteem (de termen uit de “Brinkman”) en de Nederlandse Basisclassificatie (NBC) in SKOS omgezet (Mols, 2008). In het Europeana-project, dat een grote diversiteit aan objecten uit velerlei Europese erfgoedcollecties via internet toegankelijk maakt, speelt SKOS een belangrijke rol (Isaac, 2009b). Daardoor wordt het mogelijk om optimaal gebruik te maken van de kennis die aanwezig is in al die verschillende ontsluitingssystemen die in individuele collecties zijn toegepast.

De W3C-organisatie gaf in december 2015 informatie over 39 thesauri en woordenlijsten die als SKOS datasets beschikbaar waren (SKOS 2015). Een gids over het gebruik van SKOS, gericht op informatieprofessionals, is onlangs gepubliceerd door een onderafdeling van de American Library Association (Frazier 2015).

9.6.2 Linked data

Bij de introductie van RDF en tripels in paragraaf 9.4 gaven we al aan dat intussen veel in RDF gecodeerde dataverzamelingen op internet beschikbaar zijn en dat van *linked data* wordt gesproken om deze dataverzamelingen te omschrijven. Om gegevens uit *linked data* collecties ook werkelijk te kunnen (her)gebruiken, is het wel essentieel dat ze vrij (open) toegankelijk zijn. In dat verband wordt dan ook van *linked open data* gesproken. Als onderdeel van zijn oproep om meer data vrij op internet beschikbaar te stellen, heeft Tim Berners Lee een vijf-sterren model voor open data geïntroduceerd (afbeelding 9.11). Hoe meer sterren een collectie krijgt, hoe waardevoller hij is. Pas bij 5 sterren is echt sprake van *linked open data*:

- * zet je data op het web onder een open licentie
- ** zorg dat het gestructureerde data zijn; dus geen platte tekst of PDF
- *** zorg dat het in een open format is, dus bijvoorbeeld CSV i.p.v. Excel
- **** gebruik hiervoor W3C standaarden, zoals Sparql en RDF met URI's
- ***** link je data met die van anderen

linked data

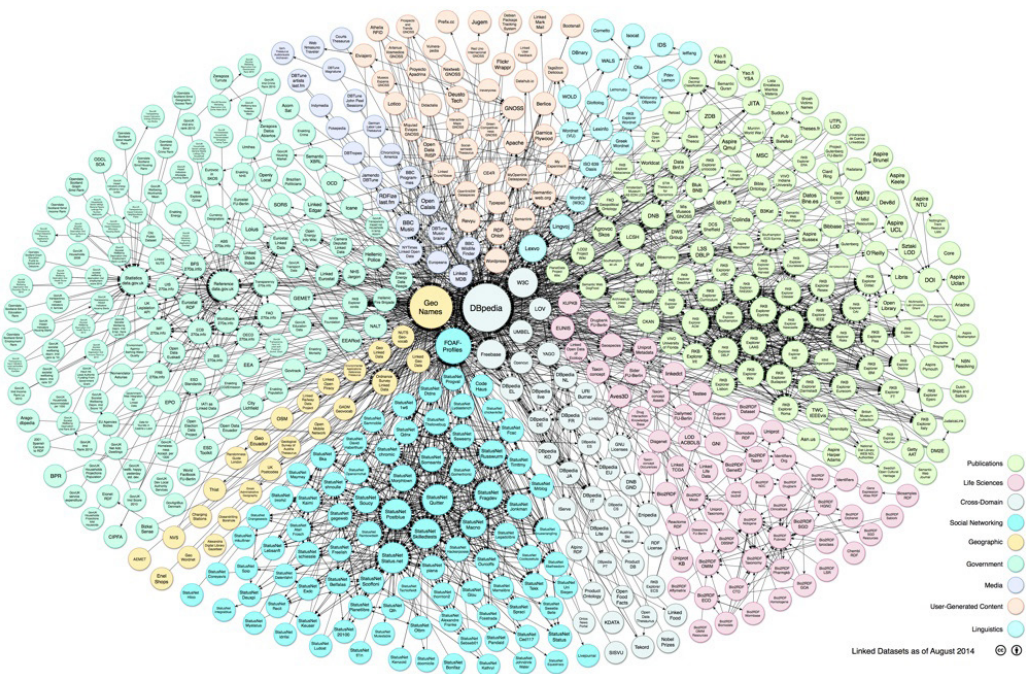
linked open data



Afb. 9.11 - Het 5-sterrenmodel voor linked open data.

DBpedia**GeoNames**

Afbeelding 9.12 toont de *Linked data cloud*, de wolk van onderling gekoppelde, open toegankelijke datacollecties. Tamelijk centraal in die wolk staat de DBpedia, een grote verzameling van gestructureerde RDF-gecodeerde gegevens die uit de Wikipedia zijn afgeleid. Een groot aantal andere systemen linkt daarheen om er geautomatiseerd extra informatie over allerlei onderwerpen aan te kunnen ontleenen. De letters DB in DBpedia zijn een aanwijzing dat het hier eigenlijk om een database-achtige extensie van het web gaat. Een andere centrale positie wordt ingenomen door *GeoNames* een wereldomvattend systeem voor geografische benamingen. Door de beschikbaarheid daarvan wordt het zelf samenstellen van geografische autorisatielijsten (zie paragraaf 9.5) veelal overbodig.



Afb. 9.12 - Databronnen in de *Linked data cloud* (situatie april 2014) zoals getoond op de site <http://lod-cloud.net/>.

Wanneer je de wolk in afbeelding 9.12 in hogere resolutie bekijkt (<http://lod-cloud.net/versions/2014-08-30/lod-cloud.svg>), is verder onder meer te zien:

- dat de BBC heel actief is op dit terrein. Ze stellen onder andere allerlei muziekgegevens beschikbaar. In hun *linked data store* zitten beschrijvingen van allerlei soorten objecten (*things*) (<http://www.bbc.co.uk/things/>);
- dat ook de *Cyc upper ontology* (zie paragraaf 9.2) zo beschikbaar is: OpenCyc;

- dat ook overheden van veel landen er belang aan hechten hun gegevens beschikbaar te stellen. Daarbij hopen ze vaak dat anderen innovatieve applicaties op basis van hun data ontwikkelen. Zie bijvoorbeeld: <https://data.overheid.nl/>, <http://www.data.gov/>, <http://opengovernmentdata.org/>;
- dat de *virtual international authority file* (VIAF), waarin een groot aantal bibliotheken samenwerkt, als *linked data* beschikbaar is (zie paragraaf 5.7).

Een andere op basis van tripels opgebouwde kennisbank is Wikidata (<https://www.wikidata.org/>). Hoewel (nog) niet getoond in de LOD-cloud, is hij een steeds belangrijker bron voor het gebruik van *linked data*. Op dit moment (augustus 2016) zijn daarin bijna 25 miljoen items via een veelvoud aan tripels beschreven.

SPARQL

Voor het opvragen van gegevens uit dergelijke dataverzamelingen is een speciale gestandaardiseerde query-taal ontwikkeld, SPARQL. Die naam is een recursief acroniem waarvan de S ook weer voor Sparql staat: *Sparql Protocol and RDF Query Language*. Voor linked data speelt Sparql de rol die SQL (Standard Query Language) voor relationele databases speelt. Voorbeelden hoe SPARQL-queries eruit zien en worden uitgevoerd zijn bijvoorbeeld te zien op <http://client.linkeddatafragments.org/>.

Uiteindelijk is het natuurlijk niet de bedoeling dat het blijft bij het beschikbaar stellen van *linked data* verzamelingen. Er moet uiteindelijk iets mee gedaan worden. Dat kan bijvoorbeeld doordat applicaties gebouwd worden die van die data gebruik maken. Ter illustratie hier een eenvoudig voorbeeld hoe *linked data* bij zoekacties toegepast kunnen worden. Stel dat ergens in een *linked dataverzameling* RDF tripels aanwezig zijn van de vorm

- Augustus {isEen} Romeinse keizer,
- Trajanus {isEen} Romeinse keizer,
- Hadrianus {isEen} Romeinse keizer, enz.

In een daartoe ingericht zoekstelsel kan voor een zoekactie naar 'Romeinse keizers' dan automatisch een SPARQL-query naar die externe dataverzameling gestuurd worden, waarmee gezocht wordt naar tripels met predicaat {isEen} en object "Romeinse keizer". De zo gevonden subjecten (Augustus, Trajanus, Hadrianus, enz.) kunnen in het oorspronkelijke zoekstelsel dan automatisch als extra zoektermen aan de zoekvraag worden toegevoegd. Een op die manier geëxpandeerde zoekactie is in feite een variant op het generiek zoeken dat bij gebruik van een thesaurus gebeurt.

Een overzicht van in Nederland ontwikkelde applicaties is te vinden in het boekje "Nederlandse Parels van Linked Data Toepassingen" (2016). Een paar andere toepassingen zijn bijvoorbeeld (De Boer 2012, 2013, 2016, Isaac 2013).

9.6.3 Knowledge Cards, Knowledge Graph en Knowledge Vault bij Google

knowledge cards

Zoekacties in Google leveren steeds vaker concrete antwoorden op voor wat vermoedelijk met de zoekvraag werd bedoeld. Dat gebeurt veelal in zogenaamde *knowledge cards* die rechts naast het gewone zoekresultaat verschijnen. Daarin wordt een aantal belangrijke kenmerken van een in de zoekvraag voorkomende entiteit opgesomd. Afbeelding 9.13 toont een voorbeeld van zo'n *knowledge card* die een gebruiker als antwoord krijgt op de simpele zoekvraag 'Rembrandt'.

Google knowledge graph

De gegevens in deze *knowledge card* zijn allemaal afkomstig uit tripels met eigenschappen van het subject Rembrandt. Al die tripels zijn opgeslagen in de *Google knowledge graph*, een voorbeeld van een *graph* zoals besproken in paragraaf 9.5. Deze *knowledge graph* bevatte in 2014 al gegevens over meer dan 700 miljoen subjects en meer dan 20 miljard kenmerken die via predicaten daaraan gekoppeld zijn. De waarden van die kenmerken, de objects, kunnen zogenaamde *literals* zijn, zoals Rembrandt's geboorte- en sterfdatum, maar heel vaak zijn dat zelf ook weer subjects waarover meer gegevens bekend zijn. In afbeelding 9.13 is dat te zien aan de gegevens die als hyperlinks zijn weergegeven.

Behalve de tamelijk generieke encyclopedische antwoorden in de *knowledge cards*, geeft Google ook vaak concrete antwoorden op heel gespecificeerde vragen. Zo levert bijvoorbeeld de zoekvraag 'melting point iron' de betreffende temperatuur in grote letters boven de gewone zoekresultaten (afbeelding 9.14). In dergelijke voorbeelden wordt dus gebruik gemaakt van slechts een enkel tripel uit de *knowledge graph*.

knowledge vault

De aanvankelijke vulling van deze gigantische *knowledge graph*, was voor een belangrijk deel afkomstig van gegevens uit de DBpedia. Daarnaast zijn echter ook diverse andere bronnen gebruikt. Meer recent worden deze gegevens ook aangevuld uit Google's zogenaamde *knowledge vault*, een kelder waarin Google gegevens verzamelt die het indirect afleidt uit de inhoud van webpagina's, ook als daar geen formele structuur in is vastgelegd. Zo kan uit een tabel in een PDF-document bijvoorbeeld worden afgeleid dat tussen gegevens uit twee kolommen een relatie bestaat, die wordt gekenmerkt door de tekst die als kop boven een kolom staat. Die wordt dus als predicaat geïnterpreteerd. Omdat deze methode natuurlijk erg gevoelig is voor misinterpretaties, moet voor elk zo afgeleid tripel wel bevestiging worden verkregen uit een voldoende aantal andere webpagina's of bronnen, voordat die gegevens van de *knowledge vault* naar de *knowledge graph* gepromoveerd kunnen worden (Dong, 2014). Hierbij kan het zeker helpen als in webpagina's wel gestructureerd enige metadata zijn opgenomen, zoals via het Schema.org metadataschema, dat in de volgende paragraaf behandeld wordt. Niettemin worden in blogs en op Twitter nog met enige regelmaat foute gegevens gerapporteerd die in Google's antwoorden voorkomen. Een overzicht van methoden om automatisch gegenereerde *knowledge graphs* te verfijnen, te corrigeren en aan te vullen is te vinden bij Paulheim (2016).

Google is overigens niet de enige zoekmachine die voor het concreet beantwoorden van gebruikersvragen, gebruik maakt van gegevens uit een *knowledge graph*. Dat is onder meer ook het geval bij de zoekmachine Bing van Microsoft en bij de Chinese zoekmachine Baidu. De door Bing gebruikte *knowledge base* heet Satori; zelf noemen ze het ook wel een “*knowledge and action graph*”. Hiervan werd in 2015 gerapporteerd dat hij van 2 miljard objecten ongeveer 25 miljard eigenschappen bevatte (He, 2015).



Rembrandt

Painter

Rembrandt Harmenszoon van Rijn was a Dutch painter and etcher. He is generally considered one of the greatest painters and printmakers in European art and the most important in Dutch history. [Wikipedia](#)

Born: July 15, 1606, Leiden, [Netherlands](#)

Died: October 4, 1669, Amsterdam, [Netherlands](#)

Influenced: [Vincent van Gogh](#), [Gerrit Dou](#), [Jan Lievens](#), [Carel Fabritius](#), [Govert Flinck](#)

Influenced by: [Caravaggio](#), [Hendrick ter Brugghen](#), [Hercules Seghers](#)

Periods: Baroque, Dutch Golden Age painting, Dutch Golden Age

Artwork

[View 15+ more](#)



The Night Watch
1642



The Anatomy Lesson of...
1632



The Return of the Prodigal...
1669

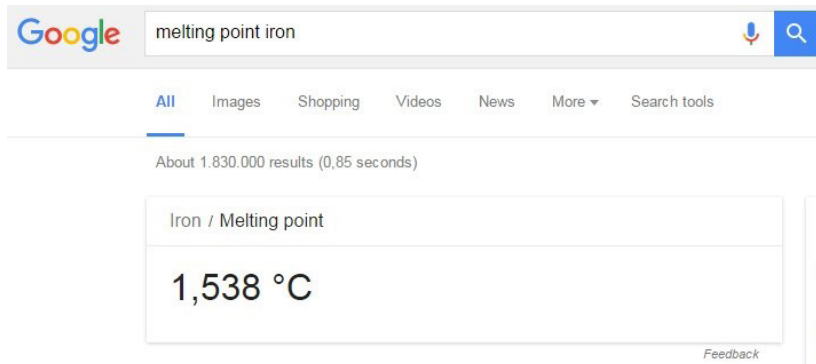


The Storm on the Sea of Galilee
1633



Danaë
1636

Afb. 9.13 - Door Google samengestelde knowledge card voor de schilder Rembrandt.



Afb. 9.14 - Feitelijk antwoord op een zoekvraag op basis van een enkel triplet uit Google's Knowledge Graph.

9.6.4 Schema.org

Schema.org

In de voorgaande paragrafen zijn al diverse voorbeelden van metadataschema's genoemd. Een tamelijk recent voorbeeld dat in het kader van het semantisch web wijdere verspreiding heeft gekregen is Schema.org. Het is vooral ontwikkeld om gestructureerd betekenis toe te kennen aan relevante gegevens die in webpagina's gepresenteerd worden. Die betekenis wordt verwerkt in de codering van webpagina's en blijft daardoor onzichtbaar voor gewone bezoekers van websites. Computerprogramma's kunnen er echter wel gebruik van maken. Deze standaard wordt dan ook actief ondersteund en mee ontwikkeld door de grote webzoekmachines Google, Bing, Yahoo! en Yandex.

Een voorbeeld van de toepassing van Schema.org, is de manier waarop restaurantbeoordelingssite IENS daar gebruik van maakt. In de broncode van de webpagina met beoordelingsgegevens van een Amsterdamse Pizzeria komen onder meer de drie fragmenten voor die in afbeelding 9.15 getoond worden. Daarbij wordt telkens via een URL verwezen naar de gegevens van een bepaald soort "ding" op de Schema.org site. In de afbeelding respectievelijk *Restaurant*, *AggregateRating* en *PostalAddress*. De eigenschappen van die dingen worden telkens gekarakteriseerd via het attribuut *itemprop* bij een HTML-code.

rich snippets

Sommige van die gegevens verschijnen ook in zoekresultaten bij Google, als zogenaamde *rich snippets*. In afbeelding 9.16 zien we daarvan *priceRange*, *ratingValue*, *bestRating* en *ratingCount* terug. Op deze gegevens kan niet gericht gezocht worden, maar ze zijn voor de zoeker wel nuttig bij de keuze welk resultaat aan te klikken. Ze dienen dus de evaluatiefunctie. Zoals we zien worden lang niet alle volgens Schema.org gecodeerde gegevens als *rich snippets* in het zoekresultaat getoond. Welke dat zijn, is een keuze van de betreffende zoekmachine. Wat niet in het zoekresultaat verschijnt, kan (en zal) door de zoekmachine overigens

wel gebruikt worden om de te tonen resultaten zo goed mogelijk te laten aansluiten bij de veronderstelde intentie van de zoekvraag.

```
<section data-restaurant-id="1501" class="restaurantHeader restaurantHeader_hasPhotoSlider restaurantHeader_hasBookingWidget"
itemscope itemtype="http://schema.org/Restaurant">
  <div class="wrapper restaurantHeader_wrapper">
    <div class="restaurantHeader_header absolute">
      <div class="photoSlider" data-plugin="photoSlider">
        <span class="photoItem"></span>
      </div>
      <div class="small-show restaurantHeader_overlay hide-on-swipe">
        <h1 class="titlePrimary restaurantHeader_title" data-restaurant-name>YamYam Trattoria Pizzeria</h1>
        <ol class="labels restaurantHeader_labels fontSmall listStyle-none">
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#keuken" data-restaurant-navigation>Restaurant</a>
</li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#keuken" data-restaurant-navigation
itemprop="servesCuisine" content="Pizzeria">Pizzeria</a></li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/duurzaamheid#restaurantNav" data-restaurant-navigation>69%
Duurzaam</a></li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#prijzen" data-restaurant-navigation
title="Menuprijs" data-plugin="tipsy" itemprop="priceRange" content="22.50"><strong>€</strong> 22.50</a></li>
        </ol><ol class="restaurantHeader_grades scoreList listStyle-none">
          <ol class="restaurantHeader_grades scoreList listStyle-none" itemprop="aggregateRating" itemscope
itemtype="http://schema.org/AggregateRating">
            <li>
              <div class="scoreMedium scoreMediumBetter fontStandOut rating">
                <span class="average" itemprop="ratingValue">8.2</span>
                <span class="isInvisible" itemprop="bestRating">10</span>
                <span class="isInvisible" itemprop="worstRating">1</span>
                <span class="isInvisible" itemprop="ratingCount">34</span>
              </div>
              <p class="fontSmall">Eten</p>
            </li>
          </ol>
          <div class="restaurantWidget_locationBox medium-6 large-12 large-last small-hide" itemprop="address"
itemscope itemtype="http://schema.org/PostalAddress">
            <div class="restaurantWidget_location">
              <div class="mapCanvas" data-map data-marker="[52.376095,4.873566]" data-icon="restaurant" data-zoom="15"
data-mouse-zoom="false" data-re-centre="true"></div>
              <a class="restaurantWidget_locationLargeMap fontSmall" href="/restaurant/1501/amsterdam-yamyam-trattoria-
pizzeria/contact#kaart" data-restaurant-navigation>Grote kaart</a>
              <div class="restaurantWidget_locationText fontSmall">
                <div class="restaurantWidget_locationAddress">
                  <i class="icon mrs left icon-location icon-light"></i>
                  <address><span itemprop="streetAddress">Frederik Hendrikstraat 90</span>,
                    <span class="restaurantWidget_locationCity">
                      <span itemprop="postalCode">1052 HZ</span>,
                      <span itemprop="addressLocality">Amsterdam</span>
                    </span>
                  </address>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</section>
```

Afb. 9.15 - Fragmenten uit de HTML broncode van een webpagina van restaurantbeoordelingssite iens.nl, waarin Schema.org kenmerken volgens de microdata-standaard zijn toegevoegd.

✓ Restaurant YamYam Trattoria Pizzeria in Amsterdam ... - iens.nl

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 8.2/10 - 34 votes - Price range: 22.50

Bij **Pizzeria** restaurant YamYam Trattoria **Pizzeria** in **Amsterdam** eten? ... Noord-Holland; >; **Amsterdam**; >; **West**; >; YamYam Trattoria **Pizzeria** Gemakkelijk een van de betere Italiaanse restaurants in **Amsterdam** vanuit dat oogpunt.

✓ Restaurant Forno in Amsterdam - Recensies | iens.nl - Waar gaan we ...

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 7.7/10 - 11 votes - Price range: 25.00

Word nooit teleurgesteld, een van de lekkerste **pizza** die er in **Amsterdam** te ie wordt opgediend door het mooiste blonde meisje van **Amsterdam-West** zou ...

✓ Restaurant Pazzi II in Amsterdam - Recensies | iens.nl - Waar gaan we ...

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 8.1/10 - 20 votes - Price range: 25.00

Een leuke **pizzeria** voor een snelle hap met de lekkerste pizzabodem tot nog toe! ... Dit is toch zeker wel de lekkerst **pizza** die je in **Amsterdam** kan eten!

Afb. 9.16 - Voorbeeld van een zoekresultaat uit Google waarin enkele van de in afb. 9-15 getoonde gegevens als "rich snippets" getoond worden.

Intussen (augustus 2016) zijn er al meer dan 2,5 miljard webpagina's die Schema.org *markup* bevatten (Bernstein, 2016). In hoofdstuk 13 wordt meer in detail ingegaan op de structuur en eigenschappen van Schema.org.

9.7 Samenvattend

"Kennisorganisatiesystemen" is de algemene benaming voor systemen waarin in enige vorm kennis wordt vastgelegd. Dat hoeft lang niet altijd bedoeld te zijn voor het rechtstreeks toegankelijk maken van informatie. De eerder behandelde taxonomieën en thesauri zijn kennisorganisatiesystemen waarvoor dat wel het geval was. Voor de in dit hoofdstuk geïntroduceerde voorbeelden van semantische netwerken en ontologieën is dat niet het geval. Wel kunnen die beide goed gebruikt worden om zoeksystemen slimmer te maken, zodat ze indirect wel kunnen bijdragen tot een betere toegankelijkheid van informatie.

Voor het formeel vastleggen van kennis, al dan niet in de vorm van metadata, introduceren we in dit hoofdstuk tripels en de RDF standaard, waarmee relaties tussen objecten, concepten en/of data kunnen worden vastgelegd. Uit grote verzamelingen van tripels kunnen hele netwerken van concepten en hun onderlinge relaties worden opgebouwd, zogenaamde *graphs*. Een toepassing van dergelijke *graphs* is onder meer de *Google Knowledge Graph*, die wordt gebruikt om gebruikers bij hun

zoekvragen directe antwoorden en concrete gegevens te verschaffen. Als toepassingen van RDF-tripels worden SKOS (voor het computerleesbaar vastleggen van thesaurusrelaties) en *linked data* (als basis voor het semantisch web) behandeld. Dit hoofdstuk wordt afgesloten met een beschrijving van Schema.org, een metadataschema om gestructureerd semantische informatie aan de inhoud van webpagina's mee te kunnen geven.

Gebruikte literatuur

Becker, P., van der Linden, M., Magrijn, H. & Sieverts, E. (2010). *Organiseer je informatie*; aan de slag met thesauri, taxonomieën, tags en topics. Leidschendam: Biblion. ISBN 9789054839545

Bernstein, A., Hendler, J., Noy, N. (2016). A New Look at the Semantic Web. Te verschijnen in: *Communications of the ACM*, September 2016. Geraadpleegd op: 29 augustus 2016 van: <https://www.merlin.uzh.ch/contributionDocument/download/9193>

De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J. and Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: the Amsterdam Museum case study. In *Extended Semantic Web Conference* (pp. 733-747). Springer Berlin Heidelberg. Geraadpleegd op: 28 juni 2016, van: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.381.2070&rep=rep1&type=pdf>

De Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J. and Schreiber, G. (2013). Amsterdam museum linked open data. *Semantic Web*, 4(3), pp.237-243. Geraadpleegd op: 28 juni 2016, van: http://semantic-web-journal.org/sites/default/files/swj293_0.pdf

de Boer, V. (2016). Connecting collections across national borders. In: *Semantic Web, Digital Humanities and ICT for Development*, 4 juli 2016. Geraadpleegd op: 28 juni 2016, van: <http://www.victordeboer.com/digital-humanities/sound-and-vision/connecting-collections-across-national-borders/>

Doerr, M., Hunter, J., & Lagoze, C. (2003). Towards a Core Ontology for Information Integration. *Journal of Digital Information*, 4(1). Geraadpleegd op: 27 juni 2016, van: <https://journals.tdl.org/jodi/index.php/jodi/article/view/92>

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmman, T., Sun, S. & Zhang, W. (2014). Knowledge vault: A web-scale

approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 601-610). ACM. Geraadpleegd op: 27 juni 2016, van: <https://cs.cmu.edu/~nlao/publication/2014.kdd.pdf>

Frazier, P.J. (2015). SKOS: A Guide for Information Professionals; A Guide to Representing Structured Controlled Vocabularies in the Simple Knowledge Organization System, ALCTS. Geraadpleegd op: 27 juni 2016, van: <http://www.ala.org/alcts/resources/z687/skos>

Garshol, L.M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all. *Journal of Information Science*, 30(4), 378-391. Geraadpleegd op: 27 juni 2016, van: <http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html>

Harper, C.A. (2006). Encoding Library of Congress Subject Headings in SKOS: Authority Control for the Semantic Web. In *International Conference on Dublin Core and Metadata Applications*. 3 - 6 October 2006. Geraadpleegd op: 27 juni 2016, van: <http://dcpapers.dublincore.org/pubs/article/view/842/838>

He, L., Shao, B., Li, Y., & Chen, E. (2015). Distributed real-time knowledge graph serving. In *Big Data and Smart Computing (BigComp), 2015 International Conference on* (pp. 262-265). IEEE. Geraadpleegd op: 27 juni 2016, van: <http://www.graphengine.io/downloads/papers/Trinity.KS.pdf>, <http://www.graphengine.io/downloads/slides/Trinity.KnowledgeServing.pdf>

Hildebrand, Michiel (2015). Linking data sources with Spinque. COMSODE June 18, 2015. Geraadpleegd op: 28 juni 2016, van: <http://www.comsode.eu/index.php/2015/06/linking-data-sources-with-spinque/>

Hill, L., Buchel, O., Janée, G., & Zeng, M.L. (2002). Integration of Knowledge Organization Systems into Digital Library Architectures. In *Proceedings of the 13th ASIST SIG/CR Workshop on "Reconceptualizing Classification Research"* (p. 62-68). Geraadpleegd op 27 juni 2016, van: <https://journals.lib.washington.edu/index.php/acro/article/viewFile/13835/11949>

Isaac, A. (2006a). OWL: Web Ontology Language. *Informatie Professional*, 10(12), 30-33

Isaac, A. (2006b). SKOS: Simple Knowledge Organisation System. *Informatie Professional*, 10(11), 40-43

Isaac, A., & Summers, E. (2009a). SKOS Simple Knowledge Organization System Primer. *W3C Working Group Note*. 18 August 2009. Geraadpleegd op: 27 juni 2016, van: <http://www.w3.org/TR/skos-primer/>

Isaac, A., Wang, S., Zinn, C., Matthezing, H., van der Meij, L., & Schlobach, S. (2009b). Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain. *IEEE Intelligent Systems*, 24(2), 76-86. Geraadpleegd op: 27 juni 2016, van:

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.674.6249&rep=rep1&type=pdf>

Isaac, A., & Haslhofer, B. (2013). Europeana linked open data – data. *europeana.eu. Semantic Web*, 4(3), 291-297. Geraadpleegd op: 28 juni 2016, van:

<http://eprints.cs.univie.ac.at/3732/1/main.pdf>

Miles, A., & Brickley, D. (2005). SKOS Core Guide. *W3C Working Draft*. 2 November 2005. Geraadpleegd op: 27 juni 2016, van:

<http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/>

Miles, A., & Bechhofer, S. (2009). *SKOS Implementation Report*. May 19th 2009. Geraadpleegd op: 27 juni 2016, van: <http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html>

Mols, B. (2008). Digitaal zoeken in meerdere collecties tegelijk. *Informatie Professional*, 12(4), 32-33

Nederlandse Parels van Linked Data Toepassingen (2016). *PlatformLinkedData.nl*

Geraadpleegd op: 27 juni 2016, van:

<http://www.geonovum.nl/sites/default/files/LinkedDataParels.pdf>

Noy, N.F., & McGuinness, D.L. (2005). *Ontology Development 101: A Guide to Creating Your First Ontology*. Geraadpleegd op: 27 juni 2016, van: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Ontologist's Handbook (2016). Cycorp. Geraadpleegd op: 27 juni 2016, van: <http://www.cyc.com/documentation/ontologists-handbook/>

Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, (Preprint), 1-20. Geraadpleegd op: 27 juni 2016, van:

<http://www.semantic-web-journal.net/system/files/swj1167.pdf>

Priem, M., de Boer, V., Hildebrand, M., Oomen, J. & Verplancke, N. (2016). *Eindrapport Gemeenschappelijke Thesaurus voor Uniforme Ontsluiting*. Taalunie. Geraadpleegd op: 28 juni 2016, van:

<http://viaa.be/assets/files/news/downloads/Eindrapport-GTUO-Taalunie.pdf>

Sanchez-Alonso, S. & Garcia-Barriocanal, E. (2006). Making use of upper ontologies to foster interoperability between SKOS concept schemes. *Online Information Review*, 30(3), 263-277

Schie, M. van (2006). RDF: Resource Description Framework. *Informatie Professional*, 10(10), 40-43

SKOS/Datasets - Semantic Web Standards (2015). Geraadpleegd op: 6 september 2016, van:

<https://www.w3.org/2001/sw/wiki/SKOS/Datasets>

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., & Katz, S. (2004). Reengineering Thesauri for New Applications: the AGROVOC Example. *Journal of Digital Information*, 4(4). Geraadpleegd op: 6 september 2016, van:

<https://journals.tdl.org/jodi/index.php/jodi/article/view/112/111>

Streefkerk, Marco (2015). Verbinden van kennisknooppunten met CultuurLINK. *DEN Kenniscentrum Digitaal Erfgoed*. Geraadpleegd op: 28 juni 2016, van:

<http://www.den.nl/blog/bericht/4898>

Trim, Craig (2015). *Semantic Relations in Wordnet*. Geraadpleegd op: 20 juni 2016, van:

<http://trimc-nlp.blogspot.nl/2015/06/python-nltk-and-wordnet.html>

KADER

XML

XML (*extensible mark-up language*) is een coderingswijze om documenten op platform-, medium- en opmaak-onafhankelijke wijze gestructureerd vast te leggen. XML kan worden gebruikt voor zowel tekstdocumenten als voor database-achtige toepassingen. In het eerste geval is het een middel om de structuur van dergelijke documenten vast te leggen. De XMLcode geeft dan bijvoorbeeld aan:

- dat een tekst begint met een titel en de naam van de auteur;
- dat een tekst uit hoofdstukken bestaat die elk met een kop beginnen;
- dat hoofdstukken als inhoud een aantal paragrafen mogen hebben, elk weer beginnend met een kopje en bestaand uit een aantal alinea's.

Daarnaast kun je met XML de betekenis van bepaalde tekstonderdelen aangeven, zoals: de aanduiding dat bepaalde woorden de namen van producten zijn, en andere de namen van bedrijven of geografische aanduidingen en dergelijke.

Bij database-achtige toepassingen dient het vaak als uitwisselingsstandaard tussen verschillende systemen. In het kader van XML kan het woord 'document' dus in heel brede zin worden opgevat.

Onderstaand voorbeeld lijkt op het soort gegevens dat vaak in databases wordt opgeslagen, al is er ook wel enig verschil. Zo bestaat dit adresdocument niet uit een lineaire reeks velden, zoals bij de meeste databases, maar uit een vertakte boomstructuur. Binnen een veld kunnen subvelden voorkomen, daarbinnen ook weer enzovoort. In het volgende (vooral didactisch bedoelde) voorbeeld van een alinea uit een tekstdocument zijn de elementen meer voor semantiek dan voor structuur gebruikt: ze geven betekenis aan woorden of zinsdelen. Dat kan onder meer nut hebben om uit dergelijke teksten specifieke indexen te genereren, om gerichte zoekingen te realiseren en voor automatische opmaak. Het zorgt echter vooral dat kennis over de inhoud op formele wijze in het document zelf aanwezig is. Dit alles maakt uitwisseling en hergebruik van gegevens ook veel eenvoudiger.

```

<adres>
  <naam>
    <voorletters>Marjolein</voorletters>
    <tussenvoeg>van der</tussenvoeg>
    <achternaam>Linden</achternaam>
  </naam>
  <organisatie>Hogeschool van Amsterdam</organisatie>
  <postbus>
    <postbusnr>1025</postbusnr>
    <plaats>
      <postcode>1000 AB</postcode>
      <plaatsnaam>Amsterdam</plaatsnaam>
    </plaats>
  </postbus>
  <bezoekadres>
    <straat>
      <straatnaam>Rhijnspoorplein</straatnaam>
      <nummer>1</nummer>
    </straat>
    <plaats>
      <postcode>1091 GC</postcode>
      <plaatsnaam>Amsterdam</plaatsnaam>
    </plaats>
  </bezoekadres>
  <contact>
    <telefoon><mobiël>06-21157033</mobiël></telefoon>
    <mail>m.van.der.linden@hva.nl</mail>
  </contact>
</adres>

```

Ook in onderstaand voorbeeld zien we dat XML-documenten veel op webdocumenten lijken. We spreken dan ook niet van velden, maar van elementen, net als bij HTML. Anders dan bij HTML kun je hier zelf bepalen hoe je de elementen (tussen punthaken) wilt noemen, waarmee je een bepaald soort documenten wilt coderen.

<alinea>

Dit jaar wordt hier en daar groots gevierd dat <persoon>George Boole</persoon> 200 jaar geleden werd geboren. Dat “daar” is nog niet eens zo zeer in zijn <land>Engelse</land> geboortestadje <plaatsnaam>Lincoln</plaatsnaam>, maar vooral in het <land>Ierse</land> <plaatsnaam>Cork</plaatsnaam>. Daar heeft <persoon>Boole</persoon> met zijn baanbrekende werk de lokale universiteit op de kaart gezet, in een tijd dat daar nog geen <publicatie>Times Higher Education</publicatie> university rankings voor nodig waren. Vanaf <jaartal>1849</jaartal> was hij daar <functie>hoogleraar</functie>. <persoon>Boole</persoon> was inderdaad de man waar onze huidige <vakterm>Booleaanse operatoren</vakterm>

naar genoemd zijn. Hoewel hij nog veel meer wiskundigs op zijn geweten heeft, zoals ontdekkingen op het gebied van differentiaalvergelijkingen en de waarschijnlijkheidsrekening, kennen wij hem natuurlijk vooral van zijn aanzet tot de verzamelingenleer. Onze **<vakterm>**AND- en OR-relaties **</vakterm>** zijn voortgekomen uit het doorsnijden en verenigen van die verzamelingen.

</aline>

Bij een collectie XML-documenten kan een apart bestand horen, een schema of DTD (*document type definition*), waarin de gebruikte elementnamen, hun onderlinge relaties, de documentstructuur en dergelijke zijn vastgelegd. De documenten zelf worden in het Engels *instances* (dus eigenlijk voorbeelden of misschien liever voorkomens) van het betreffende documenttype genoemd. Bij één DTD of schema kunnen zo duizenden of meer van die *instances* horen. Voor allerlei soorten documenten zijn intussen standaard DTD's of schema's ontwikkeld, zoals onder meer voor wetenschappelijke artikelen, literaire werken, archiefbeschrijvingen, breipatronen en (kook) recepten.

Een DTD beschrijft de structuur van een document en, zo je wilt, de functie en betekenis van elementen daarin, maar geeft geen uitsluitel over eventueel gewenste opmaak. Die zal, afhankelijk van de toepassing en het presentatiemedium, telkens anders kunnen zijn. Voor specifieke toepassingen en specifieke uitvoerproducten moet afzonderlijk op een of andere manier, bijvoorbeeld in *stylesheets*, worden aangegeven hoe de verschillende (structurele en inhoudelijke) elementen van de documenten moeten worden vormgegeven. Door gebruik van meer *stylesheets* kan eenzelfde document op heel verschillende manieren worden weergegeven. Door een wijziging in een *stylesheet* kun je de presentatievorm van duizenden documenten tegelijk wijzigen. Door gebruik van DTD's en *stylesheets*, wordt dus in feite een driedeling gemaakt, waarbij vaste documentstructuur, documentinhoud en documentopmaak worden losgekoppeld.

Voor XML is een speciale *stylesheet*-taal ontwikkeld, de *eXtensible Stylesheet Language* (XSL). Het gebruik daarvan gaat veel verder dan de hierboven beschreven (en door het begrip *stylesheet* gesuggereerde) opmaak. In feite is het een formalisme waarmee je kunt beschrijven hoe een XML-bestand, en in het bijzonder alle afzonderlijk elementen daarin, moeten worden getransformeerd of geconverteerd. In handen van programmeurs zijn die *XSL Transformations* (XSLT) een krachtig hulpmiddel, hetgeen verklaart waarom XML ook populair is als uitwisselingsstandaard voor gestructureerde gegevens.

Een paar voorbeelden van toepassingen van XML zagen we bij de eerder genoemde standaard DTD's. Verder zijn onder meer *RSS-feeds* in XML gecodeerd en wordt de in paragraaf 1.4 genoemde MODS standaard voor

metadata in XML weergegeven. Het Open Archives Initiative (OAI) is een toepassing waarmee metadata van digitale documenten via *harvesting* (letterlijk: oogsten) vrij kunnen worden opgehaald, uitgewisseld en hergebruikt. Deze metadata volgens de *Dublin Core* standaard worden in XML gecodeerd. Een algemeen toepasbare standaard om metadata (van welke soort dan ook) in XML weer te geven is het *Resource Description Framework* (RDF) dat in paragraaf 9.4 aan de orde komt.

Hoofdstuk 10

Automatisch metadateren en klasseren

Een belangrijk nadeel dat kleeft aan het handmatig toegankelijk maken van informatie zijn de hoge kosten die daarmee gemoeid zijn, zeker wanneer het om grote aantallen documenten gaat. In principe zijn er twee methoden die kunnen worden toegepast om niet meer alle materiaal handmatig te hoeven indexeren, of op zijn minst de menselijke inbreng veel minder tijdrovend te maken. De ene methode is om helemaal niet te indexeren, maar bij het zoekproces statistische en taaltechnieken toe te passen om de resultaten van *freetext retrieval* te verbeteren. Deze technieken komen in hoofdstuk 11 aan de orde. De andere methode is om de indexering door de computer te laten doen, of door de computer te laten ondersteunen. Deze methode staat centraal in dit hoofdstuk. Omdat beide methoden voor een belangrijk deel berusten op tekstanalyse, is het noodzakelijk dat het om tekst-documenten gaat (en niet bijvoorbeeld om audiovisueel materiaal) en dat van die tekstdocumenten bovendien voldoende tekst digitaal beschikbaar is. De mogelijkheden om beeldmateriaal automatisch inhoudelijk toegankelijk te maken, zijn de laatste jaren verbeterd; al zijn de beschikbare technieken hiervoor nog lang niet volwassen. Voor meer informatie hierover verwijzen we naar het kader “Automatische beeldherkenning”.

10.1 Toepassingen: automatisch ontlenen en automatisch verrijken

Een belangrijk onderscheid dat we kunnen aanbrengen bij geautomatiseerd toegankelijk maken, is het onderscheid tussen automatisch ontlenen en automatisch verrijken van tekst. Het belangrijkste verschil tussen deze twee toepassingen is dat je bij automatisch ontlenen gebruik maakt van woorden die al in de tekst van het document staan en je die specifieke woorden aan de tekst “onttrekt” of als zodanig markeert. Bij automatisch verrijken hoeven deze woorden niet in de tekst te staan en bepaalt de software door middel van regels, slimme taalkundige en statistische technieken, en ook steeds vaker technieken voor *machine learning*, welke termen kenmerkend zijn voor dat document om ze vervolgens ook daadwerkelijk aan het document toe te voegen. Daarbij kan de software de termen uit een informatietaal halen.

10.1.1 Automatisch ontlenen

Bij automatisch ontlenen maak je zoals gezegd gebruik van woorden die al (min of meer) letterlijk in de tekst van het document staan. Specifieke soorten begrippen uit het document (vaak aangeduid als entiteiten) wil je als zodanig markeren. Hiervoor kun je gebruik maken van (autorisatie) lijsten (zie hoofdstuk 5), semantische netwerken, *knowledge graphs* of gespecialiseerde collecties *linked data* (zie hoofdstuk 9) om de tekst mee te vergelijken. Het zal daarbij vrijwel altijd gaan om tamelijk specifieke (soorten) begrippen, zoals:

- namen van (al dan niet bekende) personen;
- namen van producten;
- namen van bedrijven en organisaties;
- plaatsnamen of andere geografische aanduidingen;
- gebeurtenissen of andere tijdsaanduidingen.

De (autorisatie)lijsten kunnen algemeen beschikbare lijsten zijn of lijsten die specifiek voor dit doel zijn gemaakt. Bij deze techniek kunnen makkelijk herkenningfouten worden gemaakt, zeker in het geval van homoniemen. Zo is 'leiden' een plaatsnaam, maar ook een werkwoord. En zo is 'bakker' een beroep, maar kan het ook een naam zijn. Door aan de autorisatielijsten simpele intelligentie toe te voegen in de vorm van herkenningsregels die rekening houden met bijvoorbeeld hoofdlettergebruik, kan een deel van deze herkenningfouten worden voorkomen. Ook bepaalde van de in paragraaf 10.3 genoemde taalkundige technieken kunnen hiervoor worden toegepast. Daarbij kunnen soms ook op basis van zinsontleding bepaalde conclusies getrokken worden. Verder kan met behulp van TF*IDF statistiek, al dan niet in combinatie met gedefinieerde regels (bijvoorbeeld of woorden in de titel staan), worden bepaald welke woorden belangrijk zijn en daarom ontleend worden.

Het nut om woorden die al in een tekst aanwezig zijn expliciet te markeren, ligt in het feit dat daarmee specifieke betekenis aan die woorden wordt gehecht. Zo kun je een gebruiker laten zoeken op alleen persoonsnamen, waarbij zoekresultaten niet vervuild worden met 'bakker' in de betekenis van beroep. Bovendien kan op basis van de ontleende woorden soms een specifieke functionaliteit worden toegevoegd, bijvoorbeeld een koppeling met Google-maps voor geografische aanduidingen of een koppeling met Wikipedia voor achtergrondinformatie over de betreffende entiteit. Ook wordt zo het belang van een ontleende term groter dan dat van een willekeurig ander woord uit de tekst, wat weer gebruikt kan worden om de relevantieberekeningen van een zoekmachine te verbeteren, of om zelfs alleen maar op die woorden te zoeken. Verder kunnen ontleende woorden een samenvatting van de inhoud van het document geven, die

op basis van de tekst is gegenereerd en waarbij de belangrijkste woorden uit een tekst tot lopende zinnen aan elkaar geplakt worden.

Een voorbeeld van een systeem dat deze techniek toepast, is Calais, een door Thomson-Reuters ontwikkeld systeem om automatisch allerlei soorten namen, feiten en gebeurtenissen in teksten te herkennen (Open Calais, 2016). Onder de naam Open Calais is het ook voor anderen beschikbaar (<http://www.opencalais.com/opencalais-demo/>). Afbeelding 10.1 laat het resultaat zien van een analyse van een artikel uit de New York Times, dat in hun demo-site was geplakt. Het voorbeeld laat zien dat niet alleen entiteiten worden herkend, maar dat ook onderwerpen (Engels: *topics*) worden toegekend. Dat is een vorm van verrijking die in de volgende paragraaf aan de orde komt. Met behulp van Calais vult het Australische Powerhouse Museum (Powerhouse, 2016) zijn catalogus aan met trefwoorden die ontleend zijn aan de tamelijk uitgebreide beschrijvingen van objecten uit de collectie (zie afbeelding 10.2). Een ander soortgelijk systeem dat entiteiten herkent is S4 Ontotext (<https://console.s4.ontotext.com/>).

Open Calais Demo

Open Calais demo is best viewed in Google Chrome.

The screenshot displays the Open Calais Demo interface. On the left, there's a sidebar with various filters and a list of entities. The main area shows a news article snippet with highlighted entities and topics.

Language: English

Topics:

Topic	Percentage
Politics	100%
Diplomacy / Foreign Policy (McC)	100%
International / National Security (McLK)	71%
Military Procurement (McN9)	56%
Military Conflicts (Mc2)	41%

Entities:

- ☒ City
 - Tehran, Tehran
 - Vienna, Austria
 - Washington, United States
- ☒ Country
 - Iran
 - United States
- ☒ Industry Term
- ☒ Organization
- ☒ Person
 - Bashar al-Assad
 - John Kerry
 - John Kirby
 - Mohammed Javad Zarif
 - Obama
- ☒ Position

Events & Facts:

WASHINGTON — Iran has been invited to join talks in Vienna this week with Russia, the United States and European nations on whether a political resolution is possible in the Syrian civil war. If Iran accepts, it will be the first time Secretary of State John Kerry will enter formal negotiations with Tehran on issues beyond the nuclear accord reached in July.

Russia has been pressing to include Iran, the only other major power giving military support to President Bashar al-Assad in his effort to remain in power. Senior American officials have begun to acknowledge in recent weeks that no serious discussion of a possible political succession plan in Syria can happen without Tehran's involvement.

But the American denunciation of Iran's activities in Syria, including its support for Mr. Assad's forces and for terror groups like Hezbollah, has always prevented the United States from including Iran in formal talks about the Syrian crisis.

The State Department spokesman, John Kirby, buried that policy at a briefing on Tuesday, before it was announced that Mr. Kerry would attend the meeting on Syria in Vienna on Thursday and Friday. "We anticipate that Iran will be invited to attend this upcoming meeting," Mr. Kirby said.

Afb. 10.1 - Door Open Calais herkende entiteiten in een geanalyseerde pagina uit de New York Times. Soorten entiteiten zijn met verschillende kleuren aangegeven.

Powerhouse Museum Collection Search 2.53

SEARCH | BROWSE CATEGORIES | BROWSE TAGS | BROWSE THEMES | SPECIAL COLLECTIONS | ABOUT | API/DATA

dutch

☐ show objects with images only
☐ only show objects on display in the museum

OBJECT

Womans hat with Bird of Paradise, 1948 - 1949



zoom image

Images: 01 02 03 04 05 06 07 08

Object statement
 Hat, womens, felt / Bird of Paradise, Irian Jaya / The Hague, Holland, 1948-1949

Statement of significance
 The Lesser Bird of Paradise attached to this hat was given to Captain John Zylstra by members of a highland village in 1944, in West Irian (now Irian Jaya) in gratitude for having driven away the Japanese. As a commanding officer with the Dutch East India Army, he was given three birds. He was also presented with three blue crested pigeon crowns and a Samurai sword which has since been returned to the

User tags
 Add tags
 + henry ii X

Related Subjects
 + Birds
 + Dutch East Indian Army
 + Fashion
 + Animal rights

Similar objects
 Clothing and Dress > Hats
 + 2008/33/1 Hat, with pins attached, Games...
 + 2008/140/2 Novelty top hat, 'New Millenn...
 + 2003/144/2 Hat and hatbox, women's croch...
 + 2003/144/2 Hat and hatbox, women's croch...
 + A10561 Cocktail hat, womens, sequins / b...

Auto-generated tags: [Beta]
 City
 + London X
 + the Hague X
 + The Hague X
 Province or state
 + Irian Jaya X
 + Papua X
 Country
 + Australia X
 + East Timor X
 + Indonesia X
 + Papua New Guinea X
 + The Netherlands X

Afb. 10.2 - Voorbeeld van aan een objectbeschrijving ontleende gegevens, waarmee de catalogus van het Powerhouse Museum wordt aangevuld.

Een experimentele toepassing wordt gerealiseerd in het onderzoeksportal van de KB (<http://www.kbresearch.nl/xportal/>). In de gedigitaliseerde historische krantenartikelen en boeken in de Delpher database worden entiteiten herkend op basis van hun voorkomen in onder meer de Wikidata kennisbank. Uit Wikidata en ook uit DBpedia kunnen dan aanvullende gegevens bij die entiteiten worden opgevraagd om ze bij de krantenartikelen te tonen. Bovendien kan het verwijzingen opleveren naar materiaal in bepaalde relevante andere bronnen. Verder kan het systeem op zoek gaan naar tripels waarin de betreffende entiteit voorkomt in een specifiek soort relatie. Als een tekst bijvoorbeeld de naam Mick Jagger bevat en in Wikidata wordt gevonden dat die *{behoortTot}* de Rolling Stones, dan kan die naam van de groep automatisch als relevante extra metadata, als verrijking aan de betreffende tekst worden toegevoegd. Een gebruiker die op Rolling Stones zoekt, vindt dan dus ook een tekst waarin alleen de naam van Mick Jagger genoemd wordt. Deze verrijking van documenten is het omgekeerde van de expansie van zoekvragen die in paragraaf 9.6.2 als voorbeeld van een *linked data* toepassing wordt gegeven, maar heeft wel ongeveer hetzelfde effect.

automatisch verrijken

10.1.2 Automatisch verrijken

In tegenstelling tot automatisch ontlenen, hoeven bij automatisch verrijken woorden niet letterlijk in de tekst te staan. De software bepaalt over het algemeen door middel van taalkundige en statistische technieken en vaak ook door toepassing van *machine learning* welke termen de inhoud van het document representeren, om ze vervolgens ook daadwerkelijk aan het document toe te kennen (zie paragraaf 10.3). Deze als metadata te beschouwen kenmerken zullen bijvoorbeeld in een apart databaseveld worden toegevoegd of, voorzien van specifieke markerings, in het document zelf worden opgenomen. Zulke verrijking kan allerlei verschillende vormen hebben, zoals:

- aanduidingen van één of meer categorieën uit een taxonomie of classificatie waarin het document moet worden ingedeeld;
- trefwoorden of thesaurustermen op basis van een tevoren vastgelegde autorisatielijst of thesaurus;
- andere karakteristieken, zoals een kwalificatie van een in de tekst gegeven oordeel (bijvoorbeeld bij recensies), van een daarin tot uiting komende stemming (*mood detection* of *sentiment analysis*, voor voorbeelden zie paragraaf 10.5) of van het niveau van de tekst of de beoogde doelgroep (ook in bedrijfsomgevingen).

De daarbij toegekende termen of klassen zullen vrijwel altijd uit een gecontroleerde informatietaal afkomstig zijn.

Een zeer eenvoudige vorm van automatisch verrijken, is die waarbij we op basis van eenvoudige, bedachte regels (zie ook 10.2.1.) metadata toe laten kennen. Zo kun je binnen SharePoint op basis van een map waarin een document geplaatst wordt, bepalen welke metadata aan dat document gehangen worden. Bijvoorbeeld: alle documenten die in de map 'Integriteitszaken' worden geplaatst, krijgen de metadata 'Integriteit' en 'HRM' mee (de laatste omdat die afdeling verantwoordelijk is voor integriteitskwesaties).

10.2 Methoden: rule based en getraind met voorbeelden

Of je nu automatisch gaat ontlenen of automatisch gaat verrijken, er zijn twee methoden die je hierbij kunt gebruiken. Bij de *rule based* (regel gebaseerde) methode laat je het systeem automatisch ontlenen of verrijken op basis van vooraf gedefinieerde (kennis)regels; bij het trainingsprincipe bied je het systeem een aantal voorbeelden aan (een trainingsset) op basis waarvan het systeem zelf leert en vervolgens kan ontlenen en verrijken.

kennisregels

10.2.1 Methode 1: Rule based

Bij de *rule based* methode gebruikt een computer (meestal handmatig opgestelde) kennisregels om kenmerken van woorden of van een heel document af te leiden. Zulke kennisregels kunnen gebruikt worden voor:

1. het herkennen van het belang van woorden (op basis van bijvoorbeeld hun positie in een document);
2. het herkennen van de aard van woorden (of het een persoon, plaats of bedrijf is, maar misschien ook wel taalkundig of het een werkwoord, persoonlijk voornaamwoord of zelfstandig naamwoord is);
3. het herkennen waar een document over gaat zodat je kunt verrijken; als het woord 'Olympische spelen' (of nog een van 100 andere sportgerelateerde woorden of woordstammen) er in voorkomt, gaat het over sport,

Die eerste twee toepassingen worden vooral ingezet bij het automatisch ontlenen. Daarvan gaven we in paragraaf 10.1.1 al wat voorbeelden. Voor deze toepassingen volstaan meestal relatief simpele kennisregels zoals: "Bakker met een hoofdletter is een naam"; "leiden als werkwoord is geen plaatsnaam".

De derde toepassing wordt gebruikt voor het verrijken van documenten. Daarbij wordt voor ieder onderwerp een kennisregel opgesteld, met daarin voor het onderwerp karakteristieke woorden en woordstammen. Die kennisregels kunnen tamelijk complex worden, vaak met Booleaanse combinaties waarin bijvoorbeeld allerlei synoniemen verwerkt worden. Daarbij kunnen die woorden ook van gewichtsfactoren voorzien zijn.

Als we bijvoorbeeld een kennisregel willen maken voor 'jaguar', in de betekenis van katachtig dier, dan zou die er als volgt uit kunnen zien:

{Jaguar OR Panthera onca OR Panthera} NOT {auto OR Jaguar Car OR Jaguar XJ OR Jaguar XK OR Jaguar XF OR Jaguar F-Type OR Jaguar XE OR Jaguar F-pace OR Sepecat Jaguar}.

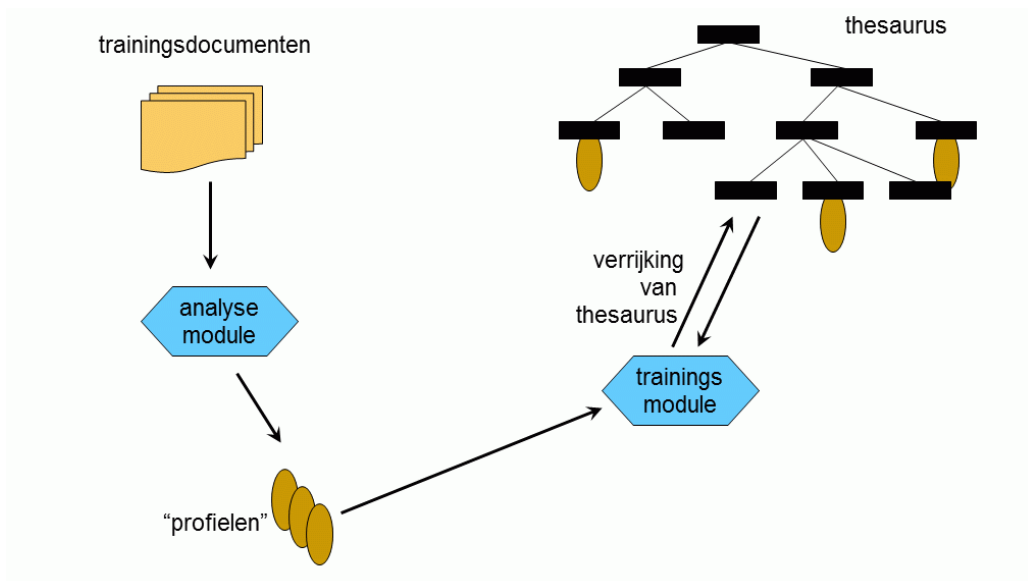
Met deze kennisregel willen we bereiken dat alleen de documenten die over jaguar in de betekenis van het dier gaan, als zodanig worden gekenmerkt en willen we documenten over Jaguar als automerk of als jachtbommenwerper uitsluiten.

trainen

10.2.2 Methode 2: getraind met voorbeelden

Een andere mogelijkheid is om het systeem te trainen met behulp van voorbeelden (zie afbeelding 10.3 en 10.4). Deze methode zet men voornamelijk in bij het automatisch verrijken. In feite is dit een vorm van *supervised machine learning* (zie kader "Machine learning"), maar dan specifiek voor tekstbestanden. Bij de trainingsmethode wordt voor iedere term uit een thesaurus of voor elke klasse uit een taxonomie of

classificatie een trainingsset met relevante documenten aangedragen. Deze documenten worden door het systeem geanalyseerd om daaruit de meest kenmerkende eigenschappen af te leiden. Dat gebeurt met toepassing van statistische en linguïstische technieken (zie paragraaf 10.3). Op die manier worden “profielen” van deze documenten gemaakt, die veelal zullen bestaan uit verzamelingen voor het betreffende onderwerp kenmerkende termen, elk met hun eigen gewicht. Het systeem verrijkt de thesaurus dan als het ware door die documentprofielen samen te voegen tot een profiel voor elk van de betreffende thesaurustermen. Op basis hiervan leert het systeem onderwerpen te herkennen, zodat nieuwe documenten van metadata kunnen worden voorzien.



Afb. 10.3 - Schematische weergave hoe profielen van trainingsdocumenten worden gebruikt om aanwezige thesaurustermen te verrijken met kenmerkende termen (Naar een voorbeeld van Joop van Gent - Irion).

vectorrepresentatie

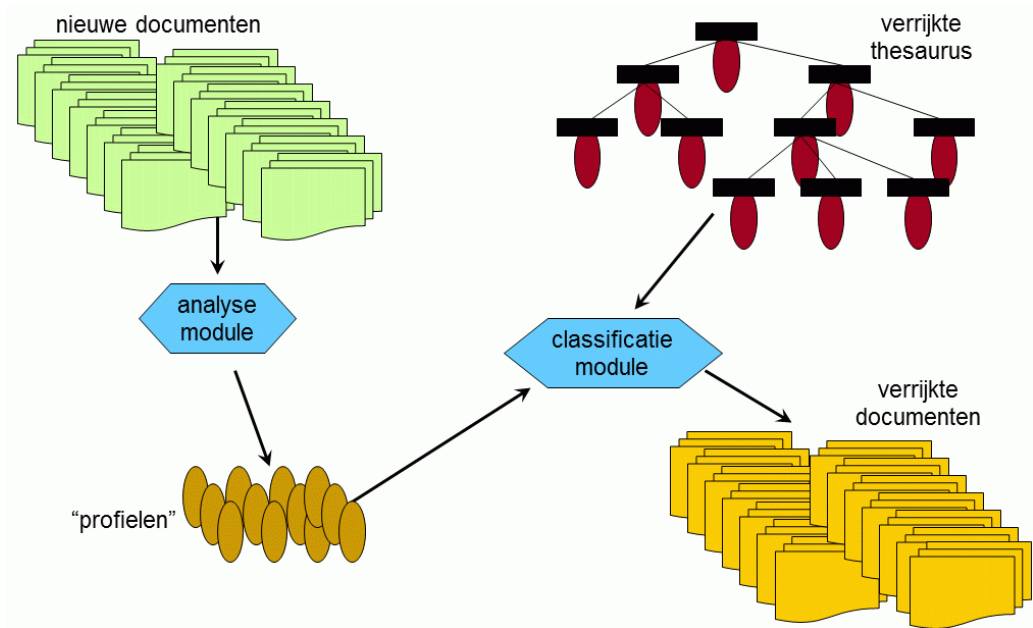
In het classificatieproces (afbeelding 10.4) moeten uiteindelijk termen of categorieën aan nieuwe documenten worden gekoppeld. Dit gebeurt door de van die documenten gemaakte profielen te vergelijken met de profielen van alle aanwezige thesaurustermen en daar de meest gelijkende uit te kiezen. Deze vergelijking kan bijvoorbeeld plaats vinden door vermenigvuldiging van een vectorrepresentatie van het profiel van een document met die van de categorieën (zie kader “Vectormodel”), waaruit de computer de beste match kiest.

classificatieproces

Bij het classificatieproces kunnen meestal drempelwaarden worden ingesteld voor de waarschijnlijkheid dat een toekenning ook werkelijk

**betrouwbaar-
heidsdrempel**

betrouwbaar is. Bij een hoge betrouwbaarheidsdrempel zal een hoog percentage van de geklasseerde documenten (vaak ruim meer dan 90%) in de correcte klasse(n) terechtkomen, maar het systeem zal ook een flink percentage (soms meer dan 20%) wegens onzekerheid terzijde leggen, zodat die documenten alsnog handmatig verwerkt moeten worden. Ook moet soms worden aangegeven of slechts één term of klasse mag worden toegekend of dat dat er meer mogen zijn.



Afb. 10.4 - Schematische weergave van het classificatieproces. Profielen van nieuwe documenten worden vergeleken met de getrainde (verrijkte) thesaurustermen en op basis daarvan geklasseerd. (Naar een voorbeeld van Joop van Gent - Irion).

De methode van het trainen met voorbeelddocumenten zal zeker aantrekkelijk zijn wanneer een collectie digitale documenten beschikbaar is, die eerder al handmatig is ontsloten op basis van de te gebruiken taxonomie of thesaurus. In een dergelijk geval kan het trainingsproces grotendeels geautomatiseerd worden.

Voor het trainen van documenten zullen meestal minimaal tien documenten per klasse als trainingsmateriaal nodig zijn, maar soms worden hiervoor ook wel grotere aantallen genoemd. Elk van die documenten dient voldoende karakteristiek te zijn voor de daarmee te trainen klasse. De tekst daarin moet ook voldoende homogeen zijn. Dat wil zeggen dat ze geen overzicht moet geven van een aantal verschillende onderwerpen die je bij het klasseren juist wilt onderscheiden. Dus bijvoorbeeld geen overzichtsartikel over allerlei sporten, als je wilt dat

het systeem die afzonderlijk leert klasseren. Het trainingsmateriaal zal dus enigszins selectief uit de al ontsloten documenten gekozen moeten worden. Als een dergelijke selectie door tijdgebrek niet kan plaats vinden, moet je er rekening mee houden dat een grotere trainingsset nodig is om toch tot betrouwbare toekenningen te kunnen komen. Wanneer nog geen ontsloten materiaal beschikbaar is, moet er rekening mee worden gehouden dat het trainen van het systeem een tijdrovend karwei kan zijn. Dat is zeker het geval bij een omvangrijke thesaurus, maar zelfs een taxonomie met bijvoorbeeld 500 klassen vereist al dat minimaal 5.000 voldoende karakteristieke documenten worden geselecteerd en handmatig worden ingedeeld of ontsloten.

10.3 Technieken: statistisch en linguïstisch

statistische methoden

TF*IDF

Met statistische methoden (vooral relatieve woordfrequenties) kan een computerprogramma bepalen wat de inhoudelijk belangrijkste en meest karakteristieke woorden uit een document zijn. Een veelgebruikte methode is om van elk woord uit een document de waarde TF*IDF te berekenen. Dat wil zeggen dat woorden alleen kenmerkend zijn voor een document, als ze daarin vaker voorkomen, maar tegelijk in weinig andere documenten aanwezig zijn. Zie voor meer informatie het kader "TF*IDF". Deze techniek kan zowel een rol spelen bij systemen die via training verrijken, als bij die welke het via regels (Engels: *rule based*) doen. Bij systemen die via training werken, kunnen de TF*IDF-waarden van de woorden uit een tekst de basis zijn voor de in de vorige paragraaf genoemde documentprofielen waarmee een systeem getraind wordt op het bepalen welke categorie of thesaurusterm moet worden toegekend. Bij *rule based* systemen beperkt de rol van deze techniek zich tot het eventueel identificeren van belangrijke woorden waarop de regels al dan niet toegepast moeten worden.

linguïstische methoden

Met linguïstische methoden kan onder meer een soort normalisatie van aanwezige termen worden bewerkstelligd, zodat diverse varianten van woorden of begrippen voor het berekenen van een TF*IDF gelijkgesteld kunnen worden. Dat kan door:

woordstam

- reductie van woorden tot hun morfologische woordstam (Engels: *stemming*), zodat voor de statistiek geen onderscheid meer wordt gemaakt tussen enkel- en meervoud, vervoegingen, verbuigingen e.d.;

decompounding

- zogenaamde *decompounding* van samengestelde woorden in hun losse componenten, zodat het systeem weet dat bijvoorbeeld een hockeytoernooi ook een toernooi is en dat woorddeel ook in de woordstatistiek kan worden meegenomen en datzelfde voor de component hockey;

syntactische analyse

- syntactische analyse, zodat equivalentie tussen zinsdelen en losse woorden kan worden vastgesteld (bijvoorbeeld: ‘besparing van energie’ = ‘energiebesparing’);

semantische kennis

- beschikbaarheid van semantische kennis over mogelijke betekenisrelaties tussen woorden, bijvoorbeeld welke woorden (vrijwel) synoniem zijn en voor de woordstatistiek samengenomen kunnen worden.

Bij systemen die met training werken zorgt dit voor meer universele profielen voor het trainings- en herkenningsproces. Daarbij blijkt het in de praktijk niet altijd nodig alle vier hier genoemde technieken toe te passen om toch voldoende bevredigende resultaten te krijgen. Welke het meest van belang zijn, hangt vaak af van de taal van de te karakteriseren documenten. Bij *rule based* systemen kunnen deze technieken toegepast worden, zodat het niet nodig is voor ieder woord uit een kennisregel ook bijvoorbeeld enkel- en meervoud, vervoegingen en verbuigingen in een OR-relatie op te nemen.

Een aantal van bovenstaande technieken wordt ook in slimme zoeksystemen toegepast en komt in paragraaf 11.2 nog iets uitgebreider aan de orde.

10.4 Kwaliteit van het resultaat

Hoewel ook met automatisch ontlenen het systeem fouten kan maken (Google Scholar ziet een hoofdstukop “V. Conclusions” op basis van bepaalde opmaak soms als een auteursnaam), moet vooral bij automatische verrijking goed getest worden of de resultaten als voldoende beschouwd mogen worden. Dat geldt zowel bij *rule based* systemen als bij systemen die gebruik maken van trainingsdocumenten. Om te bepalen of het systeem goede resultaten oplevert, zal het vaak worden uitgetest op een collectie testdocumenten. Men zal dan controleren of het classificatie-proces daarbij tot de gewenste resultaten leidt. Daarna volgt een iteratief proces van aanpassen, testen en weer aanpassen.

Bij het samenstellen van een testset moet met de volgende aspecten rekening gehouden worden:

- Bepaal hoe de testset het beste samengesteld kan worden. Is er wellicht metadatering op basis waarvan een selectie gemaakt kan worden, of moeten de testdocumenten handmatig worden uitgezocht?
- Bepaal per taxonomiecategorie welke documenten die categorie goed vertegenwoordigen. Voeg die documenten toe aan de testset en geef aan (voor zover dat niet al gebeurd is) welke categorie ze vertegenwoordigen.

- Houd er rekening mee dat documenten in meer categorieën ingedeeld kunnen en mogen worden. Voeg bewust documenten toe waarvoor dat het geval is.
- Zorg voor diversiteit in de documenten: varieer zoveel mogelijk in lengte, taalgebruik, bestandsformaten, bronnen, etc.

Leg in een testplan vast wanneer de classificatietest geslaagd genoemd kan worden. Bepaal ook welke (soorten) documenten in ieder geval goed geklasseerd moeten worden en voor welke documenten dat minder belangrijk is.

Voor het al dan niet correct klasseren worden (in iets aangepaste betekenis) ook wel de begrippen vangst en precisie gebruikt. Hierbij is voor elke categorie:

- $\text{vangst} = \frac{\text{aantal geklasseerd en relevant}}{\text{totaal aantal relevant}}$;
- $\text{precisie} = \frac{\text{aantal geklasseerd en relevant}}{\text{totaal aantal geklasseerd}}$.

Daarbij kan de volgende tabel worden gebruikt:

Aantallen per categorie	Relevant	Niet relevant	Totaal
Geklasseerd	Correct	Niet correct	Totaal geklasseerd
Niet geklasseerd	Niet correct	Correct	Totaal niet geklasseerd
Totaal	Totaal relevant	Totaal niet relevant	

Als een voor een categorie relevant document niet als zodanig geklasseerd is, is dat niet correct (een *false negative*). Als een document wel geklasseerd is, maar voor een onderwerp waarvoor dat document niet relevant is, is dat ook niet correct (een *false positive*). Deze cijfers moeten voor elke taxonomiecategorie worden bepaald.

Realiseer je hierbij wel dat 100% vangst en precisie nooit gehaald kunnen worden. In de praktijk zal dus de afweging gemaakt moeten worden tussen:

- Meer documenten geklasseerd in een categorie, zodat daarvoor zoveel mogelijk relevante documenten geklasseerd zijn. De niet-relevante documenten die ook in de betreffende categorie geklasseerd zijn, worden daarbij voor lief genomen (hogere vangst, lagere precisie, weinig *false negatives*).
- Minder documenten geklasseerd in een categorie, waarvan wel zoveel mogelijk daarvoor relevant zijn. De relevante documenten die niet in de betreffende categorie geklasseerd zijn, worden daarbij voor lief genomen (lagere vangst, hogere precisie, weinig *false positives*).

Als uit de test blijkt dat de resultaten niet goed genoeg zijn, moet gekeken worden of de niet-correct geklasseerde documenten (zowel de *false positives* als de *false negatives*) aanknopingspunten bieden wat er in de opgestelde regels of bij het trainen mis gegaan kan zijn. Voor verbetering kan soms worden volstaan met het verhogen of verlagen van bepaalde drempelwaarden. Anders zal het voor elke slecht presterende klasse nodig zijn om daarvoor de classificatieregels aan te passen of om daarvoor aanvullende of vervangende training uit te voeren, met betere trainingsdocumenten. Voor een uitgebreider beschrijving van het hele implementatietraject van een systeem voor automatische verrijking, zie Aalten (2011).

Een situatie waarin het niet zo belangrijk is dat elk document van alleen correcte categorieën wordt voorzien, is die waarbij het systeem alleen als hulpmiddel dient voor menselijke indexeerders. In zo'n geval wordt een betrekkelijk groot aantal in aanmerking komende klassen aan de menselijke indexeerder gepresenteerd, waaruit die een selectie kan maken. In de praktijk blijkt dat je daarmee beter, sneller, vollediger en consistentere categorieën of termen kunt toekennen. Dat speelt zeker in situaties waar het om niet-professionele indexeerders gaat. Zo krijgen de redacteurs van KankerNL (zie hoofdstuk 13) suggesties voor toe te kennen thesaurustermen, die ze vervolgens kunnen overnemen, negeren en/of aanvullen.

Of automatische verrijking van voldoende kwaliteit gerealiseerd kan worden, is mede afhankelijk van:

1. de kwaliteit van de gebruikte taxonomie of thesaurus: in het bijzonder de aanwezigheid van zoveel mogelijk consistente, eenduidige en onderscheidende categorieën met zo weinig mogelijk inhoudelijke overlap;
2. de omvang van de documentcollectie: hoe meer documenten, hoe beter;
3. de homogeniteit van de collectie en het domein: hoe beperkter het aantal onderwerpen in een collectie, hoe beter;
4. de homogeniteit van de documenten: hoe minder verschillende onderwerpen binnen een document, hoe beter;
5. de veranderlijkheid van het domein: hoe minder aan verandering onderhevig, hoe beter;
6. de kwaliteit van de taal in de documenten: hoe minder straat-, sms- of gesproken taal, hoe beter;
7. de gelijksoortigheid van de documenten: hoe minder verschillende soorten documenten, hoe beter;
8. de domeinexpertise binnen de organisatie: hoe meer kennis van het domein bij opbouw en onderhoud, hoe beter;
9. de diversiteit in talen: hoe minder verscheidenheid in talen, hoe beter.

Als illustratie bij het eerstgenoemde punt kan een bij de UB Utrecht uitgevoerd experiment dienen. Daarin bleek bijvoorbeeld dat het

gebruikte systeem nauwelijks onderscheid kon maken tussen artikelen op het terrein van psycholinguïstiek, die waren ontsloten met de trefwoorden 'leren lezen', 'dyslexie' en 'spelling'. Na training bleken profielen voor deze termen (misschien ook niet echt verwonderlijk) onvoldoende onderscheidend. Dit illustreert dat orthogonaliteit (weinig inhoudelijke overlap) tussen klassen uit de taxonomie of thesaurus een eis is voor succesvol automatisch klasseren.

De ervaring van een leverancier leert dat ontsluiting met dit soort systemen nog goed kan werken voor informatietalen met maximaal 5000 klassen. Anderzijds bleek uit jarenlange ervaring met automatisch indexeren ten behoeve van de PHYS database (Engelstalige bibliografische database met natuurkunde-artikelen) dat goede resultaten behaald konden worden op basis van een thesaurus van bijna 20.000 termen (Biebricher, 1988). Dat systeem diende echter vooral als ondersteuning van menselijke indexeerders, zodat eventuele dubbelzinnigheden eenvoudig handmatig rechtgezet konden worden. Ervaring daar leerde dat in de loop van de tijd hertraining van termen nodig bleek om het systeem aan te passen aan geleidelijk veranderend vocabulaire en zwaartepunten. Ook werden bevredigende resultaten gemeld voor het klasseren van *health resources* op internet met behulp van de MeSH, een zeer uitgebreide medische thesaurus die in hoofdstuk 13 kort wordt besproken.

Van de te klasseren informatie dient ook een voldoende hoeveelheid tekst digitaal beschikbaar te zijn (punt 2). Wat in dit verband voldoende is, kan per domein of documenttype verschillen. Voor teksten met zeer concreet en specifiek taalgebruik kunnen korte samenvattingen van tien regels al voldoende zijn. In andere gevallen, als bijvoorbeeld tekst met het karakter van een samenvatting ontbreekt, zal meer tekst digitaal beschikbaar moeten zijn. Van in oorsprong niet digitaal beschikbare documenten kan door scannen en optische karakterherkenning (OCR) alsnog een computerleesbare versie gegenereerd worden. Ook het web is een veelgebruikte proeftuin door de aanwezigheid van grote hoeveelheden digitale informatie. Nadeel daarbij is dat aan het daar beschikbare materiaal maar zelden termen of categorieën uit een vaste informatietaal zijn toegekend, zodat trainen en testen niet altijd eenvoudig mogelijk is. Studies waarin diverse aanpakken worden vergeleken of waarin de resultaten van soms grote aantallen onderzoeken aan automatische verrijking worden samengevat, zijn onder andere gepubliceerd door Jindal (2015), Dalal (2013), Dasari (2012) en Baharudin (2010).

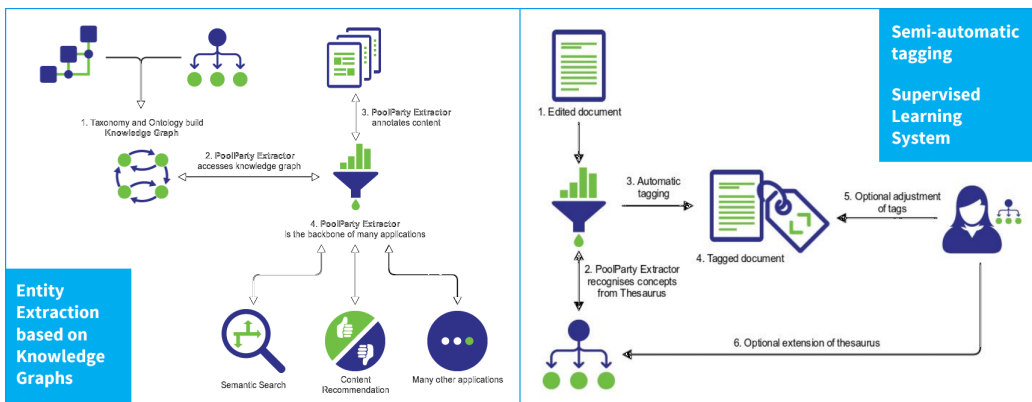
Om de hier beschreven methoden voor automatisch verrijken te kunnen toepassen, is een noodzakelijke voorwaarde dat er al een informatietaal als een classificatie of taxonomie, dan wel een thesaurus of ander gecontroleerd ontsluitingssysteem beschikbaar is. Wanneer (nog) geen classificatie of thesaurus beschikbaar is, kunnen de hier besproken methoden om profielen van digitale documenten te maken ook worden gebruikt om te bepalen welke documenten op elkaar lijken. Door toepassing van methoden uit

de retrieval-techniek, zoals het vectormodel kan dan getracht worden de hele documentcollectie in clusters uiteen te laten vallen (zie kader “Vectormodel”). In de praktijk blijkt die clustering vaak geïnitieerd te moeten worden met specifieke documenten die als kernen voor te vormen clusters moeten dienen. Op grond van woorden die voorkomen in de profielen van de documenten in zo’n cluster, kan een systeem vaak ook al een suggestie doen voor de omschrijving van die categorie. Experimenten met deze techniek zijn uitgevoerd op zelfs zeer grote informatiecollecties van wel 20 miljoen webpagina’s (Golub, 2006).

10.5 Praktijkvoorbeelden van automatische verrijking

In deze paragraaf presenteren we wat voorbeelden van systemen en toepassingen voor automatisch categoriseren. Daarbij ook systemen die je zelf online kunt uitproberen, ontwikkelingen op bepaalde onderwerpsterreinen, algemene toepassingen en wat commerciële producten.

Een voorbeeld hoe automatische verrijking in de praktijk toegepast wordt, is Climate Tagger (www.climate>tagger.net). In dit initiatief werken verschillende partijen wereldwijd samen om content over klimaat en klimaatverandering te delen, te ontsluiten en beschikbaar te stellen. Climate Tagger verzamelt hierbij ongestructureerde content uit de databases van aangesloten partijen. Vervolgens ontleent het op basis van *knowledge graphs* relevante termen en concepten, zoals geografische locaties, personen en organisaties. Daarnaast wordt de Climate Thesaurus ingezet voor semi-automatisch taggen: het systeem kent thesaurustermen toe aan documenten, maar gebruikers kunnen deze indien gewenst nog aanpassen.



Afb. 10.5 - Schematische weergave van de werking van de bij Climate Tagger gebruikte software; links voor entity extraction, rechts voor verrijking met thesaurustermen. (Ontleend aan PoolParty documentatie <https://www.poolparty.biz/>).

Afbeelding 10.5 toont schema's van de werking van de hierbij gebruikte PoolParty software. Je kunt het systeem uitproberen door stukken tekst in Engels, Spaans, Frans, Portugees of Duits in een testvenster in te plakken (<http://api.reegle.info/demo/>). Afbeelding 10.6 geeft een voorbeeld van het resultaat daarvan.

[DEMO](#)
[FAQ](#)
[DOCUMENTATION](#)
[GET API TOKEN](#)
[BACK TO CLIMATE TAGGER SUITE](#)

[back](#)

Extracted Tags

✓ Extracted Concepts

Concept	Score
▼ arctic ecosystem	
▼ global climate	
▼ temperature	
▼ climate change	
▼ wind	
▼ melting sea ice	

Synonyms:
no Synonyms

Associated Top Concepts:

- climate change theory

Containing Concept Schemes:

- Climate Compatible Development Glossary

Broader Transitive Concepts:

- climate change theory

Related:

- melting of permafrost
- melting snowpack
- antarctic ecosystem
- arctic ecosystem
- polar ecosystems

URI:
<http://reegle.info/glossary/1491>

▼ weather	
▼ melting	
▼ tops	
▼ water resources	

Locations

Country	Location
	Paris (City)

Concepts
A concept is a recognized phrase from the [Climate Thesaurus](#). Semantic information is available for each extracted concept. Via the API such additional information includes synonyms, related concepts, and higher level concepts. More data still can be requested through climatetagger's [SPARQL endpoint](#).

Locations
Locations can be either cities or countries which come along with latitude / longitude information.

Terms
Terms are plain keywords extracted using statistical algorithms such as frequency and position in a text.

Afb. 10.6 - Resultaat van de door analyse van een ingeplakt nieuwsbericht door Climate Tagger gesuggereerde thesaurustermen. Van de term 'melting sea ice' is de thesauruspresentatie opengeklapt.

Een andere toepassing is die waarbij het Ministerie van BZK papieren archieven scant en automatisch klasseert. Dat systeem wordt in paragraaf 13.8 uitgebreider beschreven.

Wie zelf iets wil uitproberen, kan gebruik maken van de demo ACT-DL, waar je Engelse of Duitse teksten kunt laten categoriseren op basis van de Dewey Decimale Classificatie (<http://act-dl.base-search.net/>). Je kunt daar kiezen om een Dewey-code te laten toekennen aan:

- een uitgeknipt stuk tekst (via *text categorizer*);
- de inhoud van een webpagina (URL via *web categorizer*);
- een PDF-document (upload via *PDF categorizer*).

Voor de meest waarschijnlijke codes geeft het systeem berekende betrouwbaarheidsscores.

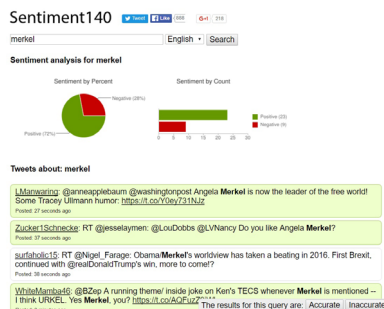
De International Press Telecommunications Council (IPTC) werkt aan een *Classification Engine* om nieuwsartikelen automatisch volgens de eigen taxonomie te classificeren. Daarbij is gekozen voor een *rule based* oplossing, die onder de naam EXTRA (EXTRaction Rules Apparatus) in een open-source-omgeving wordt ontwikkeld (<https://iptc.github.io/extra/>).

In het bedrijfsleven bestaat veel belangstelling voor automatische *sentiment detection* of *sentiment analysis*, om snel te kunnen analyseren hoe, vooral in de sociale media, over een bedrijf, organisatie, product, persoon of onderwerp wordt gesproken. Een paar vrij toegankelijke voorbeelden daarvan zijn:

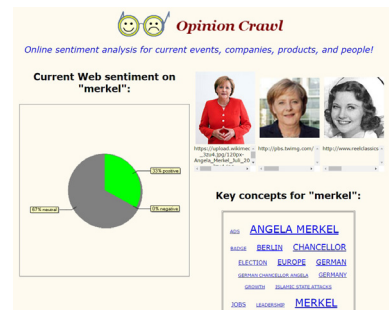
- Python NLTK Text Classification (<http://text-processing.com/demo/sentiment/>) voor het laten analyseren van eigen stukken Engelse, Nederlandse of Franse tekst van maximaal 50.000 tekens, die je in een venster kunt inplakken. Het systeem bepaalt de mate van subjectiviteit (neutraliteit ↔ polariteit) en van de polariteit het percentage positief ↔ negatief. Dit systeem is gebaseerd op *machine learning*.
- Sentiment140 (<http://www.sentiment140.com/>) analyseert berichten op Twitter, op basis van een door de gebruiker ingevoerde zoekterm. Je moet daarvoor inloggen met een Twitter-account. De analyse vindt plaats op de meest recente, Engelstalige of Spaanstalige, tweets die voor de zoekterm gevonden zijn. De percentages positief en negatief hebben alleen betrekking op de tweets waarvoor een sentiment bepaald kon worden (en tellen dus altijd op tot 100%). Dit systeem is gebaseerd op *machine learning*.
- Opinion Crawl (<http://www.opinioncrawl.com/>) analyseert nieuwsberichten uit Engelstalige bronnen, op basis van een door de gebruiker ingevoerde zoekterm. Het geeft percentages voor neutraal, positief en negatief.

- Social Searcher (<http://www.social-searcher.com/>) zoekt in een grote verscheidenheid aan sociale media (o.a. Twitter, Google+, YouTube, Instagram, Tumblr, Reddit, Vine, Dailymotion, VKontakte). Onder een tabblad *Analytics* toont het voor elk van die systemen apart het sentiment van de gevonden berichten. Het werkt alleen goed voor Engelstalige berichten.

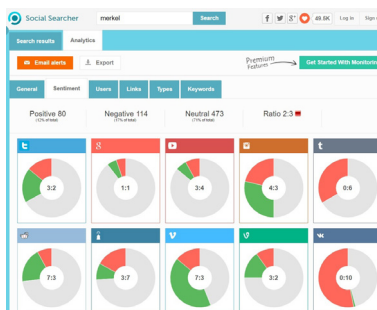
De voorbeelden in afbeelding 10.7 laten zien dat de uitkomsten van de verschillende systemen niet altijd erg eensluidend zijn. Deels zal dat komen door de verschillende bronnen die gebruikt worden, deels door verschillen in de toegepaste techniek.



a



b



c

Afb. 10.7 - Voorbeelden van de resultaten van automatische sentiment analyse. Als zoekterm was de naam "Merkel" gebruikt.

a: Sentiment140; b: OpinionCrawl; c: Social-Searcher.

Octrooien vormen een documentsoort die, op grond van een tamelijk vaste structuur en door beschikbaarheid van een algemeen gebruikt classificatiesysteem, in aanmerking komt om automatische categorisatie op uit te testen. De groep van Suzanne Verberne van de Radboud Universiteit heeft hier onderzoek aan uitgevoerd (Verberne, 2011; D'hondt, 2012, 2013). Het daar ontwikkelde systeem kan ook voor andere soorten tekstmateriaal worden toegepast (Verberne, 2016).

Een terrein waar ook veel wordt geëxperimenteerd met toepassingen voor automatische verrijking, is de geneeskunde. Enerzijds is er op dat terrein een zeer omvangrijke bibliografische database beschikbaar. Medline bevat meer dan 25 miljoen wetenschappelijke artikelen, met goed gestructureerde samenvattingen, waaraan al medische *subject headings* uit de MeSH-thesaurus zijn toegekend. Daar kan dus vrij makkelijk zowel trainings- als testmateriaal aan ontleend worden. Anderzijds verschijnen er op dat terrein zoveel nieuwe publicaties dat automatische technieken voor het toekennen van *subject headings* al snel tot grote tijd- en geldbesparing kunnen leiden. De Amerikaanse National Library of Medicine, de samensteller van Medline, gebruikt zelf inderdaad de Medical Text Indexer om menselijke indexeerders te voorzien van suggesties voor toe te kennen termen (Mork, 2013). Studies waarin diverse methoden getest en resultaten daarvan vergeleken werden, zijn onder andere gedaan door Yepes (2014) die verschillende technieken voor *machine learning* heeft uitgetest, en Trieschnigg (2009) die resultaten van een zestal verschillende systemen heeft vergeleken. Ook voor andere soorten van medische informatie op internet is automatische indexering met MeSH-terminen een populaire proeftuin, zoals we in paragraaf 10.4 al aangaven.

Algemene terreinen waarvoor deze technieken kunnen worden ingezet, zijn verder onder meer:

- spam filtering in E-mail;
- email routing (afhankelijk van het onderwerp doorsturen naar een specifieke mailbox);
- genre classificatie;
- *readability assessment* (vinden van geschikt materiaal voor verschillende niveaus, leeftijden of soorten van lezers).

Voorbeelden van producten waarin automatische categorisatie wordt toegepast zijn onder meer:

- LexisNexis: voor zijn omvangrijke nieuwsdatabases worden Engelstalige krantenartikelen automatisch van trefwoorden voorzien;
- NorthernLight: zoekmachine software die functionaliteit biedt om de interne informatie van bedrijven automatisch op onderwerpscategorieën in te delen;
- Smartlogic: idem;
- HP/Autonomy: idem.

10.6 Samenvattend

In dit hoofdstuk bespraken we de verschillende toepassingen, methoden en technieken voor automatisch metadateren en klasseren. Enerzijds maakten we daarbij onderscheid tussen toepassingen die ontlenen en

toepassingen die verrijken. Bij die eerste soort toepassingen worden woorden of begrippen die al in een document aanwezig zijn, als kenmerkend daarvoor aangewezen. Bij toepassingen die verrijken, worden documenten voorzien van termen of klassen uit een informatietaal zonder dat de betreffende woorden in de documenten zelf hoeven voor te komen. Anderzijds maakten we onderscheid tussen methoden die berusten op formele kennisregels die voor het ontlenen of verrijken zijn opgesteld, en methoden die automatisch leren door ze te voeden met grote aantallen voorbeelden via zogenaamde machine learning technieken. Daarnaast gingen we ook kort in op de hierbij toegepaste statistische en linguïstische technieken.

Bij afweging tussen handmatige en automatische ontsluiting kan een aantal factoren meespelen:

- stijgende kosten van handmatige indexering;
- een groeiende hoeveelheid te ontsluiten materiaal;
- de naar verwachting nog verder verbeterende kwaliteit van automatische indexering;
- de vraag of 100% correcte ontsluiting voor alle ontsloten materiaal even belangrijk is.

De afweging of alle documenten altijd 100% correct ontsloten moeten zijn, of dat een 80-20 regel (met 20% inspanning 80% goed resultaat) acceptabel is, zal voor elke organisatie en elke collectie anders kunnen uitvallen. Dat geldt ook voor de vraag in hoeverre tevoren kan worden ingeschat welk materiaal zodanig belangrijk is, dat daarvoor handmatige controle en nabehandeling vereist zijn. Dit zijn zaken die tevoren zorgvuldig moeten worden afgewogen voordat aan geautomatiseerde metadatering en klassering wordt begonnen.

Hoe dan ook is het belangrijk om je te realiseren dat automatisch metadateren en klasseren nooit geheel automatisch zal gaan en dat er altijd handwerk benodigd zal zijn.

Ook moet worden bedacht dat, wanneer uitsluitend geautomatiseerde technieken worden toegepast, alleen nog digitaal beschikbaar materiaal wordt ontsloten en niet-digitaal materiaal inhoudelijk onvindbaar blijft, tenzij dit materiaal gescand en door OCR computer-leesbaar gemaakt wordt. Anderzijds is de verwachting dat ook van steeds meer fysieke documenten een voldoende hoeveelheid inhoudelijke tekst, zoals inhoudsopgave, samenvatting en flaptekst van boeken (of zelfs de hele inhoud als e-book) digitaal beschikbaar komt, om geautomatiseerde ontsluiting daarop te baseren.

Verder lezen en gebruikte literatuur

Aalten, J. van (2011). Automatisch classificeren. In: *Handboek Informatiewetenschap IWABase, IV-F-900 1-16*. VakMediaNet, Alphen aan de Rijn

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1), 4-20.. Geraadpleegd op: 23 november 2016, van: <http://www.jait.us/uploadfile/2014/1223/20141223050800532.pdf>

Biebricher, P., Fuhr, N., Knorz, G., Lustig, G., & Schwantner, M. (1988). The Automatic Indexing System AIR/PHYS; from Research to Application. In *11th International Conference on Research and Development in Information Retrieval* (p. 333-342). Geraadpleegd op: 23 november 2016, van: http://www.is.inf.uni-due.de/bib/pdf/ir/Biebricher_etal_88.pdf

ClimateTagger (2016). About ClimateTagger. Geraadpleegd op: 24 november 2016, van: <http://www.climate>tagger.net/about-climate-tagger/>

Dalal, M. K., & Zaveri, M. A. (2013). Automatic classification of unstructured blog text. *Journal of Intelligent Learning Systems and Applications*, 5(2), 108. Geraadpleegd op: 23 november 2016, van: http://file.scirp.org/pdf/JILSA_2013052208594771.pdf

Dasari, D. B. (2012). Text categorization and machine learning methods: current state of the art. *Global Journal of Computer Science and Technology*, 12(11-C). Geraadpleegd op: 23 november 2016, van: <http://computerresearch.org/index.php/computer/article/viewFile/555/555>

D'hondt, E., Verberne, S., Weber, N., Koster, C., & Boves, L. (2012). Using skipgrams and pos-based feature selection for patent classification. *Computational Linguistics in the Netherlands Journal*, 2, 52-70. Geraadpleegd op: 23 november 2016, van: http://www.clinjournal.org/sites/clinjournal.org/files/4Dhondt2012_0.pdf

D'hondt, E., Verberne, S., Koster, C., & Boves, L. (2013). Text representations for patent classification. *Computational Linguistics*, 39(3), 755-775. Geraadpleegd op: 23 november 2016, van: <http://anthology.aclweb.org/J/J13/J13-3009.pdf>

Golub, K. (2006). Automated subject classification of textual web documents. *Journal of Documentation*, 62(3), 350-371. Geraadpleegd op: 23 november 2016, van: <http://www.diva-portal.org/smash/get/diva2:747753/FULLTEXT01.pdf>

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Webology*, 12(2), 1. Geraadpleegd op: 23 november 2016, van: <http://www.webology.org/2015/v12n2/a139.pdf>

Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013). The NLM Medical Text Indexer System for Indexing Biomedical Literature. In *BioASQ@ CLEF*. Geraadpleegd op: 23 november 2016, van: <http://bioasq.org/sites/default/files/Mork.pdf>

OpenCalais (2016). *About Open Calais*. Geraadpleegd op: 23 november 2016, van: <http://www.opencalais.com/about>

Powerhouse (2016). *Powerhouse Museum | Science + Design / Sydney Australia*. Geraadpleegd op: 24 november 2016, van: <http://www.powerhousemuseum.com/collection/database/menu.php>

Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11), 1412-1418. Geraadpleegd op: 23 november 2016, van: <http://publications.tno.nl/publication/101222/yf8ete/trieschnigg-2009-mesh.pdf>

Verberne, S., & D'hondt, E. (2011). Patent Classification Experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In *CLEF (Notebook Papers/Labs/Workshop)*. Geraadpleegd op: 23 november 2016, van: <http://sverberne.ruhosting.nl/papers/verberne-CLEF-IP2011.pdf>

Verberne, S. (2016). Automatische classificatie van teksten. *Presentatie bij: Kennisbijeenkomst taal- en spraaktechnologie voor de archiefsector*. 5 september 2016. Geraadpleegd op: 23 november 2016, van: <http://www.slideshare.net/suzanv/automatische-classificatie-van-teksten>

Yepes, A. J., MacKinlay, A., Bedo, J., Garnavi, R., & Chen, Q. (2014). Deep belief networks and biomedical text categorisation. In *Proceedings of the Twelfth Annual Workshop of the Australasia Language Technology Association* (Vol. 123). Geraadpleegd op: 23 november 2016, van: http://www.aclweb.org/old_anthology/U/U14/U14-1017.pdf

KADER

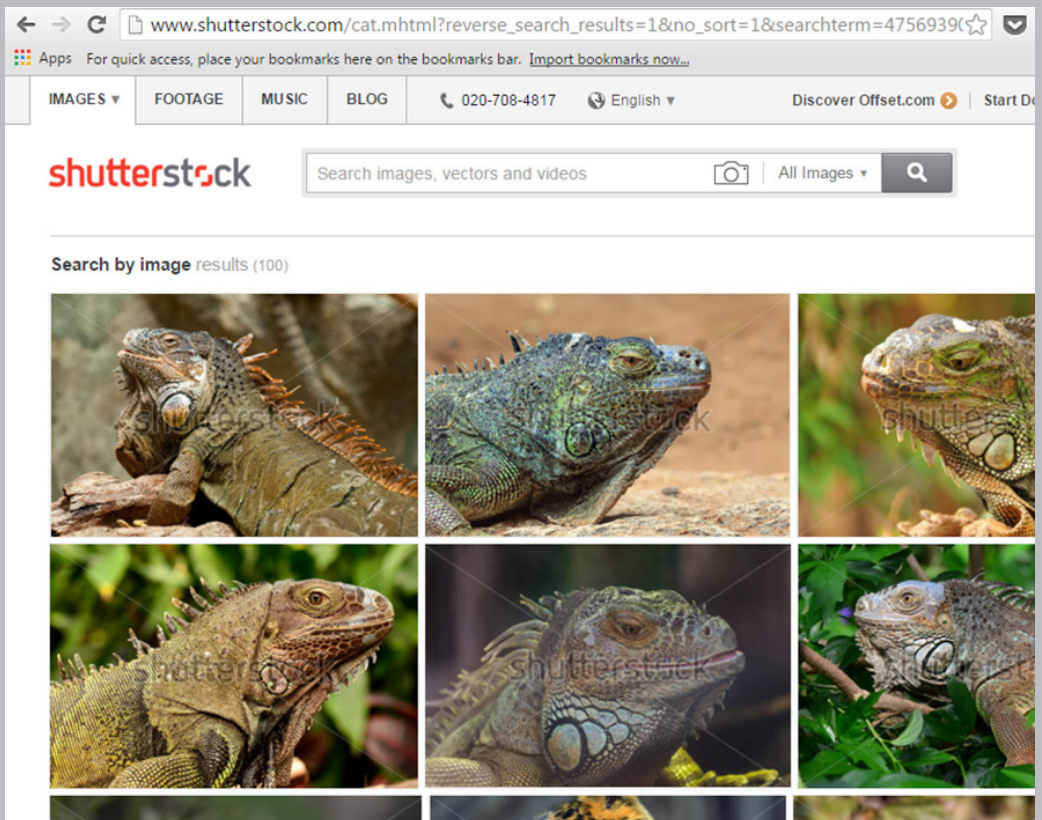
Automatische beeldherkenning

In hoofdstuk 10 bespreken we vormen van geautomatiseerde ontsluiting van documenten. Voor beeldmateriaal zijn intussen ook dergelijke technieken ontwikkeld. Het belangrijkste doel daarvan is om de voorstelling op beeldmateriaal automatisch te herkennen en de juiste metadata toe te kennen. Door toepassing van *machine learning* technieken (zie kader) worden op dat terrein de laatste tijd aardige resultaten geboekt. Wel beperken die zich over het algemeen nog tot de pre-iconografische analyse: de software herkent vormen als een paard, man of auto, maar nog maar heel zelden de betekenis of symboliek van een afbeelding op iconografisch of iconologisch niveau (zie kader “Beeldmateriaal”). Meer specifieke herkenning van bijvoorbeeld welke persoon op een foto staat, of welk merk auto, is hoogstens pas op gespecialiseerde systemen mogelijk en dan meestal alleen voor een beperkt aantal mogelijke keuzes. Zo kan Facebook de gezichten leren herkennen van de personen die regelmatig voorkomen op de foto’s die je daar uploadt. Overigens gaat de ontwikkeling op dat gebied wel verder, omdat men uit alle gezichten die op een bewakingscamera langskomen ten behoeve van surveillance en beveiliging, die van al bekende (of zelfs nog onbekende) criminelen of terroristen wil kunnen herkennen. Wie wat meer over technische achtergronden van gezichtsherkenning wil weten, moet maar eens Geitgey (2016) lezen. Dat is een aflevering uit een hele serie bijdragen over diverse toepassingen van *machine learning*.

De simpelste vorm van beeldherkenning is die waarbij niet zo zeer herkend wordt wat een afbeelding voorstelt, maar of afbeeldingen op elkaar lijken. Er zijn intussen al veel systemen op internet waar je een daarmee eerder gevonden afbeelding kunt markeren, een URL van een afbeelding kunt opgeven of een eigen afbeelding kunt uploaden, waarna je andere afbeeldingen terugkrijgt die daar in meer of mindere mate op lijken. Globaal wordt dat voor twee soorten toepassingen ingezet:

- Het vinden van andere afbeeldingen van hetzelfde onderwerp. De gevonden afbeeldingen hoeven dan alleen maar een zekere gelijkenis te hebben met de afbeelding waarvan wordt uitgegaan;
- Het vinden van vrijwel identieke afbeeldingen. Dit wordt ook wel reverse image search genoemd. Dit kan onder meer dienen om de herkomst van foto’s te achterhalen of ongeoorloofd hergebruik op te sporen. Het is ook een veel gebruikte techniek voor debunking van gemakzuchtig fotogebruik in de media en van tendentieuze onjuiste toeschrijving van foto’s op sociale media (Reid, 2016). Zie de voorbeelden in afbeeldingen 2 en 3.

De grote zoekmachines Google, Bing en Yandex bieden deze optie in hun reguliere *image search*. Als zij geen vrijwel exacte kopie kunnen vinden, kan de mate van gelijkenis van het wel gevonden materiaal vaak sterk uiteenlopen. Ook een site als Pinterest biedt deze mogelijkheid uitgaande van al gevonden afbeeldingen. Twee andere voorbeelden zoeken specifiek in collecties van stockfoto's: Shutterstock (<http://www.shutterstock.com/cat.mhtml>) in zijn eigen collectie, Pixolution (<http://demo.pixolution.de/>) in de collectie van van Fotolia. In een test, uitgaande van eigen foto's, kwamen daar heel redelijke resultaten uit (afbeelding 1).



Afb. 1 - Het resultaat van een zoekactie bij Shutterstock, waarbij een eigen foto van een leguaan als uitgangspunt genomen werd.

Een zoekmachine die alleen gespecialiseerd is in *reverse image search*, het vinden van vrijwel exacte kopieën, is Tineye (<http://www.tineye.com/>). Het voorbeeld in afbeelding 2 gaat uit van een voor een advertentie van een bekende supermarkt bewerkte foto waar Mark Rutte via de tuin van het Witte Huis Obama lijkt te bezoeken. Het resultaat toont ook andere versies van dezelfde foto, zonder Rutte, maar met een joggende Michelle Obama en met een lege tuin (vermoedelijk het origineel).



TinEye

Reverse Image Search





3 Results

Searched over **2.2011 billion** for file: <http://www.adformatie.nl/in>

- These results expire in **72 hours**. [Why?](#)
- [Share a success story!](#)
- TinEye is [free](#) to use for non-commercial purposes.

JPEG, 1571x2350, 1.9 MB

Sort by:

Best Match

Most Changed

Biggest Image



[www.freerepublic.com gallery 42 8 2837](http://www.freerepublic.com/gallery/4282837)

[Compare](#) | [Link](#)

JPEG Image
250x375, 91.6 KB



[theweek.com 26308 article](http://theweek.com/article/26308)

[Compare](#) | [Link](#)

JPEG Image
240x301, 70.5 KB

Your Image

Switch

Image Match



Your Image

Switch

Image Match



Compare your submitted image to your result image. Quickly click the 'Switch' button to switch back and forth between the two images and highlight any differences. This works best when one of the images has been modified.

Afb. 2 - Van een bewerkte foto uit een advertentie vindt Tineye andere, niet of anders bewerkte versies.

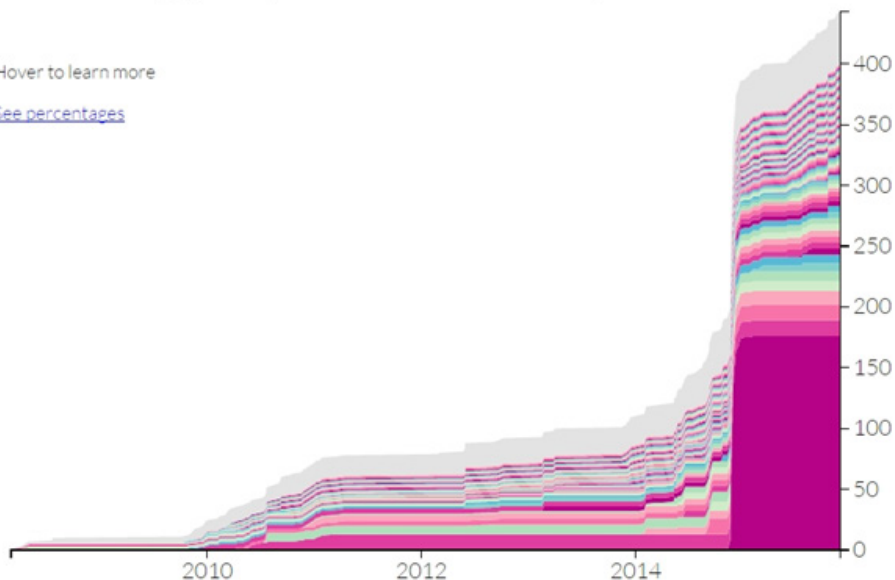
This image has appeared on
143 websites and **443**
unique pages.



Since first appearing in **March 2008** on
fadhil.blogg.se, this is how it has spread online.

Hover to learn more

[See percentages](#)

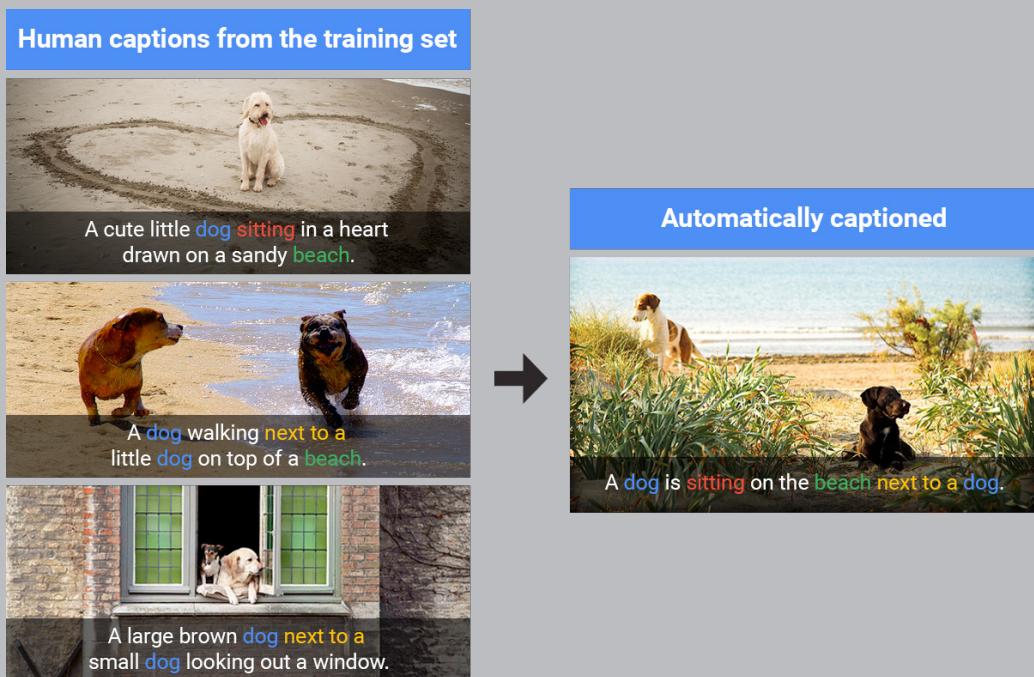


Afb. 3 - Where-on-the-net toont de verspreidingsgeschiedenis van een bij die dienst aangeboden foto.

Een dienst die specifiek gericht is op de analyse van beelden die op sociale media opduiken is Where on the net (<https://www.whereonthe.net/>). Afbeelding 3 laat zien hoe een foto die zich in 2015 viraal verspreidde als een voor recent doorgaande foto van vluchtelingen, in werkelijkheid al uit 2008 bleek te dateren.

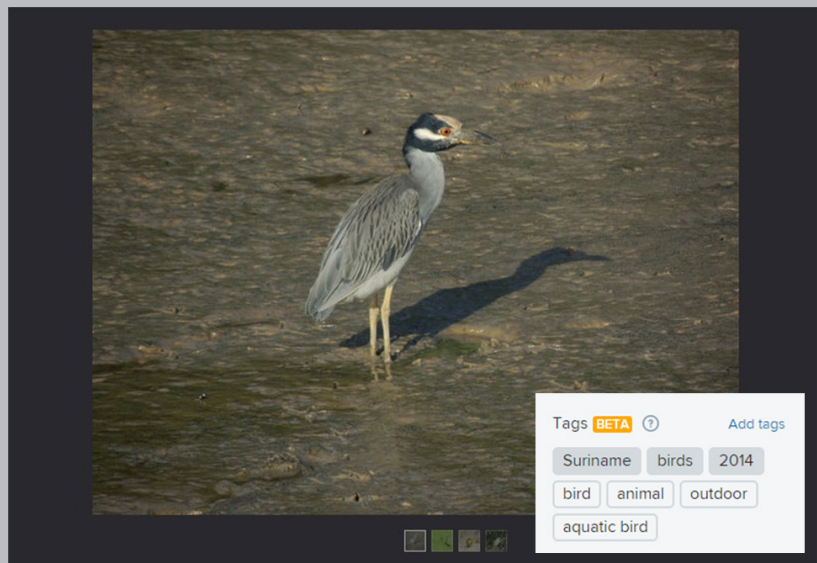
Een belangrijke stap verder zijn de systemen die ook echt kunnen herkennen wat op een foto is afgebeeld. De basistechniek daarvoor lijkt sterk op die voor het categoriseren van tekstdocumenten, zoals beschreven in hoofdstuk

10. Door een systeem een aantal voorbeelden van afbeeldingen van een bepaald onderwerp te geven, en eventueel ook een aantal negatieve voorbeelden, leert het systeem dat onderwerp ook op nieuwe afbeeldingen te herkennen. De wijze van analyse voor beeldmateriaal is uiteraard wel een heel andere dan die voor tekstdocumenten. Het gaat nu niet om een analyse van woorden, maar van pixels. Bij deze zogenaamde *feature extraction* wordt een veelheid aan karakteristieken van een afbeelding geanalyseerd. Dat leidt in feite ook tot profielen of vingerafdrukken van die afbeeldingen, zoals die in hoofdstuk 10 voor tekstdocumenten aan de orde kwamen. De eerste generaties herkenningsssoftware hebben nog veel input voor het leerproces nodig. Voor elk te herkennen onderwerp is een flink aantal zorgvuldig geselecteerde voorbeeldafbeeldingen nodig. Ook dan nog kan zo'n systeem alleen die voorstellingen herkennen waarop het getraind is. Een volgende generatie herkenningsssoftware maakt daarom gebruik van grote corpora met al (inhoudelijk) beschreven materiaal, waarin in principe bijna alle onderwerpen al voorkomen. Die beschrijvingen kunnen bijvoorbeeld de tags of zinnenetjes zijn die op uploadsites als Flickr aan foto's zijn toegekend.



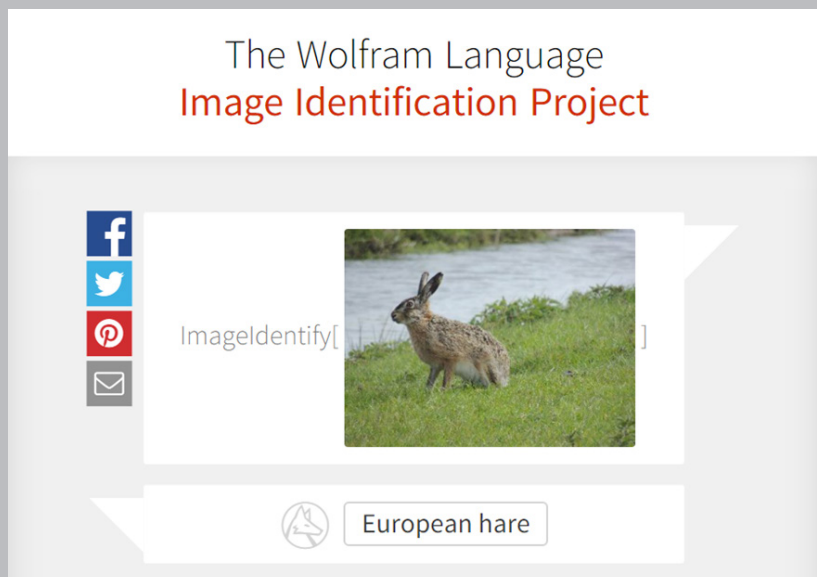
Afb. 4 - Illustratie hoe Google elementen uit andere foto's herkent en onderdelen uit beschrijvingen van die foto's combineert voor een beschrijving van een nieuwe foto.

Er zijn nog wel meer veelgebruikte systemen die voorstellingen herkennen. Een voorbeeld daarvan is de foto-uploadsite Flickr (<https://www.flickr.com/>). Naast door uploaders gegeven tags, voegt Flickr zelf ook tags toe op basis van automatische beeldherkenning. Dat zijn vaak algemene omschrijvingen als 'bike', 'people', 'animal', 'outdoor' en soms uiterlijke kenmerken als 'photo border' voor een foto met witte rand, of een sfeerkenmerk als 'serene' (wat al bijna als iconologisch kenmerk beschouwd kan worden). Een foto met veel water krijgt bijvoorbeeld drie kenmerken mee 'lake', 'creek', 'watercourse'; daarvan is er altijd wel eentje goed. Foto-eigenaars kunnen onterecht toegekende tags verwijderen of het toekennen zelfs helemaal uitzetten. Helaas kun je de door de machine bedachte tags bij een zoekactie niet uitsluiten. Aan de tags is wel te zien welke door de mens en welke door de machine zijn toegekend (afbeelding 5).



Afb. 5 - Voorbeeld van tags bij een foto op Flickr; met witte achtergrond de door Flickr automatisch toegekende tags.

Door de eerder genoemde stockfoto site Shutterstock worden voorstellingen van foto's eveneens op basis van *machine learning* herkend (Lopez, 2016). Ook de semantische zoekmachine Wolfram Alpha probeert in haar *image identification project* daadwerkelijk termen toe te kennen aan afbeeldingen die door gebruikers worden geüpload. In het voorbeeld in afbeelding 6 is te zien dat een (in Nederland gemaakte) foto van een haas correct wordt beschreven met de term 'european hare'. Toch leidt ook Wolfram niet in alle gevallen tot correcte resultaten. Voorbeelden daarvan zijn te vinden in Lowensohn (2015).



Afb. 6 - Een zelfgemaakte foto van een haas, wordt door Wolfram correct herkend.

Uit de bovengenoemde voorbeelden blijkt dat de technieken voor automatische beeldherkenning op dit moment nog tamelijk foutgevoelig zijn. Bovendien zal het in de toekomst lastig blijven om de herkenning precies te laten aansluiten bij het doel en de specificiteit van de gewenste toegankelijkheid van beeldmateriaal. Ten behoeve van gerichte zoekacties zal voorlopig dus ook nog behoefte blijven bestaan aan handwerk en aan goede afspraken over diepte en niveau van indexering, en over daarbij te gebruiken metadataschema's en informatietalen.

Verder lezen

Geitgey, A. (2016). Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning. *Medium*, 24 July 2016. <https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78>

Lopez, Napier (2016). Shutterstock now uses machine learning to help you find just the right photo. *The Next Web*, 10 March 2016. <http://thenextweb.com/insider/2016/03/10/961047/>

Lowensohn, Josh (2015). Wolfram has created a website that will identify any image you throw at it - and it's learning as you do it. *The Verge*, 13 May 2015. <http://www.theverge.com/2015/5/13/8603531/wolfram-image-identification-site-trained-by-chewbacca>

Pachal, Pete (2015). Google Photos identified two black people as 'gorillas'. *Mashable*, 1 July 2015. <http://mashable.com/2015/07/01/google-photos-black-people-gorillas/>

Reid, Alastair (2016). 4 verification case studies from #IJF16. *First Draft*, 9 April 2016. <http://firstdraftnews.com/4-verification-case-studies-from-ijf16/>

Shallue, Chris (2016). Show and Tell: image captioning open sourced in TensorFlow. *Google Research Blog*, September 22, 2016. <https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html>

KADER

Machine learning

Machine learning technieken worden al meer dan tien jaar gebruikt voor automatische classificatietaken, maar zijn anno 2016 in veel ruimere context een *buzzword* geworden. Door de beschikbaarheid van steeds meer rekenkracht vindt deze tak van kunstmatige intelligentie steeds vaker toepassing en ook voor steeds complexere problemen. Die problemen worden gekenmerkt door de aanwezigheid van zo grote aantallen parameters en afhankelijkheden dat die niet meer allemaal afzonderlijk vooraf kunnen worden onderscheiden en niet meer gemodelleerd kunnen worden met een hanteerbaar aantal “als ... dan ...” regels, zoals dat bij *rule based* systemen het geval is. Het idee van *machine learning* is om een computerprogramma het dan zelf maar helemaal te laten uitzoeken.

Het voornaamste onderscheid dat daarbij gemaakt kan worden, is dat tussen *supervised* en *unsupervised learning*. *Supervised learning* houdt in dat een systeem wordt getraind door het een heleboel voorbeelden te geven en er telkens bij te vertellen waar het een voorbeeld van is. Bij de toepassing voor documentclassificatie moet er dus al een taxonomie zijn, waarin de documenten ingedeeld kunnen worden en moeten de voorbeelddocumenten op basis daarvan zijn ingedeeld. Voor andere toepassingen kan een systeem getraind worden door handelingen heel vaak voor te doen, zodat een robot of een zelfrijdende auto vanzelf leert hoe die handelingen gedaan moeten worden. Het door IBM ontwikkelde systeem Watson is met vragen en antwoorden uit het populaire Amerikaanse TV-spel Jeopardy getraind, waarna het de eerdere superwinnaars van het spel met gemak versloeg. Nu wordt Watson vrij beschikbaar gesteld om daarmee nieuwe toepassingen te ontwikkelen, bijvoorbeeld op medisch gebied.

Bij *unsupervised learning* moet het computerprogramma het echt helemaal zelf uitzoeken. Daarbij leert het bijvoorbeeld om in beeldmateriaal voorwerpen te herkennen waarop het niet tevoren getraind is en waarvoor ook nog geen taxonomie hoeft te bestaan. Of er worden in een grote hoeveelheid materiaal clusters van op elkaar lijkende documenten geïdentificeerd, die op grond daarvan potentiële categorieën vormen, ook al waren tevoren nog geen categorieën bedacht.

Voor deze technieken gebruikt Google ook wel de benaming *deep learning*. In het najaar van 2015 heeft dit bedrijf de hiervoor gebruikte software als Open Source beschikbaar gesteld onder de naam TensorFlow. Zelf gebruikt het bedrijf hiermee ontwikkelde specifieke toepassingen. Een voorbeeld is Rankbrain dat zorgt dat het systeem geleidelijk steeds beter leert wat

zoekers met hun vragen bedoelen, zodat de ranking van zoekresultaten daarop aangepast (en dus verbeterd) wordt. DeepMind is gespecialiseerd in spellen en heeft intussen zo goed het denkspel Go leren spelen, dat het de wereldkampioen heeft verslagen.

Automatische categorisering of classificatie van tekstdocumenten is een toepassing van *machine learning* waarop in dit boek wat uitgebreider wordt ingegaan (hoofdstuk 10). Het herkennen van onderwerpen in beeldmateriaal (ook voor classificatie) is een andere toepassing die daar even genoemd wordt. Voor dat laatste worden weer heel andere, niet zo makkelijk in eenvoudige termen te beschrijven, methoden toegepast dan voor tekstclassificatie. Een paar voorbeelden van andere toepassingen zijn:

- natuurlijke taaltechniek, zoals die waarmee Google uit een zoekvraag van een gebruiker probeert af te leiden wat die gebruiker echt bedoelt;
- medische expert systemen die artsen kunnen ondersteunen bij het stellen van diagnoses;
- software voor automatische spraakherkenning;
- software voor automatische vertaling;
- de zelfrijdende auto;
- *predictive policing*, waarbij op basis van oude gegevens over crimineel gedrag, voorspellingen over (de kans op) toekomstige inbraken worden gedaan.

De technieken of algoritmes die bij *machine learning* worden toegepast zijn heel divers. Een paar voorbeelden daarvan worden hier genoemd. Omdat dit zeer specialistische informatica is, beperken we ons tot de namen, zonder verdere uitleg over de precieze werking te geven. Het zijn echter wel de technieken die ook in niet-technische literatuur over dit onderwerp wel genoemd worden:

- neurale netwerken;
- *decision trees*;
- *regression algorithms*;
- *k-nearest neighbour*;
- *Bayesian networks*;
- *support vector machines*;
-

Elk soort toepassing heeft zijn eigen aanpak en optimale daarvoor te gebruiken techniek. Voor een toepassing waarbij tekst (dus letters en woorden) geanalyseerd moet worden om daaruit af te leiden waar die tekst over gaat, moet bijvoorbeeld een andere techniek worden toegepast dan bij de analyse van beeldmateriaal waar alleen maar pixels beschikbaar zijn. Welke techniek voor een bepaalde toepassing het beste werkt kan nog niet nauwkeurig voorspeld worden. In de praktijk moet dus nog veel worden uitgetoetst.

Verder lezen

https://en.wikipedia.org/wiki/Machine_learning

<http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>

https://en.wikipedia.org/wiki/List_of_machine_learning_concepts

<https://hbr.org/2015/07/what-every-manager-should-know-about-machine-learning>

De serie “Machine learning is fun” van Adam Geitgey op Medium.com
(<https://medium.com/@ageitgey>)

Hoofdstuk 11

Verbetering van zoeken met natuurlijke taal

ontleend
indexeren
freetext retrieval

Inhoudelijke ontsluiting van documenten was een vanzelfsprekende noodzaak zolang die documenten zelf niet in digitale vorm beschikbaar waren. Ontsluiting diende dan vooral het doel de documenten terug te kunnen vinden. We bespraken al eerder dat een steeds groter deel van de inhoud van documenten, artikelen, rapporten en boeken op zijn minst gedeeltelijk digitaal beschikbaar is, waardoor deze tekst met zoekmachine- of retrievalsoftware doorzoekbaar gemaakt kan worden. Daarbij worden in principe alle woorden uit de tekst in een zoekindex opgenomen. Dat is dus de meest vergaande vorm van ontleend indexeren. Veel mensen verwachten dat met deze methode van *freetext retrieval* in principe alles terug te vinden is. Je zou dit het Google-effect kunnen noemen, omdat iedereen eraan gewend is geraakt via webzoekmachines altijd voldoende relevante informatie te kunnen vinden. Ver voor de opkomst van Google, in de jaren '80 van de vorige eeuw bestond die verwachting overigens al. In een beroemd experiment van Blair & Maron (1985) bleken advocaten op een groot Amerikaans advocatenkantoor in hun interne fulltext zoekstelsel nog geen 20% van de relevante informatie terug te vinden, terwijl ze zelf inschatten dat dat minstens 75% zou zijn. Het missen van meer dan 80% van de voor de zaak relevante documenten was in de Amerikaanse rechtspraak ook toen al volstrekt onacceptabel.

Intussen zijn we 30 jaar verder, zodat de vraag gerechtvaardigd is of *freetext* zoektechnieken intussen zo verbeterd zijn dat wel in elke situatie de gewenste informatie met voldoende vangst en precisie terug te vinden is. Daarbij is van belang dat er een levensgroot verschil is tussen het web, waar je in vele honderden miljarden webpagina's en documenten altijd wel iets relevant kunt vinden, en een intranet of database met enkele tienduizenden of honderdduizenden documenten, waar men zich niet kan veroorloven met een zoekactie een bepaald relevant document te missen, zoals bij de door Blair & Maron onderzochte gebruikers. Van de problemen die eenvoudige methoden van *freetext retrieval* in dit verband opleveren, kunnen sommige intussen voor een deel met technische middelen worden opgelost. Daar komt vooral veel taaltechnologie bij kijken. Dit hoofdstuk geeft een overzicht van dergelijke technieken.

vangst
precisie

11.1 Problemen met eenvoudige freetext retrieval

Met de eerste generatie *freetext* retrievalsystemen werden alleen documenten gevonden die exact de combinatie van zoekwoorden bevatten, zoals die in de zoekvraag waren ingetikt. Dat leverde een aantal bekende problemen op met betrekking tot vangst en precisie. Los van echte fouten die zoekers kunnen maken, zien we daarvoor vaak de hieronder genoemde oorzaken.

Oorzaken voor het missen van relevante informatie:

- In relevante documenten komen andere woordvormen (enkel-/meervoud, vervoegingen, verbuigingen) voor, dan het exacte woord uit de zoekvraag.
- In relevante documenten komen woorden in andere spelling voor.
- In relevante documenten zijn synoniemen van de zoekwoorden gebruikt.
- In relevante documenten komen specifiekere woorden voor dan het meer algemene begrip waarmee is gezocht.
- Relevante documenten zijn in een andere taal dan de woorden waarmee wordt gezocht.
- Van relevante documenten is zo weinig tekst digitaal beschikbaar en dus doorzoekbaar, dat lang niet alle in de publicatie behandelde onderwerpen en invalshoeken in die digitale tekst gerepresenteerd zijn, zoals in veel bibliotheekcatalogi (Sieverts, 2009).

Oorzaken voor het vinden van niet-relevante informatie (de in paragraaf 2.6.2 genoemde ruis):

- De letterreeks die als zoekwoord was gebruikt komt in gevonden documenten in een andere betekenis voor dan die welke door de zoeker bedoeld was; dit probleem van homoniemen of homografen kom je vooral bij acroniemen veelvuldig tegen.
- Gebruikte zoekwoorden komen in andere, niet bedoelde context voor.
- Het feit dat een document aan een Booleaanse AND-combinatie van twee zoekwoorden voldoet, wil alleen zeggen dat de zo gecombineerde woorden samen in dat document voorkomen. Maar dat is nog geen garantie dat ze daarin ook de inhoudelijke relatie hebben die de zoekvraag impliceerde: het probleem van *false coördination* (zie paragraaf 1.7).
- In lange teksten komen ook veel woorden voor die weinig te maken hebben met het werkelijke onderwerp van het document, maar waarop het document wel wordt gevonden.

Veel van die problemen spelen veel minder als je bij het zoeken gebruik kunt maken van gecontroleerd vocabulaire. Als voornaamste doel heeft dat immers om:

- terminologie eenduidig te maken en te standaardiseren, zodat geen problemen meer optreden met woordvormen; spellingsverschillen; taalverschillen; synoniemen of homoniemen/homografen; door browsen of navigeren door hiërarchieën de juiste term of categorie te helpen vinden;
- te kunnen verwijzen naar een officieel te gebruiken term (vervangend zoeken);
- zoekingen bij benadering gelijk gewicht te geven, zodat documenten alleen nog op daarvoor relevante termen worden gevonden en, bij diepte-indexering (zie paragraaf 7.2), in principe ook op alle relevante termen;
- relaties te leggen tussen begrippen, zodat zoekresultaten makkelijk (soms automatisch) kunnen worden uitgebreid met publicaties over specifiekere of deelonderwerpen (zoals bij generiek zoeken; zie paragraaf 4.5) of over verwante onderwerpen;
- (soms) precoördinatief syntactische relaties te leggen tussen begrippen die in een document aan de orde komen (zie het voorbeeld van de subheadings in de MeSH-thesaurus in paragraaf 13.2).

automatische technieken

Gezien het gebrek aan flexibiliteit van gecontroleerde ontsluiting en vooral ook gezien de kosten die verbonden zijn aan het handmatig ontsluiten van informatie is het goed te kijken in hoeverre in nieuwe generaties free-text retrievalsoftware automatische technieken worden toegepast die op zijn minst een gedeeltelijke oplossing bieden voor de hier genoemde problemen. In de volgende paragrafen komen daartoe de volgende technieken aan de orde:

- zoeken op woordstammen;
- fuzzy zoeken;
- suggereren van spellingscorrecties;
- splitsen van samengestelde woorden;
- gebruik van semantische kennis;
- relevantie ordening;
- clusteren/uitsplitsen van zoekresultaten;
- genereren van suggesties voor vraagverfijning;
- terugkoppelingstechnieken.

Daarnaast zullen we kijken naar hetgeen de laatste tijd onder de noemer “semantisch zoeken” ontwikkeld is.

11.2 Woordstammen, fuzzy zoeken, spellingscorrecties en splitsen van samengestelde woorden

trunceren maskeren

Al lang bestaat in veel zoeksystemen de mogelijkheid van trunceren of maskeren: het zoeken naar alle woorden die beginnen met een bepaald door

de gebruiker ingetikt woorddeel ('computer*' levert ook 'computeronderwijs') of waar op een bepaalde plaats in het woord een variabele letter mag voorkomen ('publi?atie' levert zowel 'publicatie' als 'publikatie').

word-stemming

Trunceren biedt een tamelijk ruwe manier om op variaties van woorduitgangen te zoeken. Bovendien moeten gebruikers zelf bedenken hoe dat te doen. Daarom wordt in zoeksystemen steeds vaker *word-stemming* toegepast. Ook Google is daar een voorbeeld van. Door bij het indexeren van tekst de daarin voorkomende woorden te reduceren tot hun morfologische woordstam (bijvoorbeeld: computer → comput) en dat bij zoekwoorden ook te doen, maakt het niet meer uit welke woordvormen toevallig in de te doorzoeken teksten voorkomen (enkelvoud, meervoud, verbuiging, vervoeging, enzovoort: computers, computing, computation, computed) en evenmin op welke woordvorm iemand zoekt. Als zodanig heeft het een positieve invloed op de vangst van zoekacties. De meest toegepaste technieken voor *word-stemming* zijn gebaseerd op regels die in principe voor (bijna) alle woorden in een taal zouden moeten gelden. Deze methode is sterk taalafhankelijk en niet voor alle talen even eenvoudig mogelijk. Zo zal bij (regelmatige) werkwoorden in het Engels alleen aan het eind van het woord iets variëren (talk - talks - talked - talking), maar in het Nederlands ook aan het begin en zelfs vaak in aanwezige klinkers (schrijven, schrijft, schreef, geschreven). Probleem met deze methode is dat er bovendien altijd uitzonderingen kunnen voorkomen, waarbij de algemene stemmingregels niet tot correcte resultaten leiden (bijvoorbeeld: communism - community - communication). Om *word-stemming* effectief te laten zijn, moet gezorgd worden dat noch *under-stemming*, noch *over-stemming* optreedt. Moet van het woord 'hypothetical' alleen de uitgang '.al', wellicht beter '.ical' of misschien zelfs '.etical' worden afgehaakt om de juiste woordstam (hypothetic / hypothet / hypoth) over te houden? En is dat bij het woord 'chemical' net zo? *Over-stemming*, het verwijderen van te lange woorduitgangen, zal over het algemeen nogal nadelig zijn voor de precisie van zoekresultaten; *under-stemming* juist weer voor de vangst. *Stemming* geeft meer verbetering van de vangst, naarmate de hoeveelheid digitaal doorzoekbare tekst van de documenten kleiner is. In lange documenten zullen de verschillende vormen van belangrijke woorden toch vaak al allemaal aanwezig zijn.

fuzzy zoeken

Een ander soort woordvarianten, met name die in spelling, kan automatisch worden meegenomen door technieken van *fuzzy* zoeken. Er bestaan verschillende technieken om de computer te laten zoeken naar woorden die wat betreft spelling of uitspraak sterk lijken op de woorden in de zoekvraag. Deze techniek kan compenserend optreden voor spelfouten in zowel de documenten als in de zoekvragen, alsmede voor spellingsvarianties en in beperkte mate voor morfologische varianten. Dit geldt ook voor OCR-fouten in gescande teksten. Hierbij hoeft de gebruiker dus niet meer te maskeren. Soms kan bij het zoeken de mate van

fuzziness worden ingesteld (ruwweg het aantal letters waarin gevonden en gevraagde woorden maximaal mogen verschillen). Bij deze techniek gaat de verbetering van de vangst overigens vaak ten koste van de precisie. Eigenlijk alleen bij vrij lange zoekwoorden zal de precisie van zo verkregen zoekresultaten voldoende goed blijven. Korte woorden waarvan één letter verschilt, hebben meestal al een heel andere betekenis (bijvoorbeeld: boek → boer – bonk – koek)

Na het intikken van een zoekvraag kunnen zoeksystemen ook met suggesties komen voor een zoekterm in iets afwijkende spelling, waarvan het systeem verwacht dat die een beter resultaat oplevert. Dat kan gebeuren in gevallen dat de oorspronkelijke term erg weinig oplevert. De suggestie voor de betere spelling zal veelal gegenereerd worden door een *fuzzy search* in de eigen index, waarbij de aantallen treffers voor potentiële afwijkende spellingen een belangrijke rol spelen. Generieke kennis van veel gemaakte tikfouten (bijvoorbeeld verwisseling van twee letters) kan daarbij ook meespelen. Een zoekmachine als Google vervangt meestal al automatisch een, in zijn ogen, fout gespelde term (“zoekresultaten voor ...”); in andere gevallen wordt alleen gesuggereerd dat wellicht beter het aangereikte alternatief kan worden gebruikt.

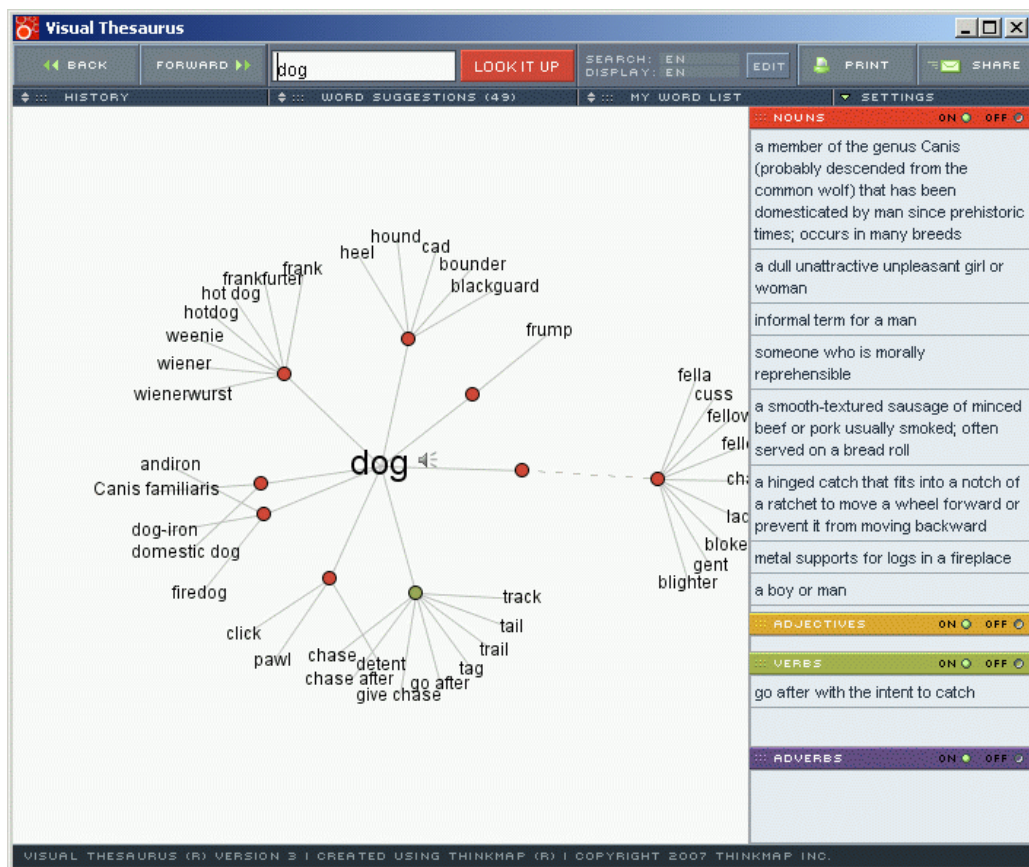
decompounding
splitsen

In het Engels komen haast geen samengestelde woorden voor. In talen als Nederlands, Duits en Deens echter wel. Daardoor zul je een document over bijvoorbeeld een ‘hockeytoernooi’ niet automatisch vinden op de zoekterm ‘toernooi’. Gewone truncatie, *word-stemming* en *fuzzy* zoeken zijn hiervoor geen remedie. Om ook in dit soort situaties de vangst te verbeteren zullen dergelijke samengestelde woorden bij het indexeren automatisch in hun afzonderlijke bestanddelen opgesplitst moeten worden, zogenaamde *decompounding*. Daarbij ligt het vaak heel subtiel voor welke woorden het zinlijk is om ze wel of juist niet te splitsen. Zo zullen de woorden ‘voorjaar’ of ‘voorhoede’ zeker niet gesplitst moeten worden, maar ‘vooraankondiging’ of ‘jubileumjaar’ weer wel. Ook het punt waar het systeem een woord moet splitsen is niet altijd makkelijk eenduidig te bepalen. In kranten gebruikte automatische afbreek-routines geven hier soms hilarische voorbeelden van. Denk bijvoorbeeld aan ‘kerst-omaatjes’ of ‘park-eerst-rook’. Google past deze methode (zeker bij het zoeken met Nederlandse woorden) ook al vaak toe. Als in je zoekvraag bijvoorbeeld het woord ‘tunnels’ voorkomt, blijkt vaak ook op ‘autotunnels’ en ‘wegtunnels’ gezocht te zijn.

11.3 Gebruik van semantische kennis

De technieken in de vorige paragraaf waren vooral gebaseerd op tamelijk mechanische manipulatie met letters en onderdelen binnen woorden, waarbij de betekenis van woorden nog geen rol speelde. Door kennis over de betekenis van woorden toe te voegen, kan zoeken ook verder worden

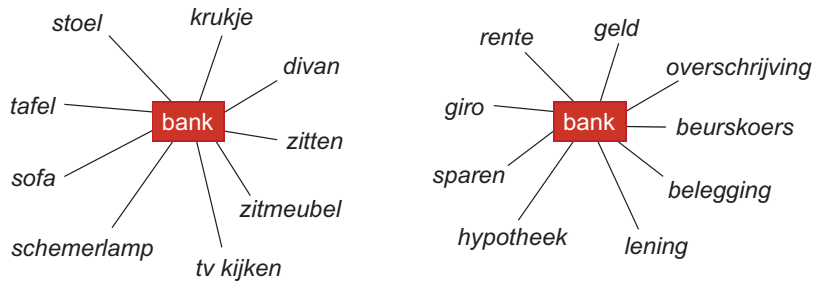
verbeterd. Dat kan onder meer door gebruik te maken van een semantisch netwerk (zie paragraaf 9.1). Enerzijds zijn in zo'n semantisch netwerk verschillende betekenissen van woorden onderscheiden. Anderzijds legt het relaties tussen woorden die op enigerlei wijze inhoudelijk gerelateerd zijn, en dat afzonderlijk voor elke betekenis van een woord. Afbeelding 11.1 toont een visualisatie van een stukje van zo'n semantisch netwerk voor de Engelse taal.



Afb. 11.1 - Visualisatie van de semantische omgeving van een woord in het Engelse Wordnet (gemaakt met VisualThesaurus software).

Door de in de omgeving van een woord voorkomende tekst in een document te vergelijken met de omgeving van dat woord in het semantisch netwerk, kan software afleiden in welke betekenis dat woord in de tekst voorkomt. Als in de buurt van het woord 'bank' woorden voorkomen als 'geld', 'beurs' of 'pinnen', kan een computer in principe afleiden dat het op die plaats in de tekst zeer waarschijnlijk om een financiële instelling zal gaan, omdat deze woorden in het semantisch netwerk op enigerlei wijze aan die betekenis van bank gerelateerd zijn.

Als er daarentegen sprake is van ‘kussens’, ‘zitten’, ‘tv-kijken’, ‘tafeltje’ of ‘schemerlamp’, dan zal het waarschijnlijk om een zitmeubel gaan. Afbeelding 11.2 geeft hier een schematische illustratie van. Als een computerprogramma op deze wijze de betekenis van multi-interpretabele woorden kan afleiden, kan dus specifiek op een gewenste betekenis worden gezocht, hetgeen goed is voor de precisie.



*“hij zat op de **bank** tv te kijken,
met een biertje voor zich op tafel”*

*“deze **bank** rekent een hoge rente
op leningen en hypotheek”*

Afb. 11.2 - Fictief schematisch voorbeeld van het disambigueren van de betekenis van een woord met behulp van een semantisch netwerk.

Ter verbetering van de vangst kan een zoekactie ook automatisch geëxpandeerd worden met verwante en synonieme begrippen uit het semantisch netwerk. Dat zijn dan de woorden die binnen een bepaalde semantische afstand van de oorspronkelijke zoekterm vallen. In principe kan een zoekstelsel deze termen ook eerst ter keuze aan de zoekster voorleggen.

Voorwaarde voor toepassing van deze techniek is uiteraard de beschikbaarheid van een semantisch netwerk van de taal (of talen) waarin de te doorzoeken documenten gesteld zijn. Voor veel talen bestaan intussen dergelijke geavanceerde digitale woordenboeken. Die bestaan zelfs meertalig, zoals het meertalige semantische netwerk dat op de Van Dale vertaalwoordenboeken is gebaseerd. Ze beperken zich echter meestal tot de gewone dagelijkse taal. Gespecialiseerd vakmatig of wetenschappelijk vocabulaire maakt daar vrijwel nooit standaard al deel van uit. Als dat van belang is, zal men een gespecialiseerde leverancier in de arm moeten nemen om dat toe te voegen.

De *graphs* die in semantische zoeksystemen worden gebruikt (zoals *Google’s knowledge graph*) kunnen nog veel gedetailleerder kennis bevatten over inhoudelijke relaties tussen begrippen, dan een semantisch netwerk (zie paragraaf 9.6.3). Daarmee kan in principe hetzelfde worden bereikt als hierboven beschreven.

11.4 Technieken voor relevantie-ordening

relevantie- ordening

De relevantie-ordening die moderne zoekmachines toepassen, is bedoeld als remedie tegen lage precisie. Dit is er immers op gericht om de resultaten met de hoogste relevantie het eerst te tonen; de precisie van de eerste tien of twintig getoonde resultaten zal dus worden verbeterd. Voor die relevantie-ordening houdt de zoekmachine onder meer rekening met de volgende factoren:

1. Hoe meer van de ingetikte zoekwoorden in een document voorkomen, hoe hoger dit scoort (zie kader "Vectormodel"); bij webzoekmachines speelt deze factor geen rol meer, omdat daar bijna altijd alle gevraagde woorden in de resultaten moeten voorkomen.
2. De plaats waar een gevraagde term in de gevonden documenten voorkomt. Als dat op een belangrijke plek in een document is (in de titel, in opvallende koppen, in de eerste paar regels) zal dit document hoger scoren.
3. Als een gevraagde term in een gevonden document herhaald voorkomt, zal dit document hoger scoren.
4. Als gevraagde woorden in een gevonden document dicht bij elkaar staan, zal dit document hoger scoren.
5. Dat laatste effect zal nog sterker gelden als de gevraagde woorden in een gevonden document ook in dezelfde volgorde voorkomen als in de zoekvraag.
6. Bij de voor deze ordening uitgevoerde berekeningen krijgen zeldzame termen zwaarder gewicht dan algemene, omdat zeldzame begrippen meer onderscheidend zijn en belangrijker, want specifieker, delen van de zoekvraag representeren (zie kader "TF*IDF").
7. Als (in een webomgeving) veel hyperlinks van andere sites naar een gevonden document verwijzen, zal dit document hoger scoren.
8. Als een gevonden document of website veel wordt bezocht, dan zal dit document hoger scoren.
9. Als een document lijkt op documenten die de gebruiker eerder geopend heeft, dan zal dit document hoger scoren (zie ook paragraaf 11.7).

Hierbij proberen de factoren 1, 2 en 3 onderscheid te maken tussen woorden die in een document inhoudelijk centraal staan en die welke daarin min of meer toevallig voorkomen. De factoren 4 en 5 zorgen dat die documenten eerder worden getoond, waarin de zoekwoorden vermoedelijk in de gewenste (syntactische) relatie tot elkaar staan. De factoren 7 en 8 zijn meer een poging om ook de factor kwaliteit (hier wellicht verward met populariteit) mee te laten spelen. Factor 9 zorgt voor personalisatie van de relevantiebepaling.

Webzoekmachines houden meestal ook nog rekening met tientallen andere verfijnde factoren bij het berekenen van relevantiescores van gevonden documenten. Veel van die factoren zijn echter vooral gericht

op het neutraliseren van ongewenste technieken van search engine optimalisatie (SEO), waarmee websitebouwers op kunstmatige wijze proberen pagina's hoog te laten scoren in de relevantieberekeningen.

11.5 Clusteren van zoekresultaten

clusters

Een manier om de precisie van een al verkregen zoekresultaat te verbeteren is automatische opdeling van dat resultaat in clusters. Die clusters kunnen gebaseerd zijn op tevoren gedefinieerde categorieën of tot stand komen op basis van statistische analyse van de tekst van de gevonden documenten. De gebruiker kan vervolgens die cluster(s) kiezen die het best aansluit(en) bij zijn/haar informatiebehoefte. Wie bijvoorbeeld op BSE zoekt, kan dan de keuze krijgen tussen clusters als 'mad cow disease' (bovine spongiform encephalopathy), 'stock exchange' (Bombay / Beirut / Boston Stock Exchange) en 'cancer detection' (breast self examination).

categorieën

Voor het werken met tevoren (door mensen) gedefinieerde categorieën dienen kennisregels beschikbaar te zijn, op basis waarvan de computer de documenten kan indelen. Die kennisregels kunnen handmatig zijn opgesteld of uit voorbeelddocumenten zijn afgeleid, op de wijze zoals dat ook bij het automatisch categoriseren van documenten gebeurt (zie hoofdstuk 10). Bij het fulltext indexeren van de documenten, worden die dan ook meteen volgens deze regels gecategoriseerd. Het resultaat van een zoekactie kan vervolgens gemakkelijk worden uitgesplitst op basis van de aan de gevonden documenten toegekende categorieën. De benaming van die categorieën is tevoren al door hun bedenker vastgelegd.

Het is ook mogelijk een zoekresultaat op te delen, zonder dat vooraf categorieën en hun karakteristieke kenmerken zijn bedacht. Daartoe kan een computerprogramma bijvoorbeeld gebruik maken van vectorrepresentaties van de documenten in het zoekresultaat (zie kader "Vectormodel"). Groepen documenten waarvan de vectoren ongeveer in dezelfde richting wijzen, vormen dan een cluster. Ook is een statistische analyse mogelijk van de woorden in de tekst van de gevonden documenten, waarbij het samen voorkomen van groepen karakteristieke woorden een maat is voor de gelijkenis tussen documenten. Deze methoden kunnen alleen worden toegepast op niet te grote zoekresultaten, dus bijvoorbeeld niet op 3 miljoen resultaten uit een webzoekmachine, maar wel op de enkele honderden resultaten die een metazoekmachine uit achterliggende echte zoekmachines heeft opgehaald (zie afbeelding 11.3). De benaming van dergelijke clusters wordt door de computer vastgesteld op grond van een meest karakteristiek woord of karakteristieke zinsnede uit zo'n cluster. Daardoor geven die benamingen niet altijd even begrijpelijke omschrijvingen van het onderwerp.



Afb. 11.3 - Een in clusters opgedeeld zoekresultaat van de metazoeckmachine Carrot2 (<http://search.carrot2.org/>). Behalve als alleen tekst, kunnen de clusters ook in “zeepbel”-visualisatie getoond worden.

11.6 Gegengereerde suggesties voor verfijnen van zoekresultaat

verfijnen

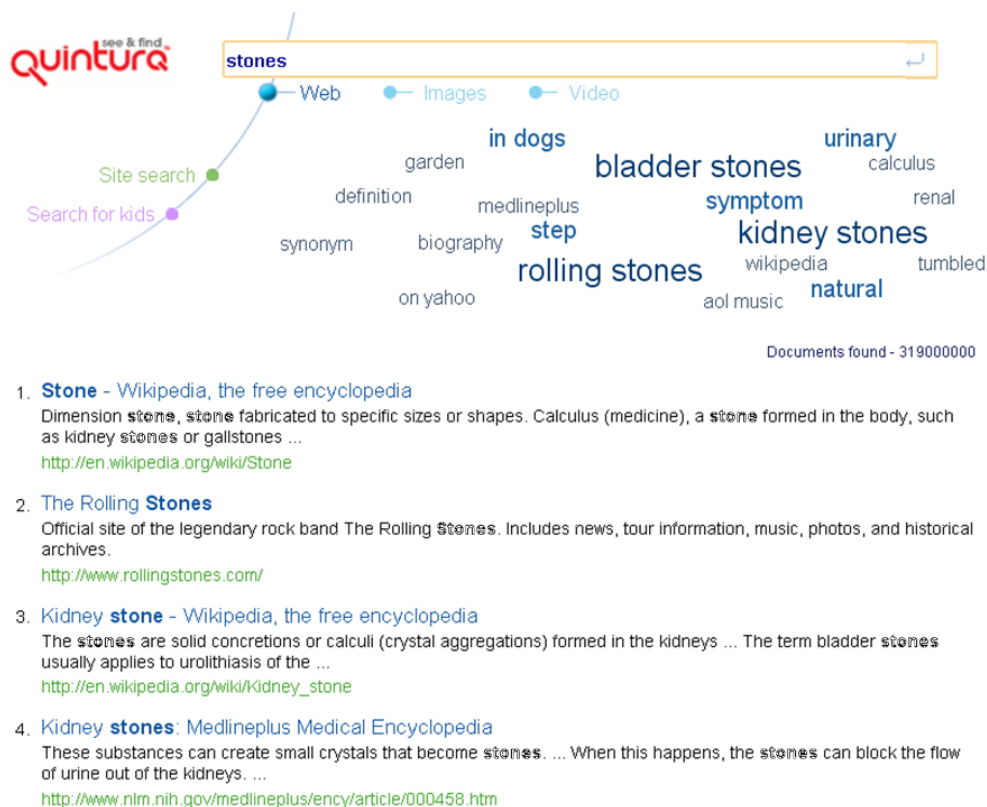
Een statistische analyse van de woorden in een zoekresultaat kan ook aanleiding geven tot een lijst karakteristieke losse woorden of zinsneden die daarin vaker voorkomen, zonder dat de gevonden documenten op basis daarvan al worden geclusterd. Voor het vaststellen welke woorden karakteristiek (of specifiek) genoeg zijn, wordt vaak de TF*IDF techniek toegepast (zie kader “TF*IDF”). Daarmee wordt voorkomen dat woorden in de keuzelijstjes terechtkomen die wel veel in het zoekresultaat voorkomen, maar zo algemeen zijn dat ze in bijna elk document veel voorkomen. De gesuggereerde termen voor verfijning van het zoekresultaat worden dan gepresenteerd in volgorde van hun TF*IDF-waarde.

De zoeker kan dan woorden of zinsneden uit het getoonde lijstje aanklikken om ze (in een AND-relatie) aan de oorspronkelijke zoekvraag toe te voegen en zo het zoekresultaat verder in te perken. De bij de Openbare Bibliotheken in Nederland en Vlaanderen gebruikte Aquabrowser, visualiseert de zo gegenereerde woorden, samen met woorden uit semantische hulpsystemen, in een soort woordenwolk (zie afbeelding 11.4).

The screenshot shows the Aquabrowser search interface. On the left is a word cloud with terms like 'presidentsverkiezingen', 'opspraak', 'presidentskandidaat', 'nominatie', 'bush', 'heksenjacht', 'hampshire', 'barack', 'bill', 'york', 'obama', 'hilarie', 'hilaria', 'hillary', 'clinton', 'zaak', 'edmund', 'hilarly', 'hilarii', 'vlucht', 'merkel', 'voorverkiezing', 'iowa', 'wankelen', 'lady', and 'buitenlandminister'. A legend at the bottom left indicates: Associate (black), Vertaling (green), Ontdekkingspad (blue), Spellingsvariant (orange), and Thesaurusterm (grey). The main search results area on the right shows the search term 'hillary' and lists several books and articles. The first result is 'Hillary' by Gail Sheehy, described as 'prijzvechter, martelaar en opportunist'. The second is 'Bill & Hillary' by Christopher Andersen, described as 'hun huwelijk'. The third is 'HILLARY FOR PRESIDENT' by an unnamed author, described as 'Ik doe mee, en ik doe mee om te winnen'. The right sidebar contains various filters and lists, including 'Verfijnen' (Refine) with tags like 'hoge berg beklimmen (1)', 'hillary (1)', 'clinton (1)', 'biografie (1)', 'de beste boeken (nrc) (1)', and 'boek (1)'. It also lists 'Alle bronnen' (All sources) and 'Bibliotheekmateriaal' (Library material).

Afb. 11.4 - Zoekresultaat in een Aquabrowser. Links een woordenwolk met zowel statistisch gegenereerde woorden, als verwante woorden uit een thesaurus of een semantisch netwerk, alsmede fuzzy gelijkende woorden. Rechts is het zoekresultaat opgedeeld op basis van formele aspecten.

Deze statistische methode is ook toepasbaar op zeer grote zoekresultaten door de analyse alleen te laten plaatsvinden op een behapbaar deel van het zoekresultaat onder de aanname dat die wel representatief zal zijn voor het hele zoekresultaat. De zoekmachine Quintura is hier een voorbeeld van (zie afbeelding 11.5).



Afb. 11.5 - Woordenwolk van statistisch gegenereerde woorden en begrippen uit een eerder zoekresultaat in de Quintura zoekmachine; hoe hoger $TF \cdot IDF$ -waarde, hoe groter de letters.

noun phrases

Hoe de computer kan bepalen wat zinnige zinsneden zijn om in de statistische analyse te betrekken, is een belangrijk veld van onderzoek. Niet elk twee- of drietal woorden uit een zin komt daar namelijk voor in aanmerking. Naast vooral in het Engels veelvuldig voorkomende *noun phrases* (opeenvolgingen van twee zelfstandig naamwoorden, zoals 'information retrieval'), kunnen dat ook complexere stukken zin zijn met voorzetsels, bijvoeglijke naamwoorden en dergelijke (zoals 'grote rode auto' of 'opwarming van de aarde'), die als geheel een concept of begrip kunnen representeren. Om deze zogenaamde *lexical phrases* te kunnen onderscheiden van willekeurige reeksen opvolgende woorden, is tamelijk geavanceerde taalkundige analyse van de teksten nodig.

lexical phrases

faceted search

Een andere wijze van verfijning die je in steeds meer zoeksystemen tegenkomt, is gebaseerd op gestandaardiseerde metadata die toch al in het doorzochte materiaal aanwezig zijn. Bij deze zogenaamde *parametric* of *faceted search* kan een zoekresultaat worden uitgesplitst op basis van bijvoorbeeld de aard van gevonden documenten (webpagina, pdf, word-document, boek, krantenartikel), van de taal van de gevonden publicaties,

van de namen van daarin frequent voorkomende auteurs, van de titels van de tijdschriften waaruit gevonden artikelen afkomstig zijn, van globale onderwerpscategorieën en dergelijke. Die uitsplitsing vermeldt vaak meteen al hoeveel na inperking overblijft van het oorspronkelijke resultaat. In het voorbeeld van de Aquabrowser in afbeelding 11.4 is rechts een opdeling in diverse facetten zichtbaar.

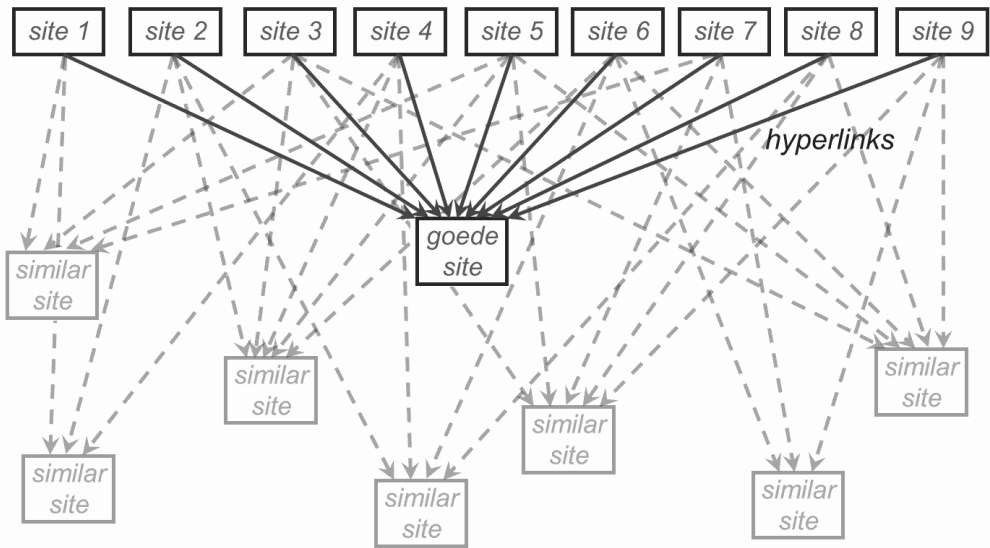
11.7 Terugkoppeling door gebruikers

terugkoppeling

more like this

Behalve de in de twee vorige paragrafen besproken suggesties voor inperkingen waaruit een gebruiker kan kiezen, zijn er ook nog andere manieren waarop terugkoppeling van het oordeel van de gebruiker tot verbetering van de precisie van zoekresultaten kan leiden. Een eerste methode daarvoor is de *more like this* functie. Daarbij kan de gebruiker aangeven dat op basis van een al gevonden en als zeer relevant beoordeeld document, verder gezocht moet worden naar documenten die daar sterk op lijken. Die gelijkenis zal veelal worden bepaald op basis van de meest karakteristieke woorden die in het voorbeelddocument voorkomen. De vectormethode is een hiervoor vaak gebruikte techniek (zie kader “Vectormodel”). Soms wordt de gelijkenis ook op heel andere wijze bepaald, bijvoorbeeld door Google die daarvoor linkpatronen tussen webpagina’s gebruikt. Afbeelding 11.6 geeft een schematische indruk hoe zoiets ongeveer werkt. Hoewel je zou verwachten dat het resultaat van deze zoekmethoden altijd tot goede precisie leidt, blijkt dat in de praktijk toch nog wel eens tegen te vallen.

Een andere wijze van terugkoppeling is die waarbij de gebruiker actief kan aanvinken welke van de zoekresultaten hij als relevant beoordeelt. Zoeksystemen kunnen echter ook zo geprogrammeerd worden dat ze zelfstandig deduceren welke documenten de gebruiker kennelijk relevant vindt. Ze registreren dan welke resultaten de gebruiker daadwerkelijk opvraagt en hoe lang hij daarnaar kijkt. Op basis daarvan zullen de meest karakteristieke woorden uit die documenten een hoger gewicht krijgen ten behoeve van toekomstige zoekacties. Documenten die deze woorden bevatten, zullen in de relevantieordening vervolgens hoger scoren dan tevoren. Op deze manier ontstaat in feite een zelflerend systeem dat geleidelijk, in de praktijk heel subtiel, steeds iets betere precisie moet opleveren, specifiek toegesneden op de persoonlijke belangstelling en voorkeuren van de gebruiker en daarmee ook mee-evoluerend. In paragraaf 11.4 werd deze methode al als één van de factoren voor relevantieordening genoemd.



Afb. 11.6 - Schematische weergave van de op linkpatronen gebaseerde methode van Google om bij een gevonden webpagina, naar “vergelijkbare” pagina’s te zoeken.

11.8 Semantisch zoeken

semantisch zoeken

De technieken in de vorige paragrafen hebben we steeds besproken vanuit het gezichtspunt van het vinden van relevante documenten. Intussen wordt er steeds vaker gesproken over semantisch zoeken. Wat simplistisch gesteld betreft dat een verzameling technieken die niet meer gericht is op het vinden van documenten, maar op het geven van antwoorden, omdat de zoekvraag begrepen wordt. Vooral bij de grote webzoekmachines is deze ontwikkeling heel prominent. Daar horen uitspraken bij als:

- “Liever een antwoord dan 10 blauwe links”: refererend aan de blauwe titels van de klassieke zoekresultaten bij de meeste webzoekmachines.
- “No strings, but things”: het gaat niet om de toevallige reeksen letters (*strings*) van woorden, maar om de betekenis ervan.

Bij het realiseren van de functionaliteit voor semantisch zoeken speelt een groot aantal van de in hoofdstuk 9 geschetste methoden en ontwikkelingen een belangrijke rol.

Een eerste essentiële taak voor een semantisch zoekstelsel is dat het moet begrijpen wat de gebruiker met een vraag bedoelt en wat diens intentie daarbij is. Daarbij spelen allerlei factoren een rol. Zo kan de locatie van de gebruiker een belangrijke vingerwijzing zijn voor de achtergronden van een vraag. En die locatie is steeds vaker al bij het stelsel bekend. Iets

soortgelijks geldt voor het gebruikte apparaat, bijvoorbeeld of dat een desktop of een mobieltje is. Verder kan herkend worden of in de zoekvraag bepaalde entiteiten, personen, plaatsen, gebeurtenissen, voorkomen op dezelfde manier als dat in documenten gebeurt (zie paragraaf 10.1.1). Daarnaast worden voor vraaginterpretatie ook natuurlijke taaltechnieken toegepast. Veel daarvan is gebaseerd op *Machine Learning* methoden (zie kader). Het door Google ontwikkelde Rankbrain algoritme is daar een voorbeeld van (Enge, 2016; Harry, 2016).

Een tweede stap is het beschikbaar hebben (en het vinden) van antwoorden op de gestelde vragen. Daarvoor worden diverse soorten kennissystemen gebruikt. Bij Google zijn dat diens *knowledge graph* en/of de *knowledge vault* (zie paragraaf 9.6.3). Voor dit soort toepassingen van semantisch zoeken worden dus geen documenten inhoudelijk toegankelijk gemaakt en er wordt ook niet naar documenten gezocht, maar rechtstreeks naar antwoorden. Begrippen als vangst en precisie in de eerder gebruikte betekenis, spelen hierbij dus geen rol. De voor de antwoorden benodigde kennis moet wel vooraf zijn vastgelegd. Zoals we in hoofdstuk 9 beschreven, wordt die kennis deels wel weer uit documenten gehaald.

Overigens is ook semantisch zoeken een containerbegrip waar diverse technieken en toepassingen onder vallen, Bij sommige van die technieken wordt ook wel degelijk nog naar documenten gezocht. Maar in de daarbij gevonden documenten worden dan bijvoorbeeld gericht die zinsneden getoond waarin het antwoord op de gestelde vraag gegeven wordt. Of zo'n antwoord wordt uit fragmenten van meer gevonden documenten samengesteld. In dergelijke systemen moet bij de eerste zoekstap dus wel naar hoge precisie gestreefd worden om de beste documenten te vinden waar de antwoorden vervolgens dynamisch uit gecreëerd worden. Een paar voorbeelden van dergelijk zoeksystemen zijn: Sensebot, Factbites, Cluuz en Kngine.

11.9 Samenvattend

Van de in de paragrafen 11.2 t/m 11.7 behandelde technieken, zijn de meeste gericht op verbetering van de precisie. Dat kan enerzijds gebeuren door technieken voor relevantieordening die in elk geval beogen de precisie van de eerst getoonde resultaten te verhogen. Anderzijds kan dat gebeuren door een iteratief proces van verfijning van de zoekresultaten, door die resultaten te clusteren, door suggesties te doen voor verfijning op basis van aanvullende zoekwoorden of door de gebruiker anderszins te laten aangeven welke resultaten het best aan zijn informatiebehoefte voldoen. Een volledig alternatief voor de precisieverbetering die optimaal toegepaste gecontroleerde ontsluiting kan bieden, vormen deze technieken echter nog niet.

Van de technieken die gehanteerd worden voor verbetering van de vangst zijn de meeste sterk taalafhankelijk, zodat toepasbaarheid daarvan afhankelijk is van de taal (talen) waarin de te doorzoeken documenten gesteld zijn. Helaas kunnen sommige van die technieken makkelijk aanleiding geven tot aanzienlijke verslechtering van de precisie van zoekresultaten. Bij een juiste toepassing van gecontroleerde ontsluiting hoeft verbetering van de vangst, bijvoorbeeld door generiek zoeken, veel minder ten koste te gaan van de precisie.

Vergelijkend onderzoek naar de prestaties van gecontroleerd vocabulaire en *freetext retrieval* onder toepassing van de meest geavanceerde natuurlijke taaltechnieken, met gebruikmaking van dezelfde bestanden of corpora, is helaas tamelijk zeldzaam. Bovendien komen daar tot op heden nog weinig eensluidende conclusies uit.

De in paragraaf 11.8 besproken semantische zoektechnieken zijn vooral gericht op het vinden van zo concreet mogelijke antwoorden. Daarin spelen retrievaltechnieken, gecontroleerde methoden voor toegankelijk maken en begrippen als vangst en precisie een veel beperktere rol.

Gebruikte literatuur

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299 Geraadpleegd op: 24 juni 2016, van:

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.145.4506&rep=rep1&type=pdf>

Engel, Eric (2016). RankBrain: A Study to Measure Its Impact. *Stone Temple - The Digital Marketing Excellence Blog*, 9 March 2016, Geraadpleegd op: 30 juni 2016, van:

<https://www.stonetemple.com/rankbrain-a-study-to-measure-its-impact/>

Harry, David (2016). Getting Your Head Around Google's RankBrain. *The Sem Post*, March 28, 2016. Geraadpleegd op: 30 juni 2016, van:

<http://www.thesempost.com/getting-head-around-googles-rankbrain/>

Sieverts, E. (2009). De mythe van de catalogus. *InformatieProfessional - Bijdragen* (29 juni 2009) <http://informatieprofessional.nl/bijdragen/2009/06/de-mythe-van-de-catalogus/>

Verder lezen

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? *Information Retrieval*, 7(3), 291–316

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing & Management*, 41(4), 873-890

Sieverts, E. (2007). Inhoudelijk toegankelijk maken van informatiecollecties in een digitale omgeving. In *Handboek Informatiewetenschappen* (aanvulling 42, p. 1-42). Alphen a/d Rijn: Kluwer

KADER

Vectormodel

Retrievalsystemen berusten bijna altijd op een *inverted file model*, wat er kort gezegd op neer komt dat de zoeksoftware een index genereert met alle woorden uit de te doorzoeken documenten. Het vectormodel is een variant daarop. Een index van alle in een systeem aanwezige termen vormt daarbij ook het uitgangspunt. Elk document wordt vervolgens in een vectorindex gerepresenteerd door een reeks 1-en en 0-en die aangeven welke indextermen wel (1) en welke niet (0) in dat document voorkomen. Als eenvoudig voorbeeld nemen we een zeer beperkt informatiesysteem waarin de volgende tien (zeer korte en soms wellicht wat onzinnige) documenten aanwezig zijn:

1. Voor de **toegankelijkheid** van deze **informatie gebruiken** we een **thesaurus**.
2. Voor **browsen gebruiken** we een **classificatie**.
3. **Zoeksystemen** worden **beter** als we een **thesaurus gebruiken** voor **toegankelijkheid**.
4. In **zoeksystemen** wordt **vaak postcoördinatie toegepast**.
5. In een **classificatie** worden **aspecten** van een **document precoördinatief gekoppeld**.
6. Als een **thesaurus** wordt **toegepast** zijn **aspecten** van de **inhoud** van een **document** niet **precoördinatief gekoppeld**.
7. **Zoeksystemen** die voor **toegankelijkheid** een **thesaurus gebruiken** worden **vaak toegepast**.
8. **Postcoördinatie** van **zoeksystemen** maakt dat je niet kunt **browsen**.
9. De **inhoud** van **informatie** in een **document** is **gekoppeld** aan de **postcoördinatie** van een **thesaurus**.
10. Hoe **beter** de **toegankelijkheid** van **informatie**, hoe **beter** we de **inhoud gebruiken**.

Bij het indexeren worden stopwoorden zoals lidwoorden, voorzetsels, hulpwerkwoorden en dergelijke overgeslagen. De index van dit kleine zoekstelseltje bevat dan de zestien termen in de linker kolom van onderstaande tabel. De kolommen daarnaast, genummerd 1 t/m 10, representeren de tien documenten in het systeem. In elke kolom in deze tabel is te zien welke woorden in een bepaald document voorkomen (in document 2 bijvoorbeeld de woorden 'browsen', 'classificatie' en 'gebruiken'), in elke rij zien we in welke documenten een bepaalde term voorkomt (bijvoorbeeld de term 'informatie' in de documenten 1, 9 en 10).

De kolommen met 0-en en 1-en in deze tabel kun je wiskundig als vectoren beschouwen, dat wil zeggen als “pijlen” die in een bepaalde richting wijzen. In dit geval zijn dat vectoren in een 16-dimensionale ruimte. Hoewel we ons visueel geen hogere dan een 3-dimensionale ruimte kunnen voorstellen, kan wiskundig ook worden gerekend met vectoren in zulke hoger dimensionale ruimtes. En dat kan ook als er geen 16 maar 16.000 indextermen zijn, dus ook in een 16.000- (of nog veel hoger) dimensionale ruimte.

Documentindex	1	2	3	4	5	6	7	8	9	10	Zoekvraag
aspecten	0	0	0	0	1	1	0	0	0	0	0
beter	0	0	1	0	0	0	0	0	0	1	1
browsen	0	1	0	0	0	0	0	1	0	0	0
classificatie	0	1	0	0	1	0	0	0	0	0	0
document	0	0	0	0	1	1	0	0	1	0	0
gebruiken	1	1	1	0	0	0	1	0	0	1	0
gekoppeld	0	0	0	0	1	1	0	0	1	0	0
informatie	1	0	0	0	0	0	0	0	1	1	0
inhoud	0	0	0	0	0	1	0	0	1	1	0
postcoördinatie	0	0	0	1	0	0	0	1	1	0	0
precoördinatief	0	0	0	0	1	1	0	0	0	0	0
thesaurus	1	0	1	0	0	1	1	0	1	0	1
toegankelijkheid	1	0	1	0	0	0	1	0	0	1	1
toegepast	0	0	0	1	0	1	1	0	0	0	0
vaak	0	0	0	1	0	0	1	0	0	0	0
zoeksysteem	0	0	1	1	0	0	1	1	0	0	0

In een zoekstelsel worden ook zoekvragen als vectoren voorgesteld, met daarin een 1 voor elke term die we vragen (en 0-en voor alle termen die we niet vragen). In de rechterkolom van de tabel is de zoekvraag ‘beter toegankelijkheid thesaurus’ zo als vector gerepresenteerd. Het zoekstelsel voert dan een wiskundige vectorvermenigvuldiging uit tussen de vraagvector en elk van de documentvectoren. Dat levert een getal op, dat een maat is voor de overeenkomst tussen de zoekvraag en elk van die documenten. In dit geval levert die vermenigvuldiging het grootste getal op voor document 3 waarin inderdaad alle drie zoekwoorden voorkomen. Dat is dus waarschijnlijk het meest relevante antwoord op onze vraag. De documenten 1, 7 en 10 waarin twee van de drie zoekwoorden voorkomen scoren wat lager (en zullen dus minder relevant zijn), die met één zoekwoord scoren nog lager, die met geen enkel zoekwoord leveren 0 op.

In de praktijk past men hierop nog allerlei verfijningen toe:

- Voor het samenstellen van de index worden woorden tot hun woordstam gereduceerd, zodat enkel- en meervoud en andere varianten geen afzonderlijke indextermen opleveren (en niet tot onnodige extra dimensies leiden).
- Aan de wijze van voorkomen van de indextermen in de documenten worden gewichten gekoppeld, zodat woorden die in een titel of trefwoordveld staan of die herhaald voorkomen, in de vectoren een hogere waarde krijgen. Daarbij worden dezelfde technieken toegepast als we ook in paragraaf 10.3 beschrijven, zoals de $TF*IDF$ -methode (zie kader). De componenten van de vectoren kunnen dan dus heel andere waarden krijgen dan alleen 0 of 1.
- Op soortgelijke wijze kunnen ook woorden in een zoekvraag een gewicht krijgen op basis van hun (vermoedelijke) belang voor die vraag.
- Voor het vermenigvuldigen worden de lengtes van de vectoren genormeerd, waardoor lange documenten (met heel veel woorden) relatief lager zullen scoren.

De uitkomsten van de vectorvermenigvuldigingen geven zo een veel genuanceerder beeld welke documenten vermoedelijk het beste antwoord op de zoekvraag geven. Wiskundig gezien zijn die uitkomsten een maat voor de hoek tussen de vraagvector en de documentvectoren; hoe groter de uitkomst, hoe kleiner die hoek, dus hoe meer de twee vectoren in dezelfde richting wijzen en hoe groter de overeenkomst met de vraag.

Ditzelfde model kun je ook toepassen om documenten te vergelijken. Hoe kleiner de hoek tussen twee documentvectoren, zoals afgeleid uit hun vermenigvuldiging, hoe meer die documenten op elkaar lijken. Zo lijken de documenten 3 en 7 enigszins op elkaar, omdat ze 4 woorden gemeen hebben. Datzelfde geldt voor de documenten 5 en 6 en voor 6 en 9. De in paragraaf 11.7 genoemde *more like this* zoekfunctie kan hierop berusten. Bovendien kan een systeem zo clusters van op elkaar lijkende documenten herkennen. Bij automatische classificatie kan deze methode ook gebruikt worden om de uit woorden bestaande profielen van te rubriceren documenten te vergelijken met de profielen van thesaurustermen (zie paragraaf 10.2).

KADER

TF*IDF

Met statistische methoden kan een computerprogramma bepalen wat de inhoudelijk belangrijkste en meest karakteristieke woorden uit een tekst zijn. Een veel gebruikte methode gaat uit van relatieve woordfrequenties. Van elk woord uit een tekst wordt daartoe de waarde TF*IDF berekend. Dat wil zeggen dat het aantal keren dat een woord in het document voorkomt, de termfrequentie (TF) wordt vermenigvuldigd met de inverse document frequentie (IDF). Dat is een ingewikkelde manier om te zeggen dat de termfrequentie wordt gedeeld door het aantal documenten waarin dat woord voorkomt. Woorden zijn dus alleen kenmerkend voor een tekst als ze daarin vaak voorkomen, maar tegelijk in weinig (andere) documenten aanwezig zijn. Heel algemene woorden als lidwoorden en voorzetsels met een hoge TF scoren dus toch heel laag, doordat door een heel hoge document frequentie gedeeld moet worden (bijv. $TF=50$; $DF=100.000$; $TF*IDF=0,0005$). Een woord dat maar één keer in een document voorkomt kan toch kenmerkend zijn, als er in totaal maar twee documenten zijn waarin het voorkomt ($TF=1$; $IDF=0,5$; $TF*IDF=0,5$).

Woord	TF	DF	TF*IDF	Rang-nummer	TF*IDF(log)	Rangnummer (log)
in	15	30.000	0,0005	10	1,14	9
de	30	40.000	0,00075	9	0,99	10
vinden	2	2.000	0,001	8	2,21	8
gemist	1	300	0,0033	7	2,52	7
relevant	2	400	0,005	6	3,12	6
zoeken	5	1.000	0,005	5	3,40	5
karakteristiek	2	40	0,05	4	4,42	3
zoekmachine	1	15	0,067	3	3,82	4
recall	1	2	0,5	2	4,70	2
pertinentie	1	1	1,0	1	5,00	1

In de praktijk worden vaak nog verfijningen toegepast. Zo wordt voor zowel TF als IDF vaak een logaritmische functie gebruikt: $TF=1+\log(F)$ en $IDF = \log(N/DF)$, waar N=totaal aantal documenten. Door die wel vaker toegepaste wiskundige truc wordt de schaal voor de TF*IDF-waardes als het ware samengedrukt. In het fictieve voorbeeld hierboven zie je dat daardoor de berekende rangorde van belangrijkheid van de woorden ook ietsje kan wijzigen.

In gevallen waar (nog) geen grote aantallen andere documenten aanwezig zijn waaruit document frequenties bepaald kunnen worden, kan in plaats daarvan ook gebruik gemaakt worden van woordfrequenties uit een ander beschikbaar groot tekstcorpus. Dergelijke getallen kunnen dan eenvoudig aan een index van dat corpus worden ontleend.

Hoofdstuk 12

Hypertext, hyperlinks en relaties

In de voorgaande hoofdstukken kwamen de begrippen linking en hyperlinks al af en toe ter sprake, waarbij er stilzwijgend van werd uitgegaan dat tegenwoordig iedereen wel weet wat daarmee wordt bedoeld. Toch gaan we er hier nog even specifiek op in, omdat hyperlinks en het daaraan ten grondslag liggende concept hypertext in feite technieken zijn waarmee informatie toegankelijk wordt gemaakt.

Omdat iedereen vertrouwd is met de hyperlinks die in webpagina's voorkomen, nemen we die als uitgangspunt voor deze introductie. Een kenmerk van hyperlinks in de eerste generatie webpagina's was dat ze blauw en onderstreept waren. Maar dat is natuurlijk maar een heel oppervlakkige karakterisering van wat een hyperlink is. Al wat wezenlijker voor een hyperlink is dat die je, als je er op klikt, bij andere informatie brengt; op een andere pagina, maar soms ook elders op dezelfde pagina. Het meest wezenlijke is echter dat zo'n hyperlink een relatie tot uiting brengt. Dit is meestal een inhoudelijke relatie die bestaat tussen het vertrekpunt en de bestemming van de link. Hyperlinks en de relaties die eraan ten grondslag liggen, brengen in feite structuur aan in informatiesystemen. En dat geldt zeker ook voor het web: er wordt dan ook vaak gezegd: *"Hyperlinks are the glue that holds the Web together"*. Een meer algemene beschouwing van relaties vormt dus een goed uitgangspunt als we het over linking en hypertext gaan hebben.

12.1 Aard van relaties

relaties

In vrijwel alle informatiesystemen komen relaties voor. Die relaties kunnen van allerlei aard zijn. Het kunnen relaties zijn:

relationele database

1. tussen woorden waarmee we dingen karakteriseren. In dit boek kwamen we daar al allerlei voorbeelden van tegen in het kader van thesauri, semantische netwerken of ontologieën.
2. tussen de "dingen" zelf. Daar kwamen we voorbeelden van tegen bij het RDF formalisme, de *knowledge graphs* en de *linked open data* die in hoofdstuk 9 aan de orde kwamen. En verder zie je iets soortgelijks bij literatuurreferenties, bij relationele databases (het woord zit al in de benaming ervan) en bij de hyperlinks waar dit hoofdstuk vooral over gaat.
3. tussen eigenschappen van objecten, zoals bijvoorbeeld bij ontologieën het geval kan zijn. Zo is in de CIDOC-CRM ontologie de eigenschap "brought into existence" bijvoorbeeld een superproperty van "has

created”, “has formed”, “brought into life”, “has produced” en “resulted in”.

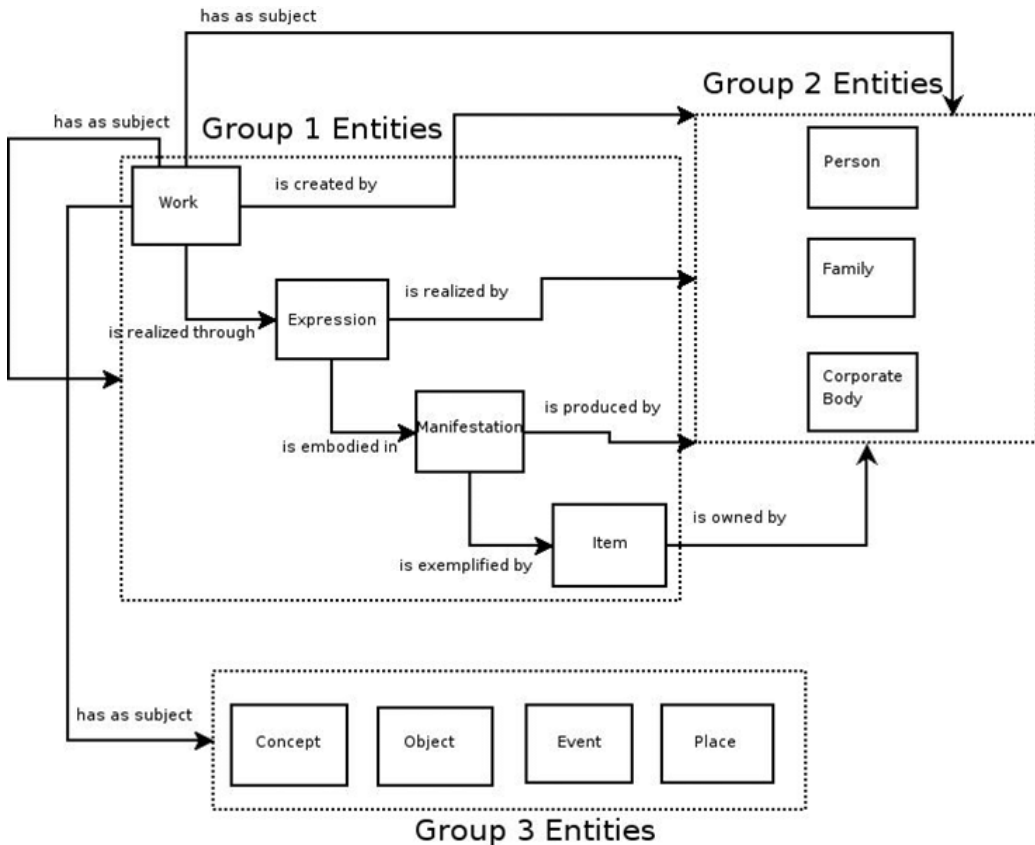
De relaties die we in de praktijk tegenkomen, kunnen ook nog op een andere manier verschillen. Het kan gaan om zeer gestructureerde relaties. Die kom je onder meer tegen in relationele databases. Als je daarmee bijvoorbeeld een uitleensysteem voor een bibliotheek bouwt, zal daarin een vaste relatie gedefinieerd worden tussen een tabel ‘uitleningen’ waarin gegevens van uitleentransacties geregistreerd worden, en een tabel ‘leners’ die gegevens over de gebruikers van de bibliotheek bevat. Voor elke individuele uitleentransactie wordt die relatie ook echt geëffectueerd. Datzelfde geldt voor de relatie tussen de tabel ‘uitleningen’ en een tabel ‘items’, waarin alle exemplaren van boeken (en eventuele andere materialen) van de bibliotheek geregistreerd zijn. Er kan niet iets worden uitgeleend zonder dat het systeem weet welk exemplaar dat was en aan wie dat is gebeurd.

Er zijn ook allerlei voorbeelden waar relaties niet hoeven te voldoen aan een verplichte structuur, maar waar ze wel standaard getypeerd worden. Een voorbeeld daarvan zijn de relaties tussen begrippen in een semantisch netwerk of in een ontologie, die we in hoofdstuk 9 tegenkwamen. Datzelfde geldt voor de in datzelfde hoofdstuk behandelde *linked open data* en voor de relatie tussen object en subject die middels predicaten in RDF-tripels getypeerd worden. Ook literatuurreferenties en de daarvoor opgezette systemen van *reference linking* zijn een specifiek voorbeeld van getypeerde relaties. Een ander voorbeeld tenslotte zijn de relaties in het FRBR-model. FRBR staat voor “*functional requirements for bibliographic records*”. Eigenlijk zijn dit maar deels *requirements* en is FRBR vooral een metadatamodel voor relaties tussen bibliografische entiteiten, een *entity-relationship* model.

FRBR

Afbeelding 12.1 toont een schema met de belangrijkste in FRBR gedefinieerde entiteiten en de relaties daartussen. Daarin zie je bijvoorbeeld dat een beschrijving van een boek in vier niveaus is opgesplitst, elk met hun eigen set metadata. In de eerste plaats is er op het meest abstracte niveau het “werk”, wat staat voor het intellectuele product, bijvoorbeeld “The catcher in the rye” van J.D.Salinger. Dat komt (iets concreter) tot uitdrukking in een of meer “expressies” van het werk, bijvoorbeeld verschillende vertalingen ervan, de Nederlandse (“De vanger in het graan”), de Franse (“L’Attrape-cœurs”), de Duitse (“Der Fänger im Roggen”) en de Finse (“Sieppari ruispellossa”). De relatie tussen het werk en de expressies ervan is dan (in het Engels) getypeerd als *is realized through*. Nog een stapje concreter zijn de “manifestaties” daarvan, bijvoorbeeld de tweede en de derde druk van de Finse vertaling. De relatie tussen een expressie en zijn manifestatie is getypeerd als *is embodied in*. Het meest concreet tenslotte zijn de individuele (in dit geval fysieke) exemplaren, de “items”, bijvoorbeeld het exemplaar dat op 3de plank links staat. *Is exemplified in* is de relatie tussen een manifestatie en de items

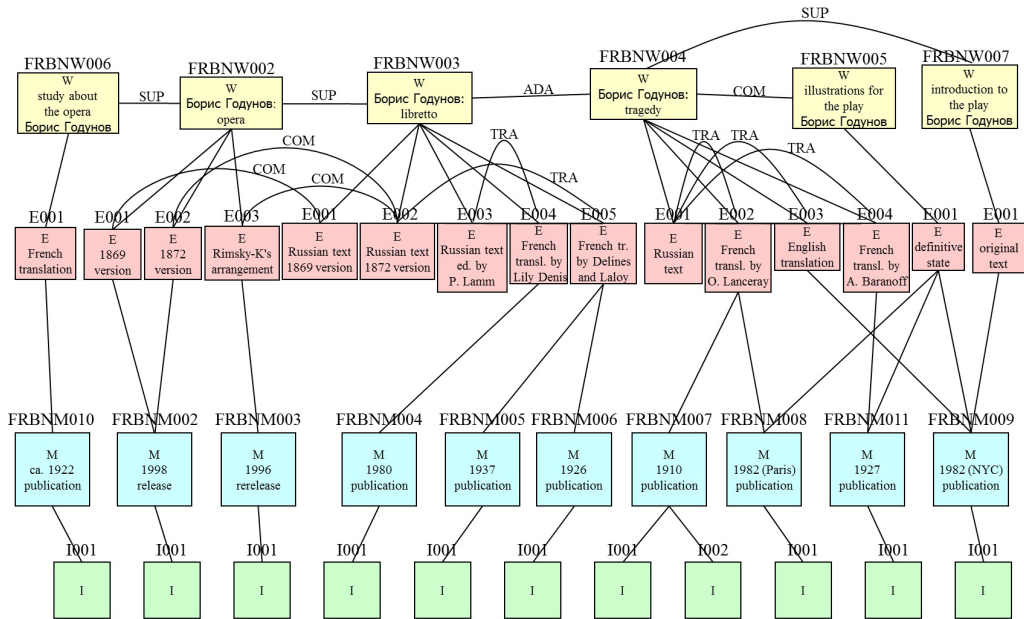
daarvan. In afbeelding 12.1 staan ook getypeerde relaties met andere (soorten) entiteiten vermeld. Zo geldt voor het werk “The catcher in the rye” dat het *is created by* een “person”, namelijk J.D.Salinger.



Afb. 12.1 - Schema met de belangrijkste entiteiten en relaties die in het FRBR datamodel gedefinieerd zijn.

Afbeelding 12.2 toont een heel complex voorbeeld hoe dit er in de praktijk kan uitzien. Je ziet daaraan dat nog wel meer gedefinieerde types relaties mogelijk zijn en ook dat dit model kan worden toegepast voor een grote variëteit aan intellectuele producten die zelf ook weer aan elkaar gerelateerd kunnen zijn. FRBR vormt de basis voor de nieuwe internationale RDA catalogiseer-standaard die in gebruik is in Worldcat en waarop ook veel bibliotheken in Nederland zijn overgegaan. Naast systemen waarin relaties strikt gestructureerd en/of getypeerd zijn, is het ook mogelijk alles vrij te laten. Dat is het geval bij hyperlinks in hypertextsystemen (zoals het web) waar de rest van dit hoofdstuk over gaat. Je kunt dus zeggen dat hypertext een meer algemeen framework is voor linking tussen willekeurige items, entiteiten of objecten, waarbij de achterliggende relaties niet noodzakelijkerwijs een strikte structuur volgen

of getypeerd hoeven te zijn. Als je een link in een webpagina aanklikt, weet je niet bij voorbaat wat er zal gebeuren, je weet alleen dat je ergens terecht zult komen. Dat in tegenstelling tot de juist gegeven voorbeelden van getypeerde relaties: bij een literatuurreferentie weet je bij voorbaat dat die naar een andere publicatie verwijst.



Afb. 12.2 - Illustratie van toepassing van het FRBR-model voor meer dan alleen puur bibliografische gegevens. Behalve de relaties met expressies, manifestaties en items, toont het ook relaties met producten die zijn afgeleid van het treurspel het treurspel "Boris Godunov" en zelf weer Werken zijn (voorbeeld ontleend aan Patrick LeBoeuf - BNF)

12.2 Hypertext

Je zou een hypertext (als ding) kunnen definiëren als een informatiesysteem:

- dat is opgebouwd uit individuele brokstukjes informatie in allerlei verschijningsvormen,
- waartussen relaties (hyperlinks) zijn gelegd,
- waardoor die tekst een systeem vormt dat niet-sequentieel georganiseerd is,
- maar via systematische en/of functionele en/of associatieve relaties tussen die componenten.

Daarbij is het aan de gebruiker te beslissen welke hyperlinks wel of niet te activeren, dus hoe door het systeem te navigeren. Omdat de brokstukjes informatie niet uit alleen tekst hoeven te bestaan maar ook beeld en geluid kunnen bevatten, wordt ook wel van hypermedia gesproken. De hier gegeven definitie is veel algemener dan alleen van toepassing op websites of het web als geheel. Hypertextsystemen bestaan dan ook al veel langer dan het web, zoals uit een historische terugblik blijkt.

Vrijwel elk overzicht over hypertext begint met de naam van Vannevar Bush. Hij was de wetenschapsadviseur van de Amerikaanse president in de periode rond de tweede wereldoorlog. In 1945 publiceerde hij in het (nog altijd in digitale vorm bestaande) blad "Atlantic Monthly" een artikel onder de titel "As we may think", waarin hij een toekomstbeeld schetste voor de overdracht van vooral wetenschappelijke informatie (Bush, 1945). Hij schreef daarin onder meer:

"Our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purposes.

[....]

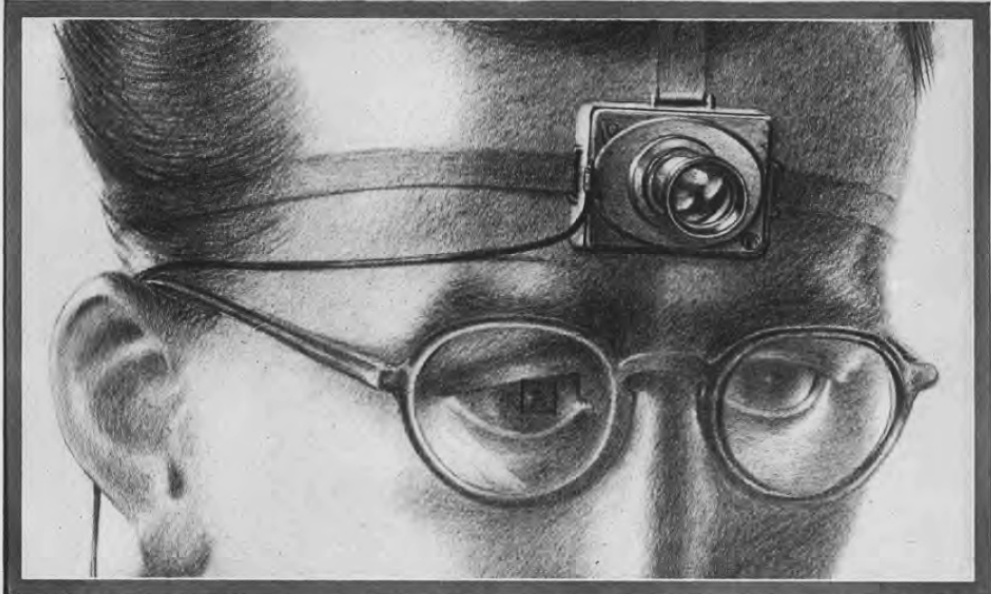
The human mind [...] operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain"

Memex

Om dit associatieve denkproces te ondersteunen, suggereerde hij de ontwikkeling van een apparaat dat hij de Memex noemde. Met dat apparaat dat nog gebaseerd zou zijn op de technologie van dat moment, fotografische microfiches, moesten tekstfragmenten op verschillende fiches door een gebruiker op een mechanische manier met elkaar in verband gebracht kunnen worden, waarbij ze permanent gelinkt werden. Daarover schreef hij:

"When the user is building a trail, he names it [...] Before him are the two items to be joined, projected to adjacent viewing positions. [...] The user taps a single key and the two items are permanently joined. [...] Thereafter, at any time, when one of these items is in view, the other can be instantly recalled merely by tapping a button."

Dat klinkt inderdaad als onze huidige hypertextsystemen. Uiteindelijk is die puur mechanische Memex er nooit gekomen. Pas toen men computers voor het verwerken van tekst ging gebruiken, kon deze opmerkelijke toekomstvisie van Vannevar Bush gerealiseerd worden.



A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD
IN WHICH MAN-MADE MACHINES WILL START TO THINK

by VANNEVAR BUSH

DIRECTOR OF THE OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

Condensed from the *Atlantic Monthly*, July 1945

This has not been a scientists' war; it has been a war in which all have had a part. The scientists, burying their old professional competition in the demand of a common cause, have shared greatly and learned much. It has been exhilarating to work in effective partnership. What are the scientists to do next?

For the biologists, and particularly for the medical scientists, there can be little indecision, for their war work has hardly required them to leave the old paths. Many indeed have been able to carry on their war research in their familiar peacetime laboratories. Their objectives remain much the same.

It is the physicists who have been thrown most violently off stride, who have left academic pursuits for the making of strange destructive gadgets, who have had to devise new methods for their unanticipated assignments. They have done their part on the devices that made it possible to turn back the enemy. They have worked in combined effort with the physicists of our allies. They have felt within themselves the stir of achievement. They have been part of a great team. Now one asks where they will find objectives worthy of their best.

• • •

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for prog-

ress, and the effort to bridge between disciplines is correspondingly superficial.

Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. If the aggregate time spent in writing scholarly works and in reading them could be evaluated, the ratio between these amounts of time might well be startling. Those who conscientiously attempt to keep abreast of current thought, even in restricted fields, by close and continuous reading might well shy away from an examination calculated to show how much of the previous month's efforts could be produced on call.

Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it. This sort of catastrophe is undoubtedly being repeated all about us as truly significant attainments become lost in the mass of the inconsequential.

Publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.

But there are signs of a change as new and powerful instrumentalities come into use. Photocells capable of seeing things in a physical sense, advanced photography which can record what is seen or even what is not, thermionic tubes capable of controlling potent forces under the guidance of

Afb. 12.3 - Een verkorte versie van het oorspronkelijke artikel van Vannevar Bush uit de *Atlantic Monthly* stond op 10 september 1945 in het Amerikaanse publiekstijdschrift *Life Magazine*.

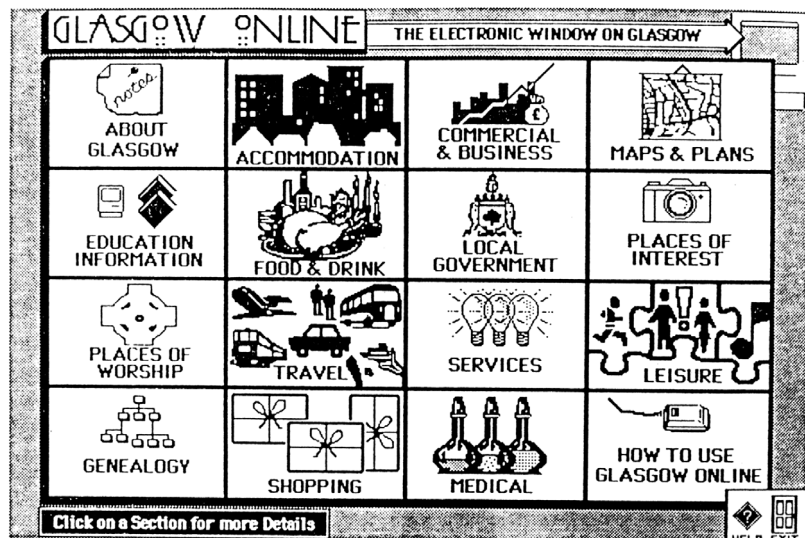
hypertext

Het woord hypertext zelf is dan ook van recenter datum. In de jaren '60 van de vorige eeuw werkte Ted Nelson aan een project Xanadu, waarin literaire teksten in een computer verwerkt werden. Daarvoor ontwikkelde hij linking mechanismen. Hoewel hij in dat verband het woord hypertext al eerder gebruikt schijnt te hebben, introduceerde hij het formeel tijdens een lezing over "Complex Information Processing" in augustus 1965, twee decennia na Vannevar Bush (Nelson, 1965):

"Let me introduce the word "hypertext" to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper. It may contain summaries, or maps of its contents and their interrelations: it may contain annotations, additions and footnotes from scholars who have examined it. Let me suggest that such an object and system, properly designed and administered, could have great potential for education, increasing the student's range of choices, his sense of freedom, his motivation, and his intellectual grasp. Such a system could grow indefinitely, gradually including more and more of the world's written knowledge."

hypercard

Het heeft daarna nog weer twee decennia geduurd voordat hypertext ook bij een iets groter publiek bekend werd. In 1987 verscheen het door Bill Atkinson bij Apple ontwikkelde computerprogramma HyperCard. Wie een Apple Macintosh kocht, kreeg dat programma erbij. Daarmee kon je op eenvoudige wijze hypertextsystemen bouwen, die bestonden uit een soort digitale bibliotheekkaartjes die je naar willekeur aan elkaar kon linken. Er hoorde ook een programmeertaal HyperTalk bij. In die periode zijn op basis van HyperCard veel leuke hypertext toepassingen ontwikkeld.



Afb. 12.4 - Glasgow Online, voorbeeld van een met HyperCard gebouwd hypertext Informatiesysteem.

HTML
HTTP

Maar al kort daarop in de jaren 1989-1991 ontwikkelde Tim Berners-Lee een hypertextsysteem voor internet. Oorspronkelijk bedoeld als intern informatiesysteem voor CERN in Genève, groeide het vanaf 1991 uit tot het World Wide Web. In de benaming van de twee belangrijkste technische standaarden waarop het web berust, verwerkte Berners-Lee het woord hypertext; HTML is de HyperText Markup Language en HTTP het HyperText Transfer Protocol. Nu, nog eens 25 jaar later, mogen we dus rustig zeggen dat hypertext intussen overal is.



Afb. 12.5 - De vier hoofdpersonen in de geschiedenis van hypertext: Vannevar Bush, Ted Nelson, Bill Atkinson en Tim Berners Lee.

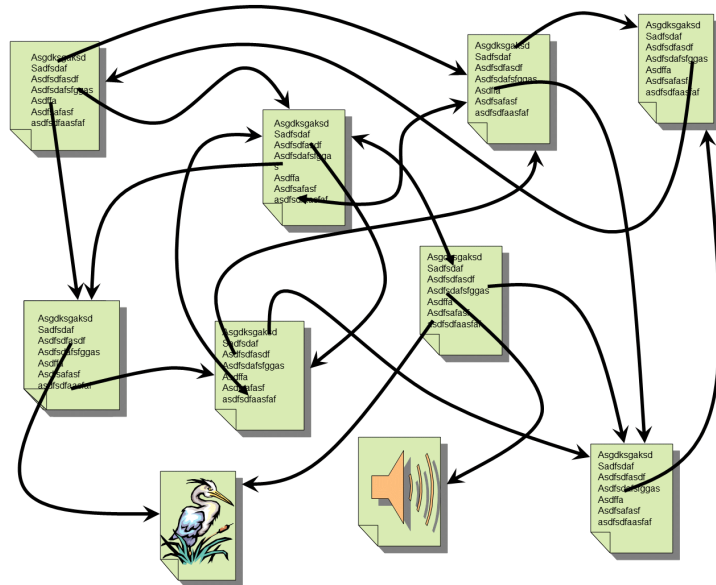
12.3 Structuur in hypertext

De eerdere definitie van wat een hypertext is, zou je kunnen visualiseren zoals in afbeelding 12.6. Dit is een voorbeeld waarin eigenlijk alleen gebruik gemaakt is van zogenaamde associatieve relaties. Enkele voorbeelden daarvan zijn hyperlinks naar:

associatieve
relaties

- een ander stukje tekst over hetzelfde onderwerp;
- uitleg van een onderwerp;
- uitleg van een woord;
- een inhoudelijke noot of verwijzing;
- informatie over een verwant onderwerp;
- een illustratieve afbeelding;
- een illustratief geluidsfragment;
- een illustratieve video;
- een literatuurreferentie.

De mogelijkheid om zulke links aan te brengen vormt de grote kracht van hypertext.



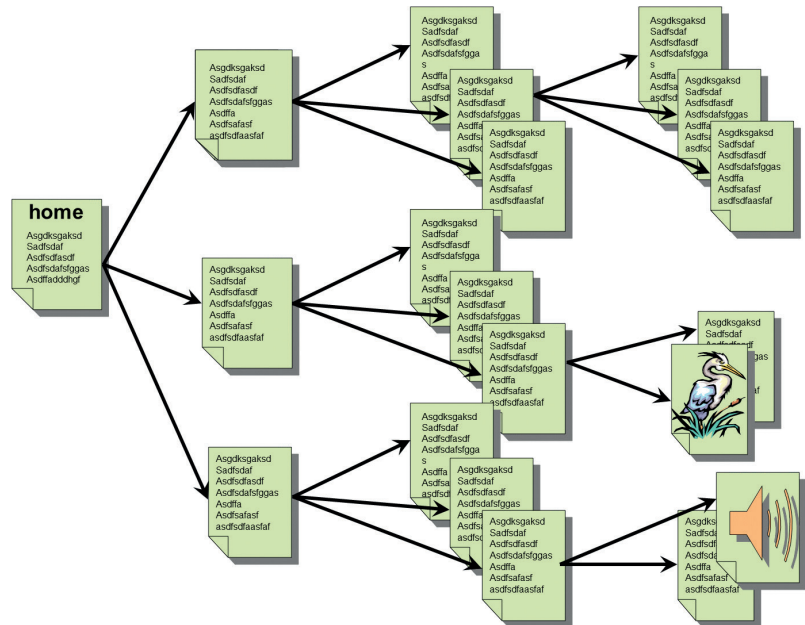
Afb. 12.6 - Visualisatie van een eenvoudige hypertextstructuur, waarin alleen associatieve relaties.

Een hypertext met alleen dit soort links dreigt echter al snel een soort spaghettistructuur te krijgen. Daarom zul je ook vaak hyperlinks voor meer systematische navigatie tegenkomen. Daarbij kun je denken aan:

- links volgens een systematische structuur van het systeem (een boomstructuur);
- link naar een volgend onderwerp;
- link naar een vorig onderwerp;
- teruglink naar het vorige scherm;
- vooruitlink om (weer) door te gaan naar het volgende scherm;
- link naar een beginscherm (homepage);
- link naar een inhoudsopgave of index;
- link naar een zoekscherm;
- link naar een (ander) algemeen systeem-specifiek scherm, bijvoorbeeld een landkaart.

Dit soort links is dan vaak op een vaste plaats op elk scherm aanwezig. Een extreme invulling hiervan is dat in een systeem alleen nog maar links voor een systematische hiërarchische structuur aanwezig zijn. Dan krijg je een situatie als in afbeelding 12.7. Daar is feitelijk geen sprake meer

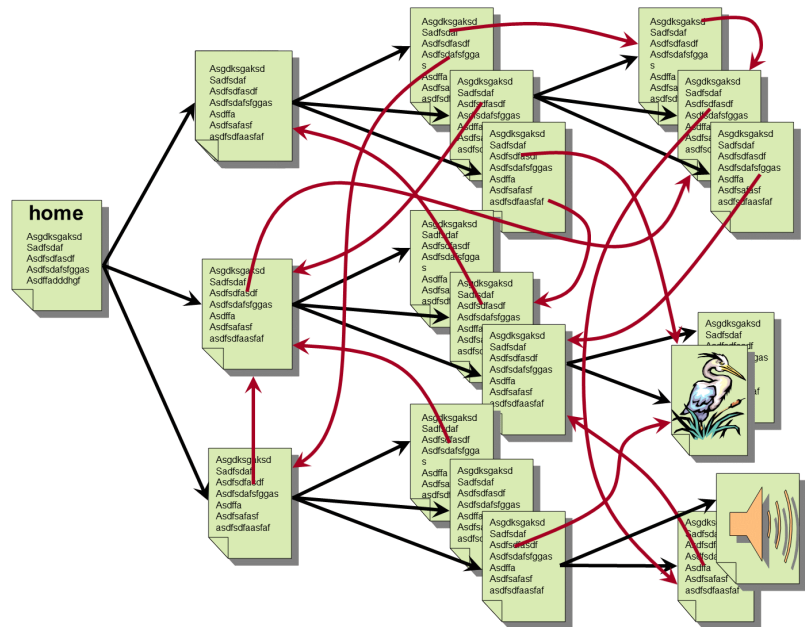
van een hypertext, maar alleen nog van een simpele taxonomie. Toch zijn er vrij veel websites die alleen dit soort links bevatten. De automatisch gegenereerde navigatiemenu's in veel contentmanagementsystemen stimuleren dat ook wel.



Afb. 12.7 - Een hypertext met alleen hiërarchische relaties is meer een taxonomie.

Een echte hypertext zal liefst een combinatie bevatten van zowel de vaste links van een systematisch hiërarchische structuur, als links voor meer associatieve navigatie. In afbeelding 12.8 is dat gevisualiseerd.

Het bouwen van een hypertext vereist (uiteraard) een weldoordacht ontwerp. Zoals bij elk informatiesysteem is daar systematiek en consistentie voor nodig. Het maken van zo'n ontwerp is arbeidsintensief, al was het alleen maar omdat je behalve over de systematische structuur, ook moet beslissen welke associatieve links je wilt aanbrengen. Voor een systeem met 100 pagina's zijn er al 100 x 99 potentiële links. Die zijn natuurlijk lang niet allemaal zinnig, maar er moet wel bedacht worden voor welke daarvan dat wel het geval is. Ook het bouwen zelf is arbeidsintensief. Er kan niet worden volstaan met het eenvoudigweg importeren van bestaand materiaal. Voor elke associatieve link moet het beginpunt (het anker) gemarkeerd en de bestemming (het doel) gespecificeerd worden.



Afb. 12.8 - Hypertext met zowel hiërarchische structuur als associatieve relaties.

Bij hyperlinks op websites spelen nog twee speciale aspecten mee. In de eerste plaats worden niet alleen links gelegd tussen de pagina's op de eigen site, maar ook naar externe pagina's, elders op het web. Die pagina's zullen dan vrijwel nooit deel uitmaken van een structuur zoals we hiervoor bespraken. Daarom is het wel aan te raden om in de vormgeving van de links herkenbaar te maken dat de eigen site wordt verlaten. Dat kan door middel van het bij de link toevoegen van een tekstaanduiding of van een daarvoor gebruikelijk standaard symbool , zoals onder meer in de Wikipedia wordt gebruikt. Ook wanneer in plaats van een tekstfragment een URL als linktekst wordt gebruikt, is daaraan (wat meer indirect) te zien dat naar een andere site gegaan wordt.

Een tweede aspect betreft het zogenaamd dieplinken. Binnen de eigen website kan daaronder worden verstaan dat niet naar het begin van een pagina wordt gelinkt, maar naar een bepaald punt ergens binnen een pagina, dat specifiek relevant is voor de context waarvandaan gelinkt wordt. De HTML markup maakt het eenvoudig mogelijk om dit te realiseren. Of dit bij links naar een externe pagina ook kan, hangt ervan af of de maker daarvan in die pagina ook al de daarvoor noodzakelijke markeringen heeft aangebracht. Onder dieplinken naar externe sites wordt meer gebruikelijk verstaan dat niet naar de home page van die site, maar rechtstreeks naar een specifieke onderliggende pagina wordt gelinkt. Dat kan betrekkelijk eenvoudig als elke pagina op de betreffende site een eigen statisch URL heeft. Het wordt soms al iets lastiger bij sites waar pagina's middels queries uit een contentmanagementsysteem worden opgeroepen

of dynamisch worden samengesteld. Hoewel vaak heel complex en niet altijd eenvoudig te reconstrueren, hebben die pagina's meestal nog wel een eigen URL. Ook bijvoorbeeld het resultaat van een Google-zoekactie kan zo in een link verwerkt worden. Er zijn echter ook voorbeelden waar alle pagina's van een site hetzelfde URL hebben. Daar worden ze via interne scripts opgevraagd. In die gevallen is dieplinken vrijwel nooit mogelijk.

Een bijzondere vorm van dieplinken is van belang bij video's. In bijvoorbeeld nieuwsuitzendingen wil je vaak niet bij het begin van die uitzending uitkomen, maar veel liever direct bij het voor de link relevante fragment, om niet eerst een kwartier van de uitzending af te hoeven spelen. Het hangt dan af van de videodienst waarnaar gelinkt wordt, of zoiets mogelijk is. Bij YouTube kan dat door in de hyperlink een bepaalde parameter aan de URL van een video toe te voegen. Met bijvoorbeeld `&start=267` geef je aan dat de video op 267 seconde van het begin moet starten.

Eerder gaven we aan dat hyperlinks in een hypertextsysteem niet noodzakelijkerwijs op formeel getypeerde relaties gebaseerd zijn. Toch zijn er op het web wel mogelijkheden om zulke typering toe te voegen. Denk daarvoor bijvoorbeeld aan de relaties die door predicaten in RDF-tripels tot uitdrukking gebracht kunnen worden (zie paragraaf 9.3.4). Via bijvoorbeeld een codering in RDFa kunnen die relaties tussen webpagina's in de attributen van de HTML-codes voor de hyperlinks verwerkt worden. Slimme software kan daar dan iets mee doen. Daarnaast wordt ook een speciale standaard ontwikkeld om op technisch andere wijze typering mogelijk te maken. Daaraan werkt IANA (Internet Assigned Numbers Authority) in samenwerking met IETF (Internet Engineering Task Force). In die standaard wordt ook gedefinieerd welke soorten links kunnen worden onderscheiden. Dat zijn er intussen al meer dan 75 (Nottingham, 2016)².

² Zie ook: <https://tools.ietf.org/html/rfc5988>

12.4 Gebruiksproblemen van hypertext

Hypertextsystemen worden gewoonlijk als gebruiksvriendelijk gezien: je hoeft alleen maar iets aan te klikken om informatie te verkrijgen. Toch worden gebruikers, zoals bij elk informatiesysteem, ook hier wel met gebruiksproblemen geconfronteerd. De belangrijkste daarvan hangen samen met de volgende drie vragen:

- Welke informatie zit er in het systeem (zoals een website)? Heeft het dus zin hier verder te gaan kijken?
- Waar zit die informatie in het systeem en hoe kom ik daar? Welke “86 links” moet ik daarvoor achtereenvolgens aanklikken?
- Waar ergens ben ik na een heleboel klikken in het systeem terecht gekomen? Het “*lost in HyperSpace*” gevoel.

Voor elk van deze drie problemen zijn wat algemene oplossingen of best-practices te geven. Sommige daarvan zijn in het algemeen van toepassing op informatiesystemen, andere zijn heel specifiek voor hypertext.

1. Wat zit er in het systeem?

Zorg dat daarover voldoende overzichtelijke informatie gepresenteerd wordt:

- Geef vooraf, op de beginpagina, de homepage, informatie over behandelde onderwerpen, niveau, soort informatie, doel, doelgroep etcetera.
- Presenteer een overzicht over het systeem in de vorm van een inhoudsopgave of een index, waarin alle (hoofd)onderwerpen genoemd worden.

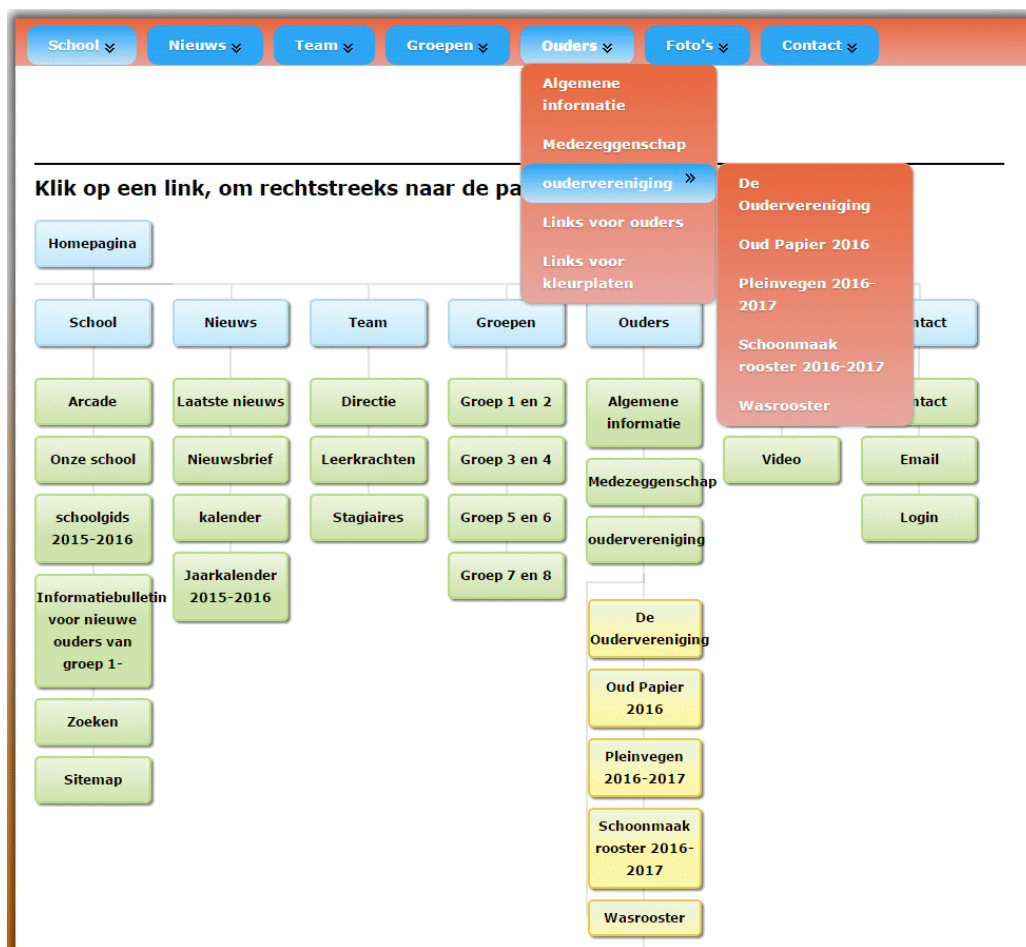
2. Waar zit het in het systeem; hoe kom ik waar ik wil zijn?

Zorg voor gerichte zoek- en navigatiemogelijkheden, zodat het niet nodig is 86 links in de juiste volgorde aan te klikken om op de gewenste plek te komen:

- Zorg dat er een systematische basisstructuur is (zie afbeelding 12.8). Daarvoor gelden dezelfde uitgangspunten als bij een taxonomie. Zie hoofdstuk 3 voor een behandeling daarvan.
- Zorg voor een overzichtelijke visualisatie van die structuur in een sitemap of siteview, zoals bijvoorbeeld met uitklapmenu's. Dergelijke visualisaties worden meestal automatisch uit een mappenstructuur binnen een contentmanagementsysteem gegenereerd, zodat ze alleen de hiërarchische structuur van de site tonen (zie afbeelding 12.9).
- Zorg voor een echte zoekfunctie.
- Zorg dat een inhoudsopgave met directe hyperlinks aanwezig is.
- Zorg dat een alfabetische index met directe hyperlinks aanwezig is.

systematische
basisstructuur

sitemap



Afb. 12.9 - Via uitklapmenu's getoonde sitemap van de website van een basisschool. In dit voorbeeld is ook nog een webpagina aanwezig die een statische representatie van de sitemap geeft.

3. Waar ben ik nu in het systeem?

Het "lost in hyperspace" syndroom was in klassieke systemen onbekend. Het is tegen te gaan door op elke pagina standaard systematische navigatielinks op te nemen:

systematische
navigatielinks

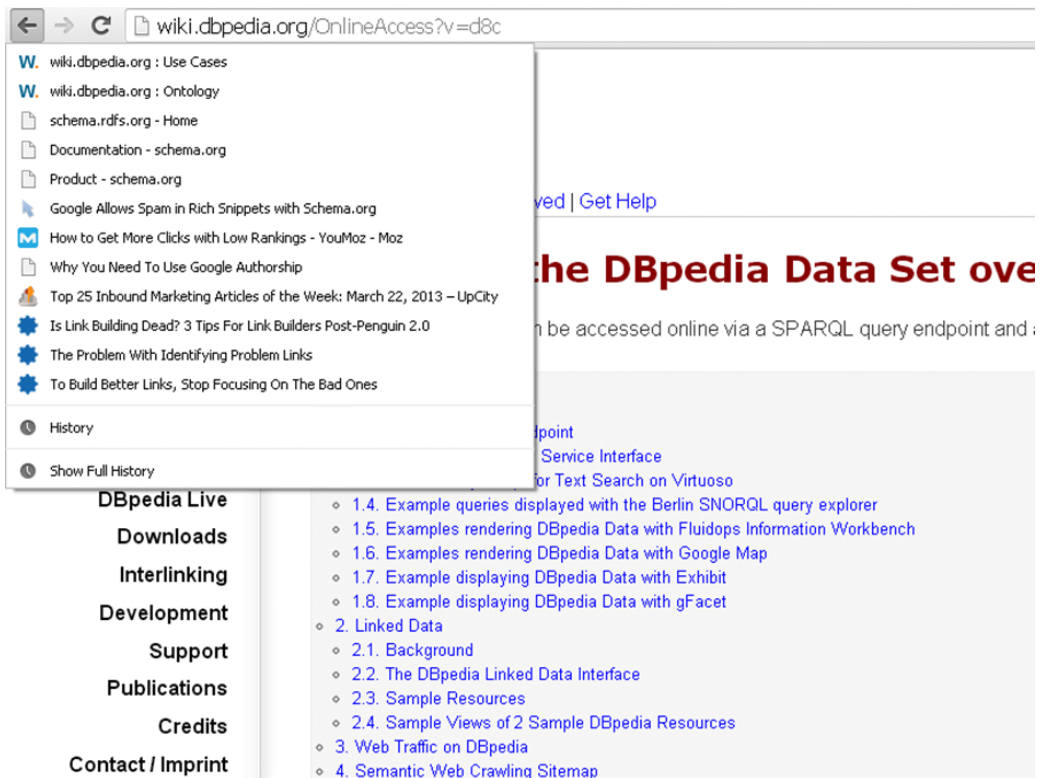
- Een standaard terugknop naar het vorige scherm. In webbrowsers is die al standaard ingebouwd om naar de vorige bekeken pagina terug te gaan.
- Een standaardknop terug naar de homepage. Op het web moet daarbij onderscheid worden gemaakt tussen de homepage van de website, die door de bouwer van de website gekozen wordt, en de homepage van de webbrowser, die (los van bezochte websites) door de gebruiker wordt ingesteld.
- Een overzicht of beeld van het gevolgde navigatiepad. Geïnspireerd op Klein Duimpje wordt dat ook wel kruimelpad genoemd.

navigatiepad

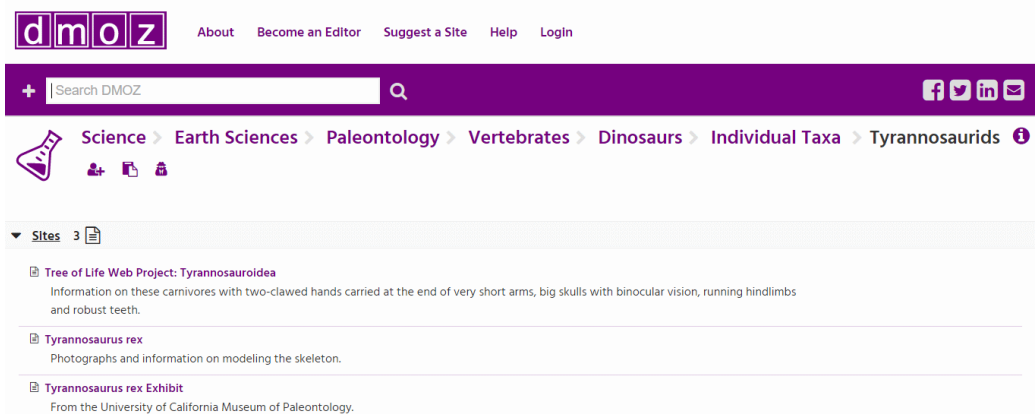
Daarbij moet onderscheid worden gemaakt tussen het echte browse-/navigatiepad dat de gebruiker gevolgd heeft en dat je bijvoorbeeld in de webbrowser kunt oproepen (afbeelding 12.10), en een indicatie waar de huidige pagina in de hiërarchische structuur van de website past (wat soms ten onrechte ook als kruimelpad wordt aangeduid. Zie hiervoor afbeelding 12.11).

visualisatie

- Een visualisatie van de omgeving van de huidige pagina, waarbij zowel de systematische structuur als een *linklist* van associatieve hyperlinks vooruit getoond wordt.



Afb. 12.10 - Voorbeeld van de via de terugknop in de Chrome browser op te roepen browsegeschiedenis.

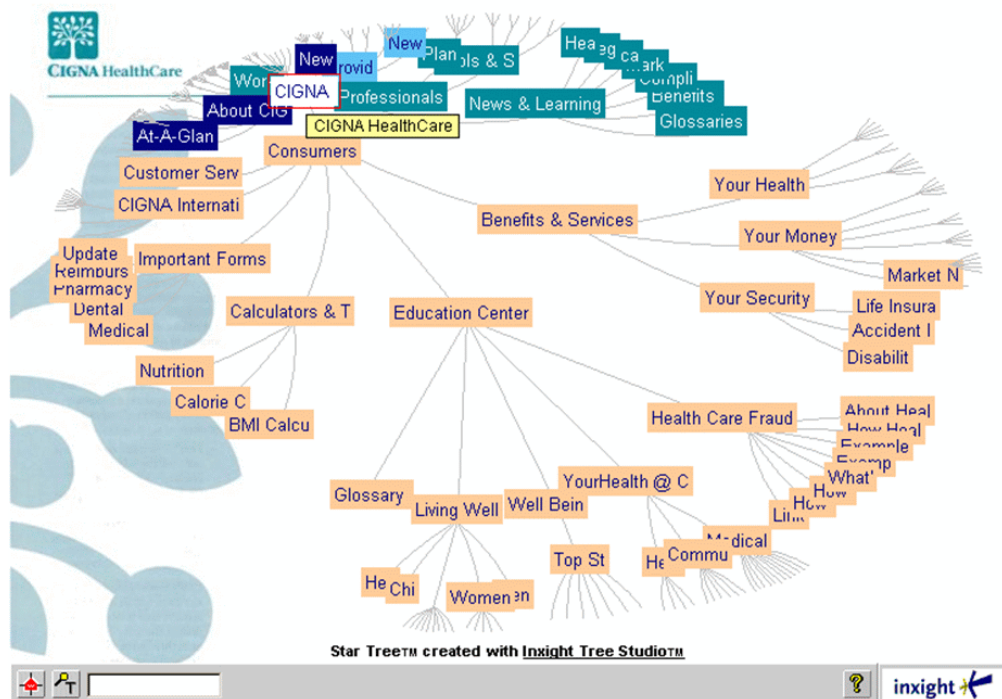
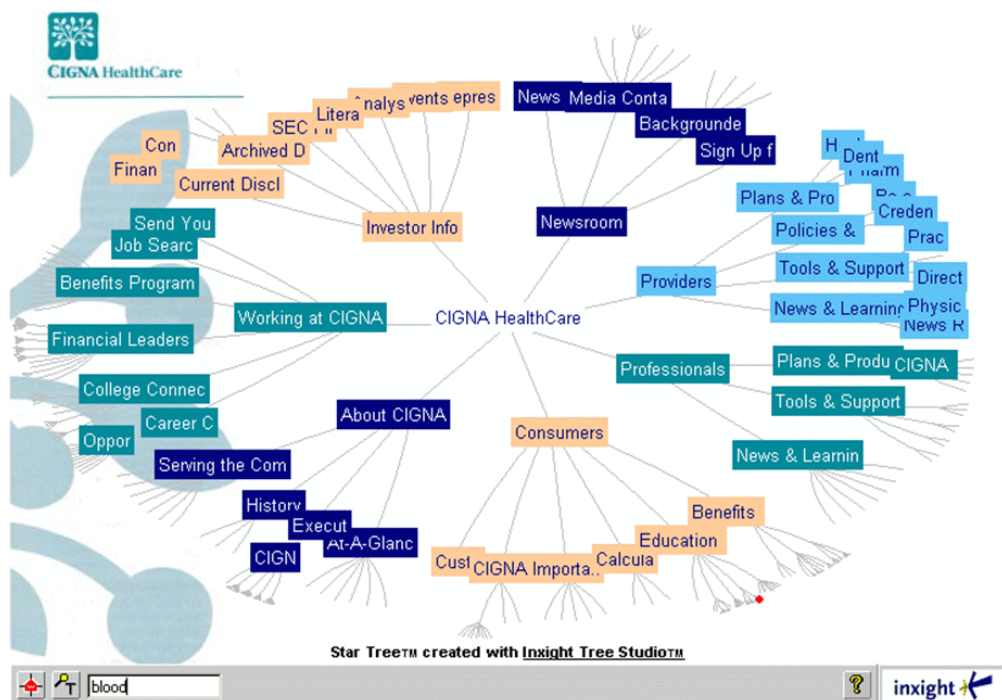


Afb. 12.11 - Voorbeeld van een oneigenlijk kruimelpad bovenaan een webpagina, dat de structuur van de niveaus boven de huidige webpagina aangeeft.

Visualisatie van de structuur van een site of van de omgeving van een pagina werd hierboven regelmatig als mogelijke oplossing genoemd. Voordeel van zo'n visualisatie is dat die snel een overzicht kan geven van de structuur van de site en de onderlinge verbanden. Toch blijft zo'n overzicht meestal afhankelijk van inhoudelijke beschrijvingen. Visualisaties laten echter weinig ruimte voor beschrijvende tekst, zodat die beschrijvingen wel kernachtig moeten zijn. De paginatitels van webpagina's zijn daar vaak te lang voor. Een beperking van veel visualisaties is bovendien dat ze meestal alleen de systematische structuur laten zien en niet de aanwezige associatieve links kunnen tonen.

Bij heel grote aantallen hyperlinks of onderliggende bestanden en mappen dreigt de visualisatie zelf al weer snel onoverzichtelijk te worden. Bedenk dat een hiërarchie van vier niveaus, waarbinnen elke pagina tien subpagina's heeft, al resulteert in een systeem met ruim 10.000 pagina's. Menu's die pas openklappen als je er met de muis overheen gaat, zoals het voorbeeld in afbeelding 12.9, proberen dat wel te ondervangen, maar die methode is weer niet zo geschikt wanneer de structuur wat dieper is met een uit meer dan twee lagere niveaus bestaande hiërarchie. Een aanpak die ontwikkeld is om het probleem van grote aantallen en vele niveaus te ondervangen, zijn de zogenaamde hyperbolische browsers. Aanvankelijk tonen die bijvoorbeeld alleen de eerste twee niveaus, rondom de beginpagina. Diepere niveaus worden alleen gesuggereerd door loshangende linkjes die als het ware over de rand van een bol hangen. Door de afbeelding met de muis in een bepaalde richting te verslepen trek je de aanvankelijk nog onzichtbare pagina's in de tegenovergelegen richting ook in beeld. Dat is te zien in het voorbeeld in afbeelding 12.12. De laatste jaren lijkt er wat minder interesse te zijn voor dit soort complexe visualisaties, wellicht omdat gebruikers steeds sterker geneigd zijn van een zoekfunctie gebruik te maken. Een bespreking van diverse visualisatietechnieken is onder meer te vinden in Shneiderman (2012).

hyperbolische browsers



Afb. 12.12 - Voorbeeld van een website waarvan de structuur middels een hyperbolische boom wordt. In de beginsituatie zijn dat alleen de eerste niveaus (boven). Door het slepen met de muis, komen over de rand hangende diepere pagina's ook in beeld (onder).

12.5 Hyperlinks maken toegankelijk

In het begin van dit hoofdstuk gaven we al aan dat hyperlinks in feite een middel zijn om informatie inhoudelijk toegankelijk te maken. Dat is ook wat formeler te onderbouwen door de constatering dat het verwijzingen zijn naar informatie-items die een of andere inhoudelijke relatie hebben met het vertrekpunt van die link en die relevantie hebben met betrekking tot inhoud en context van dat vertrekpunt. Niettemin zijn er ook duidelijke verschillen met andere vormen van toegankelijk maken zoals met taxonomieën of thesauri. Een paar verschilpunten zijn in onderstaande tabel verder uitgewerkt.

Hyperlinks	Thesaurus / taxonomie
Gevoelsmatig toegankelijk systeem	Kunstmatige informatietaal
Associatieve links kennen geen inhoudelijke beperkingen; je kunt overal naar linken	Inhoudelijk bereik beperkt; niet voor elk onderwerp is een geschikte term gedefinieerd
Associatieve links kunnen in precieze inhoudelijke context worden aangebracht	Een toegekende term of klasse geeft meestal geen indicatie in welke context dat onderwerp in de ontsloten tekst aan de orde komt
Volledige vrijheid van toepassing	Beperkte vrijheid van toepassing
Linken naar omvangrijke volledige verzameling van "alles" over onderwerp niet goed mogelijk	Kan verwijzen naar "alles" over een onderwerp
Link verwijst meestal naar één specifiek gekozen object / item	Ongewenst dat term of klasse verwijst naar slechts één document / object
Link kan verwijzen naar losse fragmenten / onderdelen van documenten	Term of klasse verwijst meestal naar gehele documenten
Toevoegen van onderwerp / informatie heeft invloed op het volledige systeem; alle potentiële links moeten beoordeeld worden	Toevoegen van een onderwerp beperkt zich in de praktijk tot het onderwerpssysteem

12.6 Hyperfictie

hyperfictie

Met de komst van hypertext, is ook hyperfictie (Engels: *hypertext fiction*) ontstaan. Hoewel men het er niet altijd over eens is wat daar heel precies onder verstaan moet worden, kun je in zijn algemeenheid zeggen dat het om vertellingen gaat die zich vertakken en waar de lezer bepaalt welke tak en dus welk vervolg hij kiest. De eerste publicatie die algemeen als hyperfictie wordt beschouwd is "Conte à votre façon", een creatie van de Franse auteur Raymond Queneau. Hij schreef dat al in 1967. De term hypertext was weliswaar al twee jaar eerder door Ted Nelson bedacht, maar diens hypertextprogramma (voor een mainframe computer) was pas

een jaar later, in 1968 voltooid. Het verhaal van Queneau, beter bekend in zijn Engelse vertaling “A Story as you like it”, was dan ook gewoon op papier uitgekomen. Op de website van de Koninklijke Bibliotheek (<https://www.kb.nl/themas/collectie-koopman/un-conte-a-votre-facon>) wordt de interactiviteit van het verhaal als volgt beschreven:

“Het verhaal is opgezet als een ganzenbordspel, hier bepaalt echter niet de dobbelsteen, maar de lezer wat er gebeurt. Op pagina 1 wordt al gevraagd of de lezer een verhaal over drie alerte erwten wil horen. ‘Zo ja, ga door naar 4. Zo nee, ga door naar 2’. De nee-zeggers krijgen op pagina 2 de vraag of ze een verhaal over bonenstaken willen horen en dan is er bij 3 nog de kans op een verhaal over drie doorsnee struiken. Als de lezer ook dat niet wil, wordt hij verwezen naar pagina 21, waar het verhaal resoluut naar een einde wordt gevoerd.”

Een echt lang verhaal is het dus niet geworden. Intussen is op het web uiteraard ook een digitale (Engelstalige) versie beschikbaar (http://www.thing.de/projekte/7:9%23/queneau_1.html). De komst van het web heeft Queneau overigens niet meer meegemaakt. In 1991 was hij al vijftien jaar dood. Niettemin zijn er ook nog wel andere interessante, op automatisering vooruitlopende producten van zijn hand (Sieverts, 2009). Een paar andere voorbeelden van hyperfictie, die wel in het hypertexttijdperk zijn geschreven:

- “*afternoon, a story*” van Michael Joyce uit 1987 (<http://www.wwnorton.com/college/english/pmaf/hypertext/aft/>),
- “*Victory Garden*” van Stuart Moulthrop uit 1992 (<http://www.eastgate.com/VG/VGStart.html>),
- “*Omphaloskepsis*” van Jay Dillemoth uit 1996 (<http://www.altx.com/hyperx/ompha/Title.html>).

Gezien het feit dat in het huidige webtijdperk vrijwel iedereen, al is het vaak onbewust, bezig is met hypertext, is het opmerkelijk dat er sinds 2000 eigenlijk nauwelijks meer iets nieuws is verschenen dat als hyperfictie aangeduid zou kunnen worden (Lafarge, 2011).

12.7 Samenvattend

Hyperlinks zijn de lijn die hypertextinformatiesystemen (zoals websites ook zijn) bij elkaar houden. Die hyperlinks zijn gebaseerd op relaties die tussen informatie-eenheden bestaan. Daarom is ook in wat algemener verband gekeken naar relaties en hoe die tot uiting komen in bijvoorbeeld relationele databases, RDF en FRBR. Voor systemen die echt op hyperlinks gebaseerd zijn, hebben we gekeken naar de manier waarop we daarmee structuur kunnen aanbrengen. Daarbij is vooral nadruk gelegd op de factoren (in feite ontwerpprincipes) die van belang zijn wanneer je op

deze manier informatie voor gebruikers toegankelijk wilt maken. Ook is besproken in welke opzichten deze manier van toegankelijk maken verschilt van de meer traditionele manieren die in eerdere hoofdstukken aan de orde kwamen.

Gebruikte literatuur

Bush, V. (1945). As we may think. *Atlantic Monthly*, 176(1), 101–108. Geraadpleegd op 6 september 2016, van: <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>

Lafarge, Paul (2011). Why the book's future never happened. In: *Salon*, 4 october 2011. Geraadpleegd op 6 september 2016, van: http://www.salon.com/2011/10/04/return_of_hypertext/

Nelson, T. (1965). A File Structure for the Complex, the Changing, and the Indeterminate. Association for Computing Machinery: Proc. 20th National Conference 1965, 84-100. <http://cs.brown.edu/courses/cs196-9/p84-nelson.pdf>

Nottingham, M., Reschke, J., Algermissen, J. (2016). *Link relation types*. Internet Assigned Number Authority. Geraadpleegd op 11 oktober 2016, van: <http://www.iana.org/assignments/link-relations/link-relations.xhtml>

Shneiderman, B., Dunne, C., Sharma, P., & Wang, P. (2012). Innovation trajectories for information visualizations: Comparing treemaps, cone trees, and hyperbolic trees. *Information Visualization*, 11(2), 87-105. Geraadpleegd op 6 september 2016, van: https://ils.unc.edu/courses/2015_spring/inls541_001/Readings/Shneiderman-Comparing-Treemaps-ConeTrees-Hyperbolic-2012.pdf

Sieverts, Eric (2009). Hij was het web vooruit. Oorspronkelijk verschenen in: *InformatieProfessional, Redactieblog*, 09 april 2009. Geraadpleegd op 6 september 2016, van: <http://sieverts.pbworks.com/f/het-web-vooruit.html>

Hoofdstuk 13

Praktijkvoorbeelden

In dit hoofdstuk zullen we een aantal voorbeelden bespreken van de ontsluitingstechnieken die in dit boek aan de orde zijn geweest. We beginnen met een aantal thesauri die in verschillende domeinen in gebruik zijn. Vervolgens besteden we aandacht aan een aantal classificaties, waarna we afsluiten met twee voorbeelden van modernere technieken: het automatisch classificeren bij het ministerie van BZK en het metadatastandaardschema.org.

13.1 AAT

De Art & Architecture Thesaurus (AAT, <http://browser.aat-net.nl>) is een wereldwijd toonaangevende thesaurus op het gebied van kunst en architectuur die gebruikt kan worden bij het toegankelijk maken van collecties op het gebied van materiële cultuur, toegepaste kunst, archeologie, archiefmaterialen en verwante vakgebieden. De AAT bestaat uit ruim 30.000 begrippen. De Nederlandstalige AAT is een vertaling en bewerking van de Amerikaanse AAT van het Getty Research Institute. De Nederlandstalige AAT is een samenwerkingsproject van het Nederlandse RKD (Rijksbureau voor Kunsthistorische Documentatie) en het Belgische Koninklijk Instituut voor het Kunstpatrimonium (KIK-IRPA). Bureau AAT is verantwoordelijk voor de vertaling en het beheer van de Nederlandstalige AAT. De medewerkers van Bureau AAT voeren, in samenspraak met deskundigen, verbeteringen door aan de data en behandelen nieuwe termvoorstellen. Bureau AAT onderhoudt contact met het Getty Research Institute en met het Nederlandstalige werkveld waaraan zij, zo nodig, ondersteuning biedt.

De Nederlandse AAT bestaat uit een gestructureerde tweetalige termenlijst (de Nederlandse en de Engelse termenlijst). Naast een volledige vertaling naar het Nederlands is de AAT ook vertaald in het Frans, Italiaans, Duits, Spaans en Chinees. Daarmee fungeert het als een internationale standaard voor het beschikbaar stellen van collectie-informatie. De thesaurus is ook te gebruiken binnen het Linked Open Data (LOD) project (Getty Institute).

Zowel in Nederland als in andere landen wordt de AAT vooral gebruikt voor het toegankelijk maken van museale en aanverwante collecties. De indeling van de AAT in zeven hoofdgroepen (facetten) sluit daarop aan. De zeven hoofdgroepen zijn:

1. Abstracte begrippen: onder dit facet vallen abstracte begrippen en verschijnselen die betrekking hebben op uiteenlopende aspecten van het menselijk denken en doen, waaronder architectuur en de kunsten in alle media en andere disciplines. Voorbeelden: schoonheid, evenwicht, metafoor, vrijheid, kapitalisme.
2. Activiteiten: omvat diverse vormen van fysieke of geestelijke bezigheden, maar ook afzonderlijke gebeurtenissen, stelselmatige reeksen handelingen, methoden en processen. Voorbeelden: archeologie, techniek, analyseren, wedstrijden, hardlopen.
3. Fysieke kenmerken: bevat de waarneembare of meetbare eigenschappen van materialen en gebruiksvoorwerpen en de eigenschappen van materialen en gebruiksvoorwerpen die zich niet als afzonderlijke componenten laten scheiden. Fysieke kenmerken zijn onder meer grootte en vorm, chemische eigenschappen van materialen, structuur en hardheid en kenmerken als oppervlakteverfraaiing en kleur. Voorbeelden: vlechtbandmotieven, randen, rond, verzadigd met water, broosheid.
4. Materialen: dit facet omvat fysieke stoffen, van zowel natuurlijke als synthetische oorsprong; van specifieke materialen die op functie zijn ingedeeld, zoals 'kleurmiddelen', tot grondstoffen die zijn gevormd of verwerkt tot producten die bij de vervaardiging van structuren of objecten worden gebruikt. Voorbeelden: ijzer, klei, lijm, emulgator, kunstmatige ivoor, freeswerk.
5. Objecten: hieronder vallen tastbare of visuele zaken die levenloos zijn en die door mensen zijn geproduceerd of vormgegeven. Dit kan uiteenlopen van afbeeldingen tot gebouwen tot geschreven documenten en van gebruiksvoorwerp tot esthetisch object. Voorbeelden: schilderijen, gevels, kathedralen, Brewster stoelen, tuinen.
6. Actoren en organismen: hier vallen mensen, groepen mensen en organisaties onder, op basis van beroep, activiteit, lichamelijke of geestelijke kenmerken, of een maatschappelijke rol, maar ook situaties en dieren en planten. Voorbeelden: grafici, landschapsarchitecten, bedrijven, religieuze ordes.
7. Stijlen en perioden: stijlen en perioden bevat de namen van kunst- en architectuurstijlen, historische perioden en kunststromingen. Voorbeelden: Lodewijk-XIV-stijl, abstract expressionisme.

In principe omvatten deze facetten alle termen die nodig zijn om door mensen gemaakte objecten te beschrijven.

Deze indeling in zeven hoofdgroepen geeft aan dat deze thesaurus gebruik maakt van classificatieaspecten. Dat is in de hele opbouw van de thesaurus te herkennen. Afbeelding 13.1 geeft hier een voorbeeld van. In de afbeelding staat de term 'straten' genoemd, met onderliggende termen, met daarbij tussen punthaakjes een categorieaanduiding. De termen tussen vishaken worden niet gebruikt om materiaal toegankelijk

te maken, maar zijn bedoeld voor raadplegers van de thesaurus. De AAT noemt dit 'gidstermen'.

```

+ straten
- <straten naar functie>
    éénrichtingsverkeerstraten
    processiewegen
    voetgangersstraten
- <straten naar locatie of context>
    + hoofdstraten
        verkeersaders (straten naar locatie of context)
        voorrangswegen
- <straten naar vorm>
    boulevards
    + doodlopende straten
        landwegen
        lanen
        plateia's
    - stegen
        angiporti
        doodlopende steegjes
    wynds
    stratengids
    stratengidsen
  
```

Afb. 13.1 - Term 'straten' in de AAT.

In dit voorbeeld is de meervoudsvorm gebruikt. De AAT heeft alle termen, voor zover mogelijk, in zowel meervoud als enkelvoud opgenomen. Dat is vooral handig voor instellingen als musea, die meestal de enkelvoudsvorm prefereren. Als voorbeeld kijken we naar de termen 'bakkerij' en 'bakkerijen'. Deze komen allebei voor in de thesaurus. 'Bakkerijen' is de voorkeursterm en 'bakkerij' de *equivalent term*.

term	bakkerij
voorkeursterm	bakkerijen
termtype	alternate descriptor
concept id	300005291

Afb. 13.2a - Niet-voorkeursterm 'bakkerij'.

term	bakkerijen
broader term	<winkels naar verhandelde artikelen> Preferred parent
alternatieve term	bakkerij
related term	bakkers
scope note	Gebouwen of gedeelten van gebouwen waar brood en andere gebakken producten worden bereid. Het kan een ruimte voor de verkoop van gebakken producten omvatten.
facet	Objecten
termtype	voorkeursterm
record type	concept
concept id	300005291
used for	bakkerswinkel bakkerswinkels broodbakkerijen broodbakkerij
bron	Van Dale Engels-Nederlands (1989)

Afb. 13.2b - Voorkeursterm 'bakkerijen'.

Een voorbeeld van een collectie die doorzocht kan worden met behulp van de AAT is die van het Rubenianum, het onderzoeksinstituut voor Vlaamse kunst van de 16e en 17e eeuw.

Wie wil zoeken op een term kan gebruik maken van “eenvoudig zoeken” op onderwerp. Als we bijvoorbeeld een zoekactie starten op ‘vaas’, komt er een scherm waar de termen worden weergegeven (een aantal komt van de AAT, maar er zijn ook andere bronnen), met het aantal beschrijvingen in de collectie. ‘Vaas’ blijkt geen voorkeursterm te zijn in de AAT. Er wordt verwezen naar ‘vazen’, waar 50 hits van zijn. Het vervangend zoeken is dus mogelijk, maar nog niet zo gebruiksvriendelijk, omdat de gebruiker zelf de voorkeursterm moet selecteren.

Andere zoekterm?

☐  Griekse vaas-schilderstijlen [Art & Architecture Thesaurus]	3
☐  Portland vaas [Uniek concept]	1
☐  VAAS = vazen [Art & Architecture Thesaurus]	50
☐  Vaassen [Plaats]	2
☐  Saint-Vaast-la-Hougue [Plaats]	1
☐  BORD [VAATWERK] = borden [vaatwerk] [Art & Architecture Thesaurus]	7

Afb. 13.3 - Eenvoudig zoeken.

Wie doorklikt op de voorkeursterm 'vazen', komt in de collectie terecht. Het volgende scherm laat een reeks werken zien waarvan de eerste drie hieronder zijn afgedrukt.

- | | | | |
|--------------------------|---|---|---|
| ☐ |  Artikel |  | Achilles and Polyxene : on a hydria in Petrograd
Beazley, J.D.
RUB |
| ☐ |  Artikel |  | L'agguato a Troilo nella ceramica Laconica
Montuoro, Paola Zancani
RUB |
| ☐ |  Boek |  | AI WeiWei
Smith, Karen; van der Zijpp, Sue-An
Rotterdam : NAI, 2008
MID |

Afb. 13.4 - Verkorte beschrijving van documenten.

Wanneer we het document in detail opvragen wordt de koppeling met de thesaurus ook zichtbaar: bij onderwerp staat immers (oa) de term 'vazen' genoemd, via welke term we gezocht hebben.

Record

Identificatie c:lvd:14159959
 Bookmark <http://anet.be/record/rub/opacrub/c:lvd:14159959/N>

Publicatie

Titel **Ai Weiwei**
 Auteur Smith, Karen
 van der Zijpp, Sue-An
 ISBN ISBN 978-90-5662-252-7
 Uitgave Rotterdam : NAI, 2008
 Omvang 95 p., afb.
 Taal Nederlands

**Onderwerp**

Art & Architecture Thesaurus installaties [beeldmateriaal]
 kunst
 vazen
 beeldhouwen
 Geografische termen China
 Historische perioden anno 1900-1999
 anno 2000-2009
 Onderwerpen (UDC) Iconografie
 Kunst
 Beeldhouwkunst
 Persoonsnaam Ai Weiwei

Beschikbaarheid

Bibliotheken Documentatiecentrum Middelheimmuseum AI 2008/575 Kunstenaarsmonografieën
 Raadpleging ter plaatse en na aanvraag

Afb. 13.5 - Detailpresentatie.

Het kan ook zijn dat we de term niet goed kennen en liever willen navigeren door de termen. Dan kiezen we bijvoorbeeld achtereenvolgens voor 'objecten', 'gebruiksartikelen' en 'uitrusting', 'houders' en 'houders naar vorm' en tenslotte 'vaatwerk'. Dan zien we een lijst met allerlei vormen (zoals ketels en putemmers) en helemaal onderaan vinden we onze vazen weer terug. Klikkend op vazen vinden we de term terug zoals hij in de thesaurus te vinden is (afbeelding 13.6). Er vallen meer specifieke begrippen (*narrower terms*) onder zoals 'alhambravazen' en 'grafvazen', maar ook de verwante begrippen (*related terms*) zoals 'urnen'.

Als we willen zoeken op vazen omdat dat de juiste term is, kunnen bijbehorende objecten getoond worden door te klikken op 'werken met betrekking tot vazen'. Dit scherm vormt de koppeling tussen thesaurus en collectie, want doorklikken op deze link, laat weer dezelfde reeks werken zien die we ook al vonden bij het eenvoudig zoeken (zie afbeelding 13.4 en 13.5).

vazen

AAT

Houders om bloemen in te zetten of voor de sier; van verschillende vorm en grootte, meestal hoger dan breed, veelal cilindrisch.

Zoeken in catalogus naar

- [Werken met betrekking tot vazen](#)
- [Werken met betrekking tot vazen en meer specifieke begrippen](#)

Zoekmogelijkheden in andere databanken

- [Art & Architecture Thesaurus](#)

Meer uitgebreide begrippen

- [vaatwerk](#)

Verwante begrippen

- [urnen](#)
- [vaasvormig](#)

Meer specifieke begrippen

- [Alhambravazen](#)
- [bouquetières](#)
- [florero's](#)
- [flower vases](#)
- [grafvazen](#)
- [hanaïke](#)
- [offering vases](#)
- [pijlenvazen](#)
- [plastische vazen \[antiek vaatwerk\]](#)
- [potpourrivazen](#)
- [vazen voor één bloem](#)

Afb. 13.6 - Thesaurusterm.

Generiek zoeken is ook mogelijk in deze thesaurus: als er niet genoeg gevonden zou worden bij vazen, zou men kunnen zoeken op 'werken met betrekking tot vazen en meer specifieke begrippen'. Omdat de *narrower terms* van 'vazen' niet toegekend blijken te zijn, bekijken we even op één niveau hoger wat het verschil is. Zoeken we op de term 'vaatwerk' via 'werken met betrekking tot vaatwerk', dan krijgen we 9 titels als resultaat. Zoeken we generiek op 'vaatwerk' via 'werken met betrekking tot vaatwerk en meer specifieke begrippen', dan is het resultaat opeens 203 titels.

Ook via het eenvoudig zoeken blijkt deze optie te bestaan: wie klinkt op de *i* vóór de term, krijgt meer informatie over de term en kan in het nieuwe scherm kiezen voor 'werken met betrekking tot vaatwerk en meer specifieke begrippen'.

- ☐ [vaatwerk \[Art & Architecture Thesaurus\]](#)
- ☐ [Extra informatie/zoekmogelijkheden \[Art & Architecture Thesaurus\]](#)
- ☐ [VAB](#)
= [Vlaamse Automobilistenbond \[Instelling\]](#)

Afb. 13.7 - Meer mogelijkheden.

13.2 MeSH thesaurus (Medical Subject Headings)

De MeSH (Medical Subject Headings) is een wereldwijd veel gebruikte thesaurus op het brede terrein van de geneeskunde. Hij wordt onderhouden door de Amerikaanse National Library of Medicine en bevat ruim 27.000 voorkeurstermen. Het is een thesaurus met veel hiërarchische niveaus waardoor zeer diepe boomstructuren ontstaan. De thesaurus is volledig gekoppeld aan een classificatie (zie afbeelding 13.8).

Otorhinolaryngologic Diseases	C09
. Ciliary Motility Disorders	C09.150
. . Kartagener Syndrome	C09.150.531
. Ear Diseases	C09.218
. . Cholesteatoma, Middle Ear	C09.218.200
. . Ear Deformities, Acquired	C09.218.271
. . Ear Neoplasms	C09.218.334
. . Earache	C09.218.350
. . Hearing Disorders	C09.218.458
. . . Hearing Loss	C09.218.458.341
. . . . Deafness	C09.218.458.341.186
. Deaf-Blind Disorders	C09.218.458.341.186.500
. Usher Syndromes	C09.218.458.341.186.500.500
. Wolfram Syndrome	C09.218.458.341.186.500.750
. . . . Hearing Loss, Bilateral	C09.218.458.341.374
. . . . Hearing Loss, Conductive	C09.218.458.341.562
. . . . Hearing Loss, Functional	C09.218.458.341.750
. . . . Hearing Loss, High-Frequency	C09.218.458.341.812
. . . . Hearing Loss, Mixed Conductive-Sensorineural	C09.218.458.341.849
. . . . Hearing Loss, Sensorineural	C09.218.458.341.887
. Hearing Loss, Central	C09.218.458.341.887.432
. Hearing Loss, Noise-Induced	C09.218.458.341.887.460
. Presbycusis	C09.218.458.341.887.772
. Usher Syndromes	C09.218.458.341.887.886
. . . . Hearing Loss, Sudden	C09.218.458.341.900
. . . . Hearing Loss, Unilateral	C09.218.458.341.950
. . Hyperacusis	C09.218.458.505
. . Tinnitus	C09.218.458.670
. . Herpes Zoster Oticus	C09.218.513
. . Labyrinth Diseases	C09.218.568
. . . Cochlear Diseases	C09.218.568.120
. . . Endolymphatic Hydrops	C09.218.568.217
. . . Meniere Disease	C09.218.568.217.500
. . . Labyrinthitis	C09.218.568.558
. . . Vestibular Diseases	C09.218.568.900
. . . . Vertigo	C09.218.568.900.883
. . Otitis	C09.218.705
...	...

Afb. 13.8 - Voorbeeld van de hiërarchie van de MESH-thesaurus, inclusief de koppeling met classificatiecodes.

Zoeken op getrunceerde classificatiecode kan dienen als alternatief voor generiek zoeken (Engels: *explode*) met thesaurustermen. Door deze

koppeling, ook voor de meest algemene categorieën van de classificatie, kunnen zodanig brede termen ontstaan dat deze in de praktijk niet worden toegekend bij het toegankelijk maken van medische literatuur.

De MeSH is een zogenaamde polyhiërarchische thesaurus, wat betekent dat een term in meer onderwerpsbomen kan voorkomen. Zo komt de term 'deafness' uit afbeelding 13.8 niet alleen voor in de boom 'otorhinolaryngologic diseases' (code C09.218.458.341.186), maar ook in die van 'nervous system diseases' onder code C10.597.751.418.341.186 en bij de 'pathological conditions' onder code C23.888.592.763.393.341.186.

Ten behoeve van de ontsluiting van publicaties, kunnen thesaurustermen worden gekoppeld met een beperkt aantal mogelijke subheadings die de context aangeven waarbinnen de term in een ontsloten publicatie voorkomt. Bijvoorbeeld: 'Aspirin - Adverse effects of Alzheimer disease – Diagnosis'. Op deze wijze kan een zekere mate van pre-coördinatie worden bewerkstelligd. In afbeelding 13.9 zijn de bij de term 'deafness' beschikbare subheadings zichtbaar.

De voornaamste toepassing van de MeSH thesaurus is het toegankelijk maken van medische literatuur in de bibliografische database Medline. Deze bevat ruim 25 miljoen publicaties. Daarnaast wordt de MeSH ook wel gebruikt voor andere toepassingen op medisch gebied. Van de database Medline bestaat een gratis via internet aangeboden versie onder de naam PubMed (www.ncbi.nlm.nih.gov/pubmed). Deze is gericht op de niet-informatiespecialist. Op deze site kan men ook de MeSH thesaurus verder exploreren (www.ncbi.nlm.nih.gov/mesh).

Belangrijke kenmerken van de combinatie PubMed-MeSH zijn:

- Er is een groot aantal ingangs- of verwijstermen in de thesaurus vastgelegd. Bij gebruik van zo'n woord als zoekterm wordt automatisch (ook) op de corresponderende thesaurusterm gezocht (zie afbeelding 13.10). Op deze wijze wordt automatisch vervangend zoeken gerealiseerd.
- Bij elke (al dan niet automatisch gegenereerde) thesaurusterm waarop wordt gezocht, vindt standaard een generieke zoekactie plaats. Daarbij worden alle lagere niveaus van de thesaurusterm meegenomen. Wie dit niet zou willen, moet tamelijk diep in het systeem deze automatische *explode* uitzetten (zie afbeelding 13.9).

Deafness

A general term for the complete loss of the ability to hear from both ears.

Year introduced: DEAF-MUTISM was heading 1963-1979

PubMed search builder options

[Subheadings:](#)

- | | | |
|--|--|--|
| ☐ analysis | ☐ epidemiology | ☐ physiopathology |
| ☐ anatomy and histology | ☐ ethnology | ☐ prevention and control |
| ☐ blood | ☐ etiology | ☐ psychology |
| ☐ cerebrospinal fluid | ☐ genetics | ☐ radiography |
| ☐ chemically induced | ☐ history | ☐ radionuclide imaging |
| ☐ classification | ☐ immunology | ☐ radiotherapy |
| ☐ complications | ☐ instrumentation | ☐ rehabilitation |
| ☐ congenital | ☐ legislation and jurisprudence | ☐ statistics and numerical data |
| ☐ cytology | ☐ metabolism | ☐ surgery |
| ☐ diagnosis | ☐ microbiology | ☐ therapy |
| ☐ drug therapy | ☐ mortality | ☐ ultrasonography |
| ☐ economics | ☐ nursing | ☐ urine |
| ☐ education | ☐ organization and administration | ☐ veterinary |
| ☐ embryology | ☐ pathology | ☐ virology |
| ☐ enzymology | ☐ physiology | |

☐ Restrict to MeSH Major Topic.

☐ Do not include MeSH terms found below this term in the MeSH hierarchy.

Tree Number(s): C09.218.458.341.186, C10.597.751.418.341.186, C23.888.592.763.393.341.186

MeSH Unique ID: D003638

Entry Terms:

- Bilateral Deafness
- Deafness, Bilateral
- Hearing Loss, Complete
- Complete Hearing Loss
- Prelingual Deafness
- Deafness, Prelingual
- Deaf-Mutism
- Deaf Mutism
- Deafness, Acquired
- Acquired Deafness
- Hearing Loss, Extreme
- Extreme Hearing Loss

See Also:

- [Hearing Aids](#)
- [Lipreading](#)

[All MeSH Categories](#)

[Diseases Category](#)

[Otorhinolaryngologic Diseases](#)

[Ear Diseases](#)

[Hearing Disorders](#)

[Hearing Loss](#)

Deafness

[Deaf-Blind Disorders](#)

[Usher Syndromes](#)

[Wolfram Syndrome](#)

[All MeSH Categories](#)

[Diseases Category](#)

[Nervous System Diseases](#)

[Neurologic Manifestations](#)

[Sensation Disorders](#)

[Hearing Disorders](#)

[Hearing Loss](#)

Deafness

[Deaf-Blind Disorders](#)

[Usher Syndromes](#)

[Wolfram Syndrome](#)

[All MeSH Categories](#)

[Diseases Category](#)

[Pathological Conditions, Signs and Symptoms](#)

[Signs and Symptoms](#)

[Neurologic Manifestations](#)

[Sensation Disorders](#)

[Hearing Disorders](#)

[Hearing Loss](#)

Deafness

[Deaf-Blind Disorders](#)

Afb. 13.9 - Voorbeeld van de (poly)hiërarchie van een thesaurusterm op de PubMed-site. Getoond worden ook de subheadings die bij deze term van toepassing zijn. Het automatisch generiek zoeken (explode) in PubMed kan in dit scherm - net onder de subheadings - middels een vinkje worden "uitgezet".

Search Details

Query Translation:

```
"rhinitis, allergic, seasonal"[MeSH Terms] OR ("rhinitis"[All Fields] AND "allergic"[All Fields] AND "seasonal"[All Fields]) OR "seasonal allergic rhinitis"[All Fields] OR ("hay"[All Fields] AND "fever"[All Fields]) OR "hay fever"[All Fields]
```

Search

URL

Result:

[14687](#)

Afb. 13.10 - Voorbeeld van automatisch vervangend zoeken in PubMed; als zoekterm is 'hay fever' gebruikt; automatisch wordt ook op de corresponderende (gelede) thesaurusterm 'rhinitis, allergic, seasonal' gezocht.

13.3 CAB-thesaurus

De CAB-thesaurus wordt samengesteld door CAB International, voortgekomen uit het Britse Commonwealth Agricultural Bureau. De thesaurus bevat ongeveer 157.000 voorkeurs- en 126.000 verwijstermen voor algemene inhoudelijke concepten op het brede terrein van de biowetenschappen en wat beperkter op het terrein van de sociale wetenschappen. Daarnaast bevat de thesaurus geografische termen en bijna 230.000 termen voor organismen volgens de taxonomie van het planten- en dierenrijk. Op de website van CAB International kan men door de thesaurus bladeren (www.cabi.org/cabthesaurus/). Afbeelding 13.11 toont daar een voorbeeld van. Voor het verkrijgen van zoekresultaten is op die website echter een wachtwoord nodig.

CAB Thesaurus

Home

All terms A-Z

Browse by subject / category

Search Thesaurus :



Language

Search type

Results format

Terms per page

Match terms *

Term type *

* Available only to search type "term begins with text"

[Help with search >>](#)

All terms A-Z

[All](#) | [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#) |

[<< Back](#) | [Show Relationships](#)

Influenzavirus A add

```

: : : : : organisms
: : : : : organisms
: : : : : pathogens
: : : : : viruses
: : : : : RNA Viruses
: : : : : ssRNA Viruses
: : : negative-sense ssRNA Viruses
: : Orthomyxoviridae
: influenza viruses
: : : : : organisms
: : : : : organisms
: : : : : pathogens
: : : : : viruses
: : : : : RNA Viruses
: : : : : ssRNA Viruses
: : : negative-sense ssRNA Viruses
: : Orthomyxoviridae
: Influenzavirus A
: . Influenza A virus
: . . avian influenza A viruses
: . . equine influenza A viruses
: . . swine influenza A viruses
  
```

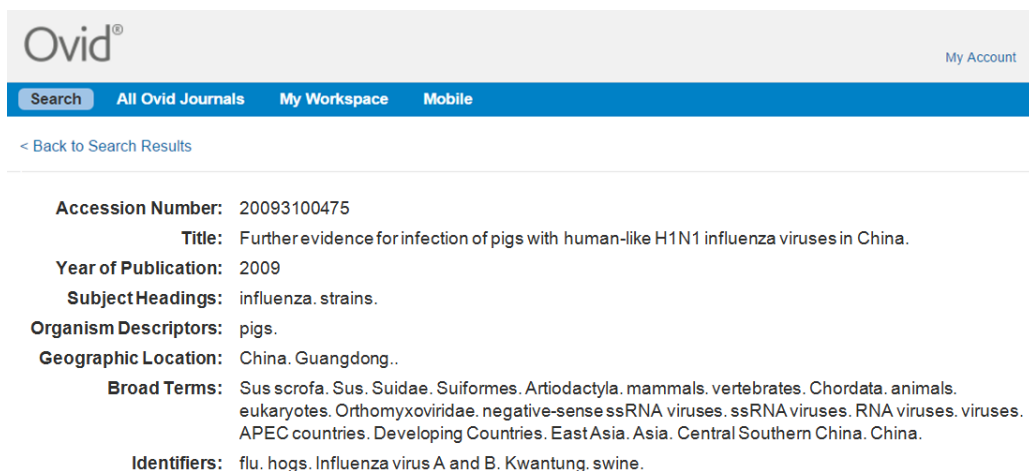
Afb. 13.11 - Hiërarchie van de CAB-thesaurus rond de term 'influenzavirus A'.

De thesaurus wordt gebruikt voor het toegankelijk maken van de bibliografische database CAB Abstracts, waarop via diverse aanbieders een licentie te nemen is. In deze database worden aparte velden gebruikt voor:

- de algemene thesaurustermen (descriptoren/*subject headings*),
- de geografische thesaurustermen,
- de taxonomische thesaurustermen (organismen).

Opmerkelijk is daarnaast nog een apart veld *broad terms* waarin niveau-indexering wordt toegepast. In dat veld staan overkoepelende (bredere) termen, op meer niveau, van zowel toegekende algemene descriptoren, als van geografische termen, als van taxonomische termen. Een artikel over 'varkensgriep in de provincie Guangdong in China' zal dan de thesaurusterm 'influenza', de geografische termen 'China' en 'Guangdong' en de taxonomische term 'pigs' meekrijgen. En daarnaast onder meer de *broader terms* 'RNA viruses', 'viruses', 'East Asia', 'Asia', 'Developing

Countries', 'Suidae', 'mammals', 'vertebrates', 'animals' (zie afbeelding 13.12). Deze wijze van indexeren maakt het mogelijk ook generiek te zoeken met behulp van zoeksystemen waar die functionaliteit op software-niveau ontbreekt. Wie op de *broader term* 'East Asia' zoekt zal bijvoorbeeld ook artikelen over (alleen) China, Korea of Japan vinden. Wie op de (geografische) thesaurusterm 'East Asia' zoekt, zal alleen artikelen over Oost-Azië als geheel vinden.



Ovid® My Account

[Search](#) [All Ovid Journals](#) [My Workspace](#) [Mobile](#)

[< Back to Search Results](#)

Accession Number: 20093100475

Title: Further evidence for infection of pigs with human-like H1N1 influenza viruses in China.

Year of Publication: 2009

Subject Headings: influenza. strains.

Organism Descriptors: pigs.

Geographic Location: China. Guangdong..

Broad Terms: Sus scrofa. Sus. Suidae. Suiformes. Artiodactyla. mammals. vertebrates. Chordata. animals. eukaryotes. Orthomyxoviridae. negative-sense ssRNA viruses. ssRNA viruses. RNA viruses. viruses. APEC countries. Developing Countries. East Asia. Asia. Central Southern China. China.

Identifiers: flu. hogs. Influenza virus A and B. Kwantung. swine.

Afb. 13.12 - Voorbeeld van (een deel van) een beschrijving uit de database CAB Abstracts, dat de verschillende in de tekst genoemde typen indexering toont. Identifiers zijn vrije trefwoorden; dus niet uit de thesaurus afkomstig.

13.4 Thesaurus Zorg en Welzijn en het gebruik binnen kanker.nl

De Thesaurus Zorg en Welzijn (TZW) bevat de terminologie van het sociaal domein (welzijn, onderwijs, opvoeding) en de zorg. Het product is 25 jaar geleden ontstaan bij het toenmalige NIZW (Nederlands Instituut voor Zorg en Welzijn) en na een tussenstop bij de kennisinstituten Movisie, Nederlands Jeugdinstituut en Vilans is het nu Stichting Stimulansz die de TZW doorontwikkelt. Inmiddels bevat de TZW ruim 35.000 termen. De thesaurus onderscheidt voorkeurs- en verwijstermen en legt hiërarchische en associatieve relaties tussen de voorkeurstermen vast. Een groot gedeelte van de thesaurustermen is voorzien van *scope notes* (definities). Door middel van specifieke coderingen kunnen deelselecties en afzonderlijke lijsten gemaakt worden.

De TZW wordt op dit moment door twee medewerkers onderhouden. Het onderhoud dat zij verrichten, bestaat uit het toevoegen van nieuwe

termen en het wijzigen van bestaande termen. Bij nieuwe termen kan het gaan om nieuwe ontwikkelingen, maar ook om het uitdiepen van een bepaald onderwerp of om aanvullingen op verzoek van klanten. Wijzigingen kunnen bestaan uit veranderingen in schrijfwijze, veranderingen in voorkeursterm en synoniem, verbetering of toevoeging van relaties tussen termen en aanvulling of verbetering van *scope notes*. De thesaurus wordt afgenomen door verschillende partijen uit het sociale domein en de zorg, waaronder GGD'en, MEE-organisaties, gemeenten en ziekenhuizen.

Klanten die de TZW, of een (op maat gemaakte) deelselectie, willen gebruiken in één of meerdere systemen, zoals een DMS, CMS en/of zoekmachine, sluiten een licentieovereenkomst af en ontvangen zes keer per jaar een update. Zij ontvangen het thesaurusbestand dan in het door hen gewenste exportformaat, meestal XML of SKOS-RDF. De updates bestaan uit een (volledig) nieuw thesaurusbestand en een overzicht van nieuwe termen en van wijzigingen die invloed hebben op trefwoordtoekenning. Door middel van een login kan de klant de volledige TZW raadplegen via de website. Voor overige belangstellenden is een beperkt gedeelte van de TZW ter inzage beschikbaar als voorbeeld. Daarnaast bestaat de mogelijkheid om een inlogabonnement op de TZW af te nemen, puur ter raadpleging van de thesaurus als woordenboek.

Stimulansz biedt ook de Geografische Thesaurus Zorg en Welzijn aan. Deze is ontwikkeld om organisatie- en adresgegevens en andere informatiebronnen op geografische kenmerken te ontsluiten. De Geografische Thesaurus bevat de namen van Nederlandse gemeenten, bijbehorende plaatsen en regionale indelingen. Er is direct te zien onder welke regio's een gemeente valt, bijvoorbeeld onder welke jeugdzorgregio, veiligheidsregio of zorgkantorregio. De Geografische Thesaurus is gratis te raadplegen via

<http://www.thesauruszorgenwelzijn.nl/geothesaurus/index.htm>

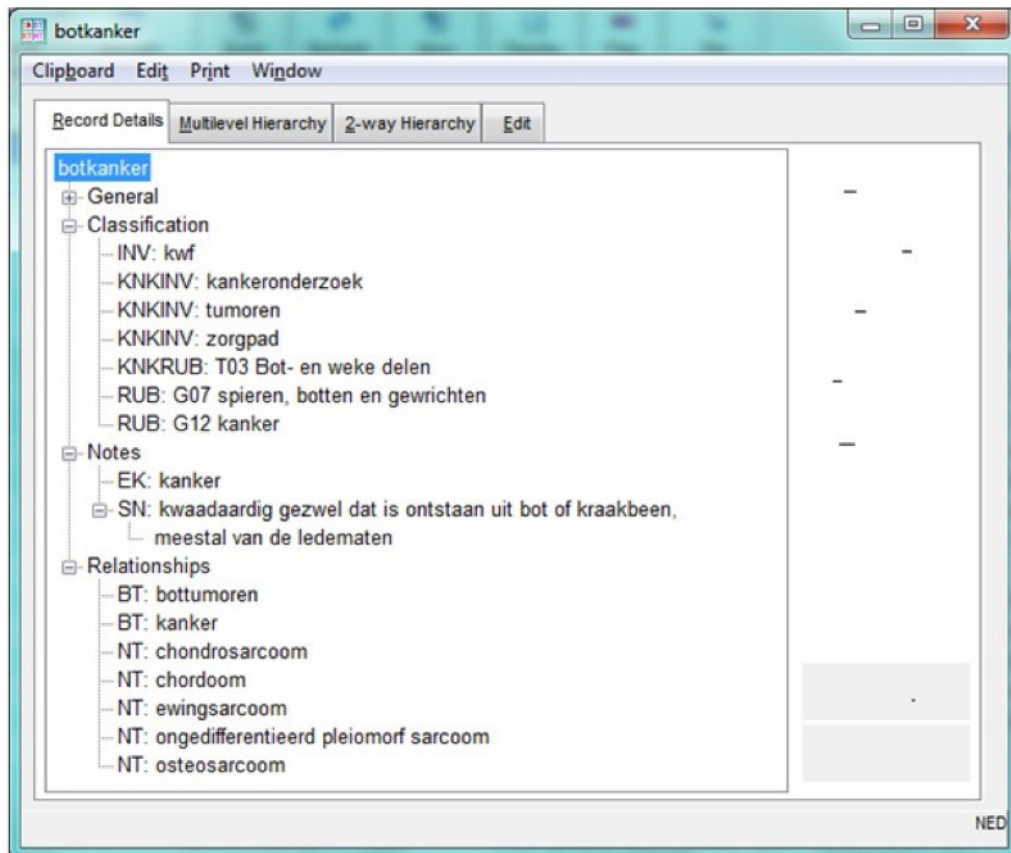
Er zijn plannen om de geografische thesaurus beschikbaar te stellen als skos/rdf. Daarmee is het dan beschikbaar als bestand, zodat het niet alleen ter inzage te gebruiken is.

De TZW wordt onderhouden in het thesaurusprogramma Multites. Dit programma voldoet aan de eisen die we aan thesaurussoftware stellen, zoals automatische foutcontroles, relaties automatisch wederkerig maken, etc.

De belangrijkste toepassingsmogelijkheden van de TZW zijn:

- eenduidige inhoudelijke metadatering (toekenning van trefwoorden/tags) bijv. in contentmanagementsystemen, projectdatabanken, sociale kaarten;

- verrijking van zoekmachines voor zoekondersteuning: inzetten van synoniemen voor vervangend zoeken, verdere zoeksuggesties op basis van de overige relaties;
- woordenboekfunctie, begrippenlijsten.



Afb. 13.13 - Een thesaurusterm in Multites.

Thesaurus Zorg en Welzijn binnen kanker.nl

De Thesaurus Zorg en Welzijn wordt onder andere ingezet bij kanker.nl. Kanker.nl is een digitaal platform over kanker waar (ex-)patiënten (maar ook naasten, hulpverleners en andere geïnteresseerden) informatie op maat kunnen vinden, in contact kunnen komen met anderen die geraakt zijn door kanker en ervaringen en kennis kunnen uitwisselen. Om informatie zo optimaal mogelijk vindbaar te maken en voor de eenheid van taal, gebruikt kanker.nl de thesaurus oncologie, een deelselectie uit de Thesaurus Zorg en Welzijn.

Voor kanker.nl zijn er binnen de TZW specifieke coderingen gedefinieerd die de mogelijkheden van de TZW vergroten. Voor kanker.nl zijn dit onder meer coderingen voor kankersoorten, behandelingen, gevolgen en onderzoeken. Daarnaast maakt kanker.nl gebruik van de hiërarchische, associatieve en gelijkwaardigheidsrelaties die per concept zijn vastgelegd. De patiëntvriendelijke term is hierbij de voorkeursterm met daaraan toegevoegd zo veel mogelijk naamsvarianten als synoniem. Daarnaast zijn er specifieke associatieve relaties aangemaakt voor de relatie tussen kankersoorten en resp. behandelingen, onderzoeken en gevolgen van kanker.

De informatie op het platform kanker.nl is onderverdeeld in vijf onderdelen: bibliotheek, discussiegroepen, personen, blogs en organisaties. Bij alle onderdelen van het platform wordt de thesaurus oncologie ingezet. Zoals in afbeeldingen 13.14 en 13.15 te zien is, gaat het hierbij om verschillende toepassingen die nu gerealiseerd zijn of die naar alle waarschijnlijkheid in de toekomst gerealiseerd zullen worden (de nummers corresponderen met de nummers in de afbeeldingen):

1. Zoeksuggesties: op basis van relaties die in de thesaurus zijn gelegd, krijgt de gebruiker in de toekomst na zijn zoekvraag suggesties voor andere zoekvragen.
2. Begripsaanduidingen: de *scope note* uit de thesaurus wordt gebruikt als definitieaanduiding van de zoekterm.
3. Vervangend zoeken: de zoekvraag van een gebruiker wordt automatisch uitgebreid met synoniemen uit de thesaurus om zo de opbrengst van zijn zoekvraag te vergroten.
4. Facetten: de verschillende coderingen uit de thesaurus (kankersoorten, behandelingen, onderzoeken en gevolgen) worden als facet aangeboden, zodat de gebruiker daar zijn zoekresultaten op kan verfijnen.
5. Automatisch relateren: op basis van overeenkomstige tekst, waaronder gelijke thesaurustermen, wordt gerelateerde informatie getoond.
6. Dynamische navigatie: op basis van de relaties die in de thesaurus zijn gelegd tussen o.a. kankersoorten en gevolgen, wordt in de toekomst de navigatie dynamisch opgebouwd (basaalcelcarcinoom toont bijvoorbeeld minder gevolgen dan leukemie).
7. Personalisering: bij de registratie als deelnemer wordt de thesaurus gebruikt voor het invullen van diverse onderdelen (o.a. kankersoort, behandelingen, onderzoeken, gevolgen) en op basis hiervan wordt informatie op maat aangeboden en kan gefilterd worden op lotgenoten.
8. *Federated search*: op basis van kankersoorten worden wetenschappelijke onderzoeken (*trials*) uit een externe database van het IKNL getoond (niet in schermafbeelding).

The screenshot shows the kanker.nl website interface. At the top, there's a navigation bar with links like 'Inloggen', 'Registreren', and a search bar containing 'borstkanker'. Below this, there are tabs for 'Bibliotheek', 'Discussiegroepen', 'Personen', 'Blogs', and 'Organisaties'. The main content area shows search results for 'borstkanker'. On the left, there's a sidebar with filters under 'gevonden in' (blogs, discussiegroepen, bibliotheek) and 'behandelingen' (chemotherapie, bestraling, etc.). The main results area includes a dropdown menu for 'borst' with suggestions like 'borstkanker', 'borsten', and 'borstamputatie'. There are three red boxes highlighting specific features: '1. Zoeksuggesties' points to the dropdown menu; '2. Begripsaanduidingen' points to the definition of 'borstkanker'; and '3. Vervangend zoeken' points to the 'Er is ook gezocht op:' section. A fourth red box labeled '4. Facetten' points to the left sidebar filters.

1. Zoeksuggesties

2. Begripsaanduidingen

3. Vervangend zoeken

4. Facetten

Afb. 13.14 - Resultatenpagina kanker.nl.

Binnen kanker.nl maakt een redactie nieuwe informatiepagina's aan en zij houdt de teksten up to date. Zij kent ook thesaurustermen aan de informatie toe. De informatiespecialist controleert deze. Zijn er nieuwe termen nodig, of moeten thesaurustermen aangepast worden, dan bekijkt de informatiespecialist van kanker.nl wat de consequenties van deze wijzigingen zijn voor álle onderdelen waarbij de thesaurus betrokken is. Na overleg met de thesaurusredactie worden vervolgens de wijzigingen in de Thesaurus Zorg en Welzijn doorgevoerd. De thesaurus wordt ook ingezet bij het automatisch classificeren van *user generated content* zoals blogs en discussiegroepen.

kanker.nl Inloggen Registreren Zoeken

Bibliotheek Informatie op maat | Discussiegroepen Vragen & antwoorden | Personen Vind mensen zoals jij | Blogs Persoonlijke verhalen

7. Personalisering

Basaalcelcarcinoom (huidkanker)

Toon colofon

Het basaalcelcarcinoom (BCC) is een vorm van huidkanker. Het ontstaat uit de cellen van de opperhuid. Ongeveer 80% van alle huidkankerpatiënten heeft een basaalcelcarcinoom.

Het basaalcelcarcinoom kan overal ontstaan, maar komt het meest voor in het gezicht en op het (kale) hoofd.

Ongeveer 80% van de mensen met huidkanker heeft een basaalcelcarcinoom. Andere vormen van huidkanker zijn

- plaveiselcelcarcinoom (ongeveer 15%)
- melanoom (ongeveer 5%)
- zeldzame tumoren van de huid, waaronder het merkelcelcarcinoom (minder dan 1%)

Groeiwijze

Basaalcelcarcinoom is de minst kwaadaardige vorm van huidkanker. Het groeit langzaam en zaait bijna nooit uit. Toch is het belangrijk ook een basaalcelcarcinoom grondig te behandelen. Wordt er niets aan gedaan, dan kan het dieper de huid ingroeien. Het kan dan de weefsels onder de huid aantasten, zoals kraakbeen en botweefsel. Of doorgroeien naar andere belangrijke weefsels zoals de ogen. Als het basaalcelcarcinoom dieper is ingegroeid in andere weefsels en als het groter is, wordt de kans op genezing kleiner. De behandeling is dan ingrijpender en moeilijker.

Dieper de huid ingroeien komt vooral voor in de zogenoemde H-zone gezicht. Dit is rond de ogen, de neus, de bovenlip, de oren en de zijwangen.

Type tumor

Uw arts beoordeelt welk type tumor u heeft om de keuze van de behandeling te bepalen. Hij bekijkt:

- hoe de tumor groeit:
 - als een bolletje: compact, nodulair
 - dik als een olievlék: oppervlakkig
 - wortelstelsel van een boom of als een spinnetje: sprietiger
 - kleine bolletjes: micronodulair
- de grootte van de tumor
- of het een eerste huidtumor is of een huidtumor die is teruggekeerd

Hoog risico

6. Dynamische navigatie

Op slaan in mijn bibliotheek

3x opgeslagen

Print Delen

Start discussie of vraag

5. Automatisch relateren

Vragen? Mail of bel ons

Kanker.nl Infolijn

205 deelnemers hebben basaalcelcarcinoom ingevuld op hun profiel

3 openbare profielen

Word ook deelnemer en

Artikelen uit de bibliotheek

- Brochure Huidkanker (KWF)
- Wat is kanker
- Ontrafelen hoe huidkanker ontstaat - Magazine Kracht
- Brochure Kanker...en hoe moet het nu met mijn kinderen? (KWF)
- Kanker naast andere ziekten

Meer artikelen (7)

Afb. 13.15 - Artikelpagina kanker.nl.

kanker.nl Hester Activiteiten Berichten Zoeken

Bibliotheek | Discussiegroepen | Personen | Blogs | Organisaties

Mijn profiel

Over mij

Gebruikersnaam Hester Deze naam kun je niet wijzigen

Mijn foto of afbeelding Foto instellen

Mijn relatie tot kanker Overige familie en vrienden

Geboortedatum 15 januari 1994

Vul je geboortedatum in, of alleen je geboortjaar.

Geslacht Vrouw

6. Dynamische navigatie

Kankersoort en diagnose

Onderzoek

Behandeling

Gevolgen

Niet meer beter worden

Meer over mij

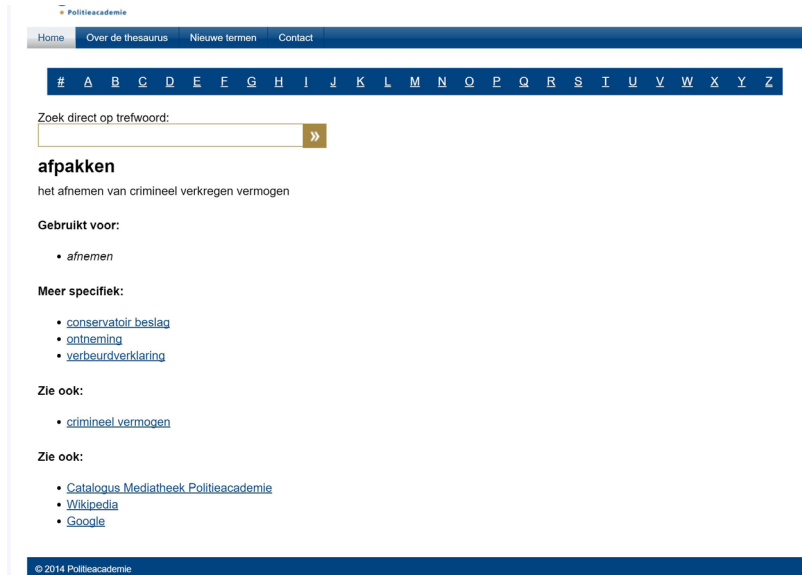
Afb. 13.16 - Deelnemersprofiel.

13.5 Thesaurus politiekunde

De thesaurus politiekunde van de Politieacademie (thesaurus.politieacademie.nl) is, zoals de beheerders het zelf aangeven, standaard voor trefwoordtoekenning binnen de politiewereld en bevat ruim 8000 termen. Deze thesaurus wordt onder andere gebruikt bij het op onderwerp doorzoeken van de catalogus van de mediatheek van de Politieacademie.

De Politieacademie is een van de weinige Nederlandstalige instellingen die hun mediatheekcollectie via het internet doorzoekbaar maken met behulp van een thesaurus. Er zijn weliswaar meer Nederlandstalige thesauri vindbaar op het internet, maar in de meeste gevallen is alleen de thesaurus raadpleegbaar en is het niet mogelijk om daarmee ook daadwerkelijk via het internet in een retrievalsysteem naar informatie te zoeken.

De thesaurus politiekunde is gericht op de politiewereld en dus een specialistische thesaurus. Dat houdt ook in dat er veel specialistische termen gebruikt worden en dat de betekenis van de termen door het vakgebied, ook zonder dat dit bijvoorbeeld door een *scope note* wordt toegelicht, is afgebakend. Het is interessant om te bekijken in hoeverre er bij het doorzoeken van de catalogus (<https://www.politieacademie.nl/kennisenonderzoek/kennis/mediatheek-catalogus>) gebruik wordt gemaakt van de speciale mogelijkheden die een thesaurus biedt, zoals het generiek zoeken en het vervangend zoeken. In afbeelding 13.17 staat de thesaurusterm 'afpakken' met zijn verwijzingen afgebeeld. Zoals te zien in deze afbeelding worden niet de standaard thesaurusverwijzingen NT, BT, RT, Use en UF gebruikt, maar hanteert de Politieacademie 'Meer specifiek', 'Gebruikt voor' en 'Zie ook'.



Afb. 13.17 - Een thesaurusterm met zijn verwijzingen uit de thesaurus van de Politieacademie (thesaurus.politieacademie.nl)

De gebruiker die in deze catalogus zoekt op 'afnemen' krijgt automatisch alle publicaties met de term 'afpakken'. Er wordt dus automatisch vervangend gezocht.

In de zoeksoftware is generiek zoeken niet mogelijk. Een zoekactie op de term 'afpakken' levert alleen maar documenten die geïndexeerd zijn met die term en niet met een van de daaronder liggende nauwere termen. Je kunt wel vanuit het overzicht, zoals dat in afbeelding 13.17 staat afgebeeld, zoeken op een van de onderliggende termen door er op te klikken.

Om er toch voor te zorgen dat informatie over specifiekere onderwerpen gevonden wordt bij het zoeken op een ruimere term, is er niveau-indexering toegepast. Dat wil zeggen dat aan een document meer termen uit een termenhiërarchie zijn toegekend.

In afbeelding 13.18 is het artikel "Afpakken werkt" zowel met de term 'afpakken' als met de nauwere termen 'verbeurdverklaring' en 'ontneming' geïndexeerd. Als generiek zoeken mogelijk zou zijn geweest dan zou de mediatheek van de Politieacademie kunnen volstaan met zo specifiek mogelijk indexereren. In dit geval had dan alleen de term 'afpakken' toegekend hoeven te worden. De redactie van de thesaurus heeft hier echter bewust niet voor gekozen.

Volledige beschrijving



Materiaalsoort



Titel Afpakken werkt

Auteur Henriette van Wermeskerken

Uitgave Amsterdam : Politieacademie, 2015

Bron Blauw, jrg. 11 (2015), nr. 2

Collatie p. 34-37

Link [Klik hier om deze titel in te zien \(alleen mogelijk voor politiemedewerkers met PKN account\)](#)
Om deze titel in te zien heeft u de Adobe Reader nodig. Deze is gratis te downloaden op <http://www.adobe.com>

Rubriek 2230 financiële opsporing

Trefwoord verbeurdeverklaring ; ontname ; inbeslagneming ; financiële opsporing ; afpakken

Samenvatting Binnen de eenheid Oost-Nederland heeft afpakken een hoge prioriteit. Herhaaldelijk afpakken samen met partners vormt hierin de rode draad.

Vindplaats DIGITALE MEDIATHEEK : Online beschikbaar

Niet uitleenbaar

Details
beschikbaarheid

Soortgelijke
werken

Afb. 13.18 - Beschrijving uit de catalogus van de Politieacademie, die het gebruik van niveau-indexering illustreert.

13.6 DDC, UDC en SISO

In een tijd waarin we informatie vooral toegankelijk maken met behulp van op tekst gebaseerde zoekmachines en metadata of ordening in de vorm van menu's en mappen, lijken classificaties met (alfa)numerieke codes nauwelijks meer relevant. Toch wordt er nog dagelijks door miljoenen mensen gebruik gemaakt van enkele belangrijke classificatieschema's omdat deze met name voor fysieke plaatsing van documenten zeer bruikbaar zijn. Bovendien zijn de ordeningsprincipes waarop classificaties berusten nog zeer bruikbaar in digitale omgevingen. Niet voor niets hebben we in hoofdstuk 6 dan ook aandacht besteed aan de principes van de traditionele classificatieleer. In deze paragraaf bespreken we een paar classificaties die in het dagelijks leven in veel informatie-omgevingen een rol spelen. Het belangrijkste en meest gebruikte classificatieschema van de wereld is de Dewey Decimale Classificatie (DDC). Deze classificatie wordt in vrijwel alle bibliotheken in de Verenigde Staten gebruikt en ook daarbuiten zijn er veel Angelsaksische bibliotheken die (een variant op) het systeem hanteren. De DDC bestaat al meer dan honderd jaar, maar ontwerper en naamgever Melville Dewey was niet de eerste die een classificatie ontwierp. De classificatie die hij in 1876 uitgaf was echter wel de eerste waarbij een overzichtelijk decimaal systeem werd toegepast. De classificatie kenmerkt zich verder door codes die minimaal drie cijfers bevatten. Bij verdere onderverdeling worden na drie cijfers punten geplaatst zodat de codes leesbaar en hanteerbaar blijven.

Oorspronkelijk was de DDC een zuiver enumeratieve classificatie, maar Dewey paste al snel een *ad on* toe op plaatsen waar een indeling naar vorm praktisch was. In de moderne versie van de DDC zijn verschillende hulptabellen toegevoegd zodat de classificatie een analytisch-synthetisch karakter heeft gekregen. De Library of Congress in Washington voegt DDC codes toe aan werken in haar collectie en aan werken waarvan bekend is dat ze gaan verschijnen (*cataloguing in publication*). Andere bibliotheken kunnen deze codes overnemen ten behoeve van hun eigen catalogus.

De redactie en verspreiding (via betaald abonnement) van de DDC is in handen van de internationale bibliotheekfirma OCLC. Zij biedt zowel een papieren als online versie van de classificatie aan. Voor bibliotheken met minder dan 20.000 items is er een *abridged version* die minder gedetailleerd is uitgewerkt. De DDC is een belangrijke inspiratiebron geweest voor de, in Europa meer gebruikte, Universele Decimale Classificatie (UDC).

De ontstaansgeschiedenis van de UDC is nauw verbonden met het streven van de heren Paul Otlet en Henri Lafontaine in de eerste helft van de 20e eeuw om alle vastgelegde kennis van de wereld te registreren in een kaartstelsel. Dit streven kwam voort uit idealistische motieven: door kennis te nemen van elkaars kennis en cultuur zou er meer begrip tussen de volkeren optreden zodat een wereldoorlog als die van 1914-1918 niet meer zou kunnen ontstaan. Om de vele kaartjes systematisch op onderwerp te kunnen ordenen, ontwikkelden ze hun classificatie, waarvoor ze Melvin Dewey om toestemming vroegen enkele van zijn principes (zoals het decimale stelsel) te mogen overnemen. Inmiddels is de classificatie verkrijgbaar in 40 talen en vooral nog in gebruik bij wetenschappelijke bibliotheken. De UDC is, net als de DDC, in feite een enumeratief schema met een sterk analytisch-synthetisch karakter. In de notaties komen combinaties van onderwerpsaspecten voor zoals '799.13 Vissen met netten in de binnenwateren'. Het sterk analytisch-synthetisch karakter komt naar voren door het gebruik van algemene hulptabellen, zoals die van plaats, tijd en vorm. De hulptabellen hebben eigen stukjes notatie, zoals aanhalingstekens en haakjes en kunnen overal gebruikt worden.

796.33	Voetbal
796.33(492)	Voetbal in Nederland.

Het analytisch-synthetisch karakter blijkt ook uit het toestaan van combinaties van hoofdgetallen en het voorkomen van bijzondere hulpgetallen. Deze hulpgetallen mogen alleen binnen een bepaalde context gebruikt worden. Zo kan het onderwerp 'Sportverenigingen' weergegeven worden door de rubrieken '796.06 (Algemene organisatie van de sport)' te koppelen met '061.23 (Verenigingen)', waardoor de code '796.06 : 061.23' ontstaat. Door deze, soms zeer complexe, combinaties

van codes is de UDC in tegenstelling tot andere classificaties uiteindelijk minder geschikt als plaatsingsschema. In de praktijk wordt het nogal eens gebruikt om de inhoud van een document zo specifiek mogelijk uit te drukken. In een digitale omgeving kan de UDC worden gebruikt om onderwerpen en concepten als *linked data* aan elkaar te verbinden.

Voor het gebruik in openbare bibliotheken is de UDC te uitgebreid en te complex. In Nederland gebruiken de grotere openbare bibliotheken en hogeschoolbibliotheken een eigen Nederlandse classificatie: Schema voor de Indeling van de Systemetische catalogus in Openbare bibliotheken (SISO).

De SISO bestrijkt evenals de UDC alle wetenschapsgebieden, maar is daarin veel minder gedetailleerd. Dat blijkt ook uit het feit dat SISO codes in principe maximaal 5 cijfers bevatten. Ook is de SISO, vanuit pragmatisch oogpunt minder zuiver in de onderverdeling van wetenschapsgebieden. Het onderwerp 'Atheïsme', dat eigenlijk onder de filosofie hoort, is bijvoorbeeld ondergebracht bij 'Godsdienst' omdat men ervan uitgaat dat gebruikers het daar zullen zoeken. Behoudens enkele hulptabellen (zoals land), is het in de SISO niet mogelijk om codes met elkaar te verbinden, zoals bij de UDC het geval is. Ondanks het vereenvoudigde karakter van de SISO, wordt het voor kleinere openbare bibliotheken nog te complex gevonden. Daar wordt het alternatieve plaatsingsschema PIM (Presentatiesysteem Informatieve Media) gehanteerd dat werkt met pictogrammen in plaats van codes. Naast grotere openbare bibliotheken gebruiken ook bibliotheken van Hogescholen de SISO of varianten daarop.

13.7 Iconclass

De Iconclass is een classificatiesysteem voor beschrijving van beeldmateriaal dat wereldwijd vooral voor kunsttoepassingen wordt gebruikt. Het is in Nederland ontwikkeld op basis van pionierswerk van de Leidse kunsthistoricus professor Henri van de Waal (1910-1972) en vanaf het begin van de jaren '70 van de vorige eeuw gepubliceerd. Op dit moment is het in beheer bij het Rijksbureau voor Kunsthistorische Documentatie (RKD). Op internet kan door de classificatie gebrowseed worden (www.iconclass.org/help/outline). Vervolgpagina's bieden ook een zoekfunctie om te vinden in welke rubriek(en) bepaalde onderwerpen zijn ondergebracht. Dat is zeker niet overbodig, want voor wie niet heel vertrouwd is met het systeem, is het vaak niet zonder meer duidelijk op welke plaatsen in de boomstructuur een bepaald onderwerp is ondergebracht. Meer inhoudelijke informatie over de wijze van opbouw van de Iconclass is te vinden op de Iconclass website (<http://www.iconclass.nl/>). Hoewel Iconclass formeel een classificatie heet, heeft het niet alle kenmerken die in dit boek aan classificaties worden toegeschreven. Zo kent het bijvoorbeeld maar in beperkte mate precoördinatieve aspecten.

De Iconclass gebruikt een gemengd decimaal-alfabetische notatie. Na twee cijfers volgt een letter en daarna weer een (per rubriek wisselend) aantal cijfers. Een voorbeeld uit de rubriek 'Bijbel' laat zien dat dat tot in groot detail kan gaan:

7	Bible
71	Old Testament
71H	story of David
71H7	David and Bathsheba (2 Samuel 11-12)
71H71	David, from the roof (or balcony) of his palace, sees Bathsheba bathing
71H713	Bathsheba receives a letter from David
71H7131	Bathsheba (alone) with David's letter

Eén van de 10 hoofd rubrieken herbergt specifiek abstracte begrippen. In afbeelding 13.19 is daarvan een voorbeeld te zien. Dit betekent dat de Iconclass ook voor ontsluiting op iconologisch niveau kan worden gebruikt (zie kader "Beeldmateriaal"). Een voorbeeld hoe dit wordt toegepast is te zien aan de geïllustreerde handschriften van de Koninklijke Bibliotheek die worden gevonden in de rubriek 'Geluk' (56B2), zoals getoond in afbeelding 13.20.

Allegorie op de vrede, matigheid en vrolijkheid.
Andriessen, Juriaan

0 Abstract, Non-representational Art

1 Religion and Magic

2 Nature

3 Human Being, Man in General

4 Society, Civilization, Culture

5 Abstract Ideas and Concepts

56 Emotion

568 Joy and Sorrow
abstract idea · allegory · emotion · idea · personification

568(+0) (+ variant)

568(+1) (+ personification)

568(+2) (+ allegorical scene, i.e. two or more personifications involved in an action)

568(+3) (+ symbolical representation of concept)

568(+4) (+ emblematical representation of concept)

568(+5) (+ 'exemplum', representation of exemplary (historical) event)

568(+6) (+ situation or event (especially in genre-painting) with symbolical connotation)

5681 Pleasure, Enjoyment, Joy; 'Allegrezza', 'Allegrezza da le medaglie', 'Allegrezza, letitia e giubilo', 'Diletto', 'Piacere', 'Piacere honesto' (Ripa)

5682 Happiness, Felicity

5683 Satisfaction, Contentment; 'Contento' (Ripa)

5684 Amusement

6 History

7 Bible

8 Literature

9 Classical Mythology and Ancient History

Afb. 13.19 - Iconclass kent ook rubrieken voor abstracte begrippen als emoties die door afgebeelde personen tot uiting worden gebracht.

Highlights... Browse... Search... **Medieval Illuminated Manuscripts** Introduction Background Home

Full reference manuscripts

(1-5 of 11) for (Iconclass = "56B2")

1. The Hague, KB, 76 E 6 fol. 119r



Divine Grace carries the Pilgrim to the other side in the ship of everlasting Bliss

Fol. 119r: column miniature

119r
50x55

iconclass
11L54134
11Q62
11R3
46C24(+63)
56B21

2. The Hague, KB, 78 D 4a fol. 13r



Wealth before Fortune's castle; Eur (Happiness) welcomes newcomers

Fol. 13r: column miniature

13r
110x100

iconclass

3. The Hague, KB, 78 D 1 fol. 16v



Fortune and her brothers Eur (Happiness) and Meseur

Fol. 16v: column miniature

16v
120x80

iconclass
11L61
12F211
31A2331
31A3191
31D15(+34)

Browse by subject

using the iconclass classification system

- Religion and Magic
- Nature
- Human Being, Man in General
- Society, Civilization, Culture
- Abstract Ideas and Concepts
 - 56 Emotion
 - 56B Joy and Sorrow [show images < 25](#)
 - 56B1 Pleasure, Enjoyment, Joy; 'Allegrezza', 'Allegrezza da le medaglie', 'Allegrezza, letitia e giubilo', 'Diletto', 'Piacere', 'Piacere honesto' (Ripa) [show images < 5](#)
 - 56B2 Happiness, Felicity [show images < 25](#)
 - 56BB Joy and Sorrow (opposing concepts) [show images < 25](#)
- History
- Bible
- Literature
- Classical Mythology and Ancient History

Happiness in an Hortus Conclusus; Fortune turning her wheel (background)	Dispute on the great number of pagan gods; God the Father, Virtue and Happiness (background)
Fol. 203v: column miniature	Fol. 208r: column miniature
203v 120x80	208r 120x80
iconclass 11L61 12F211 31A2331 31A3191 31D15(+34)	iconclass 11C23 11L532 11L61 12B13 31A25552(+932)

Afb. 13.20 - Een collectie Middeleeuwse geïllumineerde handschriften van de Koninklijke Bibliotheek is met behulp van Iconclass ontsloten (<http://www.kb.nl/manuscripts/browse>).

Sommige rubrieken van de classificatie worden verder onderverdeeld met woorden tussen haakjes. Afbeelding 13.21 toont daar een voorbeeld van uit een dierenrubriek. Daarnaast is er een onderverdeling mogelijk met zogenaamde *keys*, die wel wat lijken op hulptabellen. Ook daarvan is een voorbeeld te zien in afbeelding 13.21: de hulprubrieken (+0) tot en met (+9) rechtsboven. Het zijn geen algemene hulptabellen, aangezien de betekenissen van de hulpgetallen (de *keys*) per reeks codes kunnen verschillen. In bepaalde gevallen kunnen deze *keys* nog weer tot langere combinaties worden gecombineerd.

De H. Hieronymus
Key, Willem

0 Abstract, Non-representational Art
1 Religion and Magic
2 Nature
25 earth, world as celestial body
25F animals
25F2 mammals
25F23 beasts of prey, predatory animals

25F23(...) beasts of prey, predatory animals
(with NAME)
animal · earth · mammal · nature · predatory animal
· world

3 Human Being, Man in General
4 Society, Civilization, Culture
5 Abstract Ideas and Concepts
6 History
7 Bible
8 Literature
9 Classical Mythology and Ancient History

25F23(...) (+ variant)
25F23(...) (+1) (+ animals used symbolically)
25F23(...) (+2) (+ sex and age of animals; propagation of animals)
25F23(...) (+3) (+ anatomy of animals)
25F23(...) (+4) (+ animal behaviour)
25F23(...) (+5) (+ animal(s) in motion; positions, expressions of animals)
25F23(...) (+6) (+ disease and death of animal(s))
25F23(...) (+7) (+ animals ~ biological investigation)
25F23(...) (+8) (+ man and animal)
25F23(...) (+9) (+ products of animals)
25F23(BADGER) beasts of prey, predatory animals: badger
25F23(BEAR) beasts of prey, predatory animals: bear
25F23(CHEETAH) beasts of prey, predatory animals: cheetah
25F23(ERMINE) beasts of prey, predatory animals: ermine
25F23(FERRET) beasts of prey, predatory animals: ferret
25F23(FOX) beasts of prey, predatory animals: fox
25F23(HYENA) beasts of prey, predatory animals: hyena
25F23(ICHNEUMON) beasts of prey, predatory animals: ichneumon
25F23(LEOPARD) beasts of prey, predatory animals: leopard
25F23(LION) beasts of prey, predatory animals: lion
25F23(LYNX) beasts of prey, predatory animals: lynx
25F23(MANGOUSTE) beasts of prey, predatory animals: mangouste
25F23(MARTEN) beasts of prey, predatory animals: marten
25F23(OTTER) beasts of prey, predatory animals: otter
25F23(PANTHER) beasts of prey, predatory animals: panther
25F23(SABLE) beasts of prey, predatory animals: sable
25F23(TIGER) beasts of prey, predatory animals: tiger
25F23(WEASEL) beasts of prey, predatory animals: weasel
25F23(WOLF) beasts of prey, predatory animals: wolf
25F23(...) fabulous animals ~ predatory animals (with NAME)

Afb. 13.21 - In de Iconclass is ook ruimte voor verfijning van categorieën met behulp van woorden, zoals hier in de rubriek roofdieren.

Anders dan gebruikelijk bij de meeste classificaties, maakt de Iconclass maar in beperkte mate gebruik van precoördinatie. De onderverdeling

van rubrieken is voor het grootste deel gebaseerd op puur hiërarchische (generieke en geheel-deel) relaties. Wanneer toch aspecten uit verschillende homogene groepen in één rubriek gecombineerd worden, is dat omdat die specifieke combinatie een speciale betekenis heeft of in een bijzondere context voorkomt, zoals het eerdere voorbeeld van rubriek 71H7131 voor *Bathsheba (alone) with David's letter*. Dit heeft tot gevolg dat met Iconclass ontsloten materiaal vaak veel verschillende codes toegekend krijgt. De aan de items in afbeelding 13.20 toegekende codes illustreren dat. Deze karakteristiek van de Iconclass zien we uiteraard ook aan de rubrieken zelf. Zo bestaan er bijvoorbeeld codes waarmee

Medieval Illuminated Manuscripts Introduction Background Home

Browse by subject
using the iconclass classification system

- 1 Religion and Magic
- 2 Nature
- 3 Human Being, Man in General
 - 31 man in a general biological sense
 - 31A the (nude) human figure; 'Corpo humano' (Ripa)
 - [show images >25](#)
 - 31A2 anatomy (non-medical) [show images >25](#)
 - 31A25 postures and gestures of arms and hands
 - [show images >25](#)
 - 31A253 postures and gestures of arms and hands in relation to the head [show images >25](#)
 - 31A2536 hand(s) towards the mouth [show images < 25](#)
 - 4 Society, Civilization, Culture
 - 5 Abstract Ideas and Concepts
 - 6 History
 - 7 Bible
 - 8 Literature
 - 9 Classical Mythology and Ancient History

ICONCLASS RKD - Rijksbureau voor Kunsthistorische Documentatie
Subject browsing provided by Mnemosyne

Psalm 38 ('I said, I will take heed to my ways, that I sin not with my tongue')

David points to his mouth, his tongue stuck out, looking up to God

Fol. 256r: column miniature
256r
90x90

iconclass
31A2536
31A2552
31B6233
71H611
71X1(Ps. 38)1

Afb. 13.22 - Iconclass kent ook rubrieken voor het beschrijven van houdingen die mensen of dieren op afbeeldingen aannemen. Voorbeeld uit geïllumineerde handschriften van de Koninklijke Bibliotheek (<http://www.kb.nl/manuscripts/browse>).

je houdingen kunt beschrijven die mensen of dieren op afbeeldingen aannemen, los van hetgeen daarop verder te zien is. Afbeelding 13.22 laat daar een voorbeeld van zien. Zulke codes zullen dus vrijwel altijd in combinatie met andere woorden toegekend. Op basis van deze kenmerken moeten we misschien wel concluderen dat de Iconclass een classificatie is die sterke thesaurus-aspecten vertoont.

Naast de eerdergenoemde toepassing voor de geïllumineerde handschriften van de Koninklijke Bibliotheek, heeft ook onder meer het Rijksmuseum een experimentele IconClass browser voor de afbeeldingen

met metadata van de stukken uit hun collectie, die onder een CC0 Creative Commons licentie beschikbaar zijn. Deze browser kan uitgetest worden op <http://him.arkyves.org/RIJKSMUSEUM/>.

13.8 Automatisch klasseren bij het ministerie van Binnenlandse Zaken en Koninkrijksrelaties³

In hoofdstuk 10 hebben we een aantal technieken besproken voor automatische metadatering en classificering van documenten. Doc-Direkt, een shared service-organisatie van het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, probeert dergelijke technieken toe te passen bij haar werkzaamheden. Als centrale dienst binnen het Rijk voor de archiefselectie bepaalt de organisatie welke documenten (dossiers) permanent bewaard moeten worden en welke (en op welke termijn) vernietigd moeten worden. Naast de selectie van papieren bestanden wordt Doc-Direkt geconfronteerd met nieuwe uitdagingen: de overheid produceert niet alleen meer papieren dossiers, het creëert ook elektronische dossiers en bestanden. De traditionele, handmatige, methode om documenten te selecteren en te klasseren is hiervoor niet toepasbaar gezien het volume en de complexiteit.

De huidige selectiemethode gaat namelijk uit van 1054 officiële Selectiedocumenten⁴: lijsten van alle taken die een organisatie moet uitvoeren. Voor elke taak is een bewaartermijn of de noodzaak van een eeuwige opslag van alle gekoppelde bestanden bepaald. Er is echter een selectiedocument voor elk afzonderlijk beleidsterrein en elk document beschrijft het terrein aan de hand van honderden taken. Alleen door moeizaam en handmatig de papieren dossiers te doorgronden, kunnen deze nu worden gekoppeld aan de overeenkomstige taak.

In het geval van elektronische bestanden, zoals gescande papieren documenten, databases, file-shares en mailservers, is handmatig selecteren als proces niet haalbaar. Het gevolg hiervan is dat de grip op de digitale verzameling verloren gaat:

- Veel data is nu ongebruikt, onbeheerd, ongestructureerd en soms zelfs onbekend.
- Oude software moet in de lucht gehouden worden om een minimale toegang voor enkele medewerkers mogelijk te maken.

De data heeft hiermee grotendeels zijn informatiewaarde verloren.

³ Tekst grotendeels geschreven door Marcel van Vlaanderen, Doc-Direkt / Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

⁴ Deze selectielijsten zijn o.a. te vinden via: <http://www.handelingenbank.info/hb.php>

Waarde uit data

Om weer grip te krijgen op de data, deze beschikbaar te stellen en toch elektronische dossiers te kunnen selecteren en klasseren, ontwikkelt Doc-Direkt nieuwe methoden en systemen. Het resultaat is dat Doc-Direkt het Rijk:

- ontlast bij het beheren van digitale data;
- inzicht geeft in de waarde en risico's van de data;
- informatie (meerwaarde) kan genereren uit deze data door o.a. het uitvoeren van analyses en presenteren van patronen.

In 2015 voerde zij hiervoor een *Proof of Concept* uit, waaruit bleek dat een combinatie van instrumenten (tekstanalyse, extractie van gegevens, classificatie en ranking) beschikbaar en geschikt zijn om te gebruiken. In 2016 is dan ook een project gestart om dergelijke instrumenten te implementeren. De uitdaging is om de juiste combinatie van *search engine*, complexe queries, entiteit- en patroonherkenning, toevoeging van semantiek (toevoegen van externe bronnen), *text mining*, *machine learning* etc. te vinden. Dit gaat verder dan hetgeen al gebeurt met E-discovery. Het gaat immers niet alleen om het identificeren van relevante documenten bij een vraag binnen een vaste set, maar ook om het per object aanvullen van de metadata voor het creëren van context en classificatie voor gebruik en beheer. Het beoogde resultaat wordt gevormd door elektronische bestanden die voldoen aan de eisen van het Elektronische Depot van het Nationaal Archief en die geraadpleegd kunnen worden door wetenschappers en andere gebruikers⁵.

Als de techniek bruikbaar kan worden ingezet bij bestanden die enige tijd geleden zijn gecreëerd, dan heeft het ook de potentie om ingezet te worden binnen de primaire processen van organisaties. Bij het gebruik van documentmanagementsystemen/ zaakvolgsystemen is één van de kenmerken dat voor elk document bepaald moet worden bij welk dossier en/of welke zaak het behoort. Dat moet nu bepaald worden door eindgebruikers of ondersteunende medewerkers. Dit is niet alleen tijdrovend en foutgevoelig, ook is het lastig om een document tegelijk in meer dan één verzameling (zoals de werkvoorraad van een individuele medewerker en de archiefordening) op te nemen. De automatische metadatering kan documenten mogelijk al op het moment van creatie of ontvangst zelfstandig toevoegen aan die verzamelingen die de organisatie nodig heeft voor bedrijfsvoering, verantwoording en behoud van cultureel erfgoed.

Creatie van het nieuwe proces

Om dit te bewerkstellingen ontwikkelt Doc-Direkt een proces voor:

5 Zie o.a.: <http://www.gahetna.nl/collectie/archief>

- automatische metadata-extractie uit de documenten/bestanden (zoals datums, titels, onderwerpen, verzender/ontvanger, namen van organisaties);
- automatische metadata-verrijking door de koppeling met externe bestanden, databases en bronnen, o.a. als ondersteuning bij het maken van de selectiebeslissing (zoals het actorenregister, vertaaltabellen);
- het naar XML-bestanden wegschrijven van de toegevoegde metadata zodat deze door andere processen kunnen worden verwerkt.

Het einddoel is dat met de verkregen data de objecten beschreven kunnen worden volgens de richtlijnen uit het “Toepassingsprofiel Metagegevens Rijksoverheid”⁶. Op basis hiervan kunnen vervolgens specifieke metadataproducten van Doc-Direkt, het Nationaal Archief en andere organisaties geproduceerd worden. Een bijvangst van het produceren, beheren en ontsluiten van metadata is dat het tegelijkertijd door de gebruikte technieken ook mogelijk is om geavanceerd fulltext door de data te zoeken. Dat klinkt basaal, maar veel (oudere) digitale verzamelingen zijn nu niet of slechts zeer rudimentair doorzoekbaar en daardoor onbruikbaar.

Doc-Direkt wil deze oplossing, zoveel als mogelijk, realiseren met de relevante modules van de HPe IDOL-suite (*voorheen Autonomy*) zoals:

- *Full-text and Advanced Search (IDOL);*
- *Eduction;*
- *Automatic Query Guidance;*
- *Taxonomy Generation;*
- *Classification;*
- *Knowledge graph.*

Het realiseren van de gewenste eindproducten vraagt om de gecombineerde inzet van deze middelen in meerdere stappen. De modules worden ingezet voor:

- Identificatie van kerndocumenten in de verzameling op basis van tekstuele en vormkenmerken (nota, minute, memo, etc.);
- Extractie uit de documenten van herkenbare elementen zoals: data, indicatie van werkprocessen en de namen van organisaties, functies en personen;
- Toekennen van extra wegingsfactoren aan termen, gevonden in de kerndocumenten, aangescherpt door te werken met ankers en zonering binnen deze documenten (zo bevatten de eerste of de eerste twee pagina’s van een document veelal de kern);
- Vergelijken van herkende data met externe validatiebronnen zoals het Actorenregister en de selectie-instrumenten.

6 Zie: <http://www.nationaalarchief.nl/informatiebeheer-archiefvorming/metadata>

Automatiseren = handwerk

Het inrichten van selectieprocessen (het configureren van de software en regels) voor het juist identificeren van de juiste objecten kan niet zonder aanzienlijke menselijke inzet. Er moeten voorbeeldsets worden samengesteld van objecten die behoren bij de verschillende categorieën die later automatisch gevonden moeten worden. Op basis van deze sets kunnen vervolgens leersets voor zelflerende software of relevante termen voor opname in taxonomieën bepaald worden.

Deze, kostbare, menselijke bijdrage aan het instellen van het instrument maakt dat de techniek (voorlopig) alleen rendabel kan zijn indien er een grote hoeveelheid data verwerkt kan worden en/of het instrument lang ingezet kan worden.

Digitaal werken = veranderen

De uitdaging voor Doc-Direkt ligt maar deels in de techniek. Naast de al genoemde problematiek van het selectie-instrument ligt het deels ook in de hoeveelheid benodigde metadata-elementen en de te behalen kwaliteitseisen van deze metadata. Het blijkt dat, zoals in zoveel digitaliseringstrajecten, verwachtingen uit de papieren situatie niet 1:1 te realiseren zijn in de digitale situatie. Hierdoor ontstaat de noodzaak opnieuw kritisch te beoordelen waarom bepaalde metadata en kwaliteitscriteria worden gevraagd. Welk doel heeft die metadatawaarde en kan dat doel ook op een andere wijze behaald worden?

Een mooi en uitdagend voorbeeld van deze discussie is het meegeven van betekenisvolle beschrijvingen van de inhoud van objecten. In de papieren situatie wordt deze omschrijving taalkundig juist geformuleerd door een medewerker. Hoewel de analysesoftware goed in staat is om aan te geven welke onderwerpen het meest voorkomen, welke organisaties betrokken zijn en welke periode/ datum het betreft, is het nog niet in staat hiervan een samenhangend verhaal te maken. De vraag is dan: "Is dat erg gezien de doelgroepen?". Vergeet niet dat de enige ontsluiting in de oude (papieren) wereld, de metadata op archief- en dossierniveau was. In de digitale wereld kan iedereen ook door de volledige tekst van de informatie/documenten zoeken.

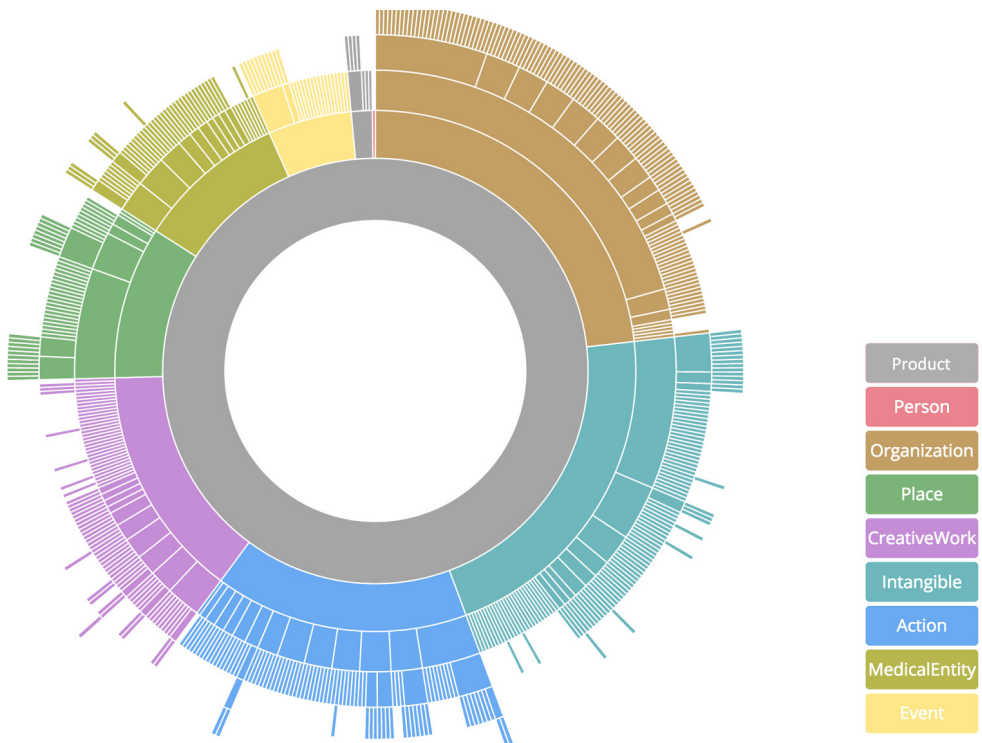
13.9 Schema.org

Schema.org is een metadataschema dat is ontwikkeld om gestructureerd betekenis toe te kennen aan relevante gegevens die in webpagina's gepresenteerd worden. Die betekenis wordt verwerkt in de codering van webpagina's en blijft daardoor onzichtbaar voor gewone bezoekers van websites. Computerprogramma's kunnen er echter wel gebruik van maken. Daarbij moet vooral gedacht worden aan zoekmachines die hierdoor de betekenis van gegevens in webpagina's kunnen herkennen.

Deze standaard wordt dan ook actief ondersteund en mee-ontwikkeld door de grote zoekmachines Google, Bing, Yahoo! en Yandex.

Schema.org bevat een algemeen kernvocabulaire, een collectie van ongeveer 600 zogenaamde “types”. Daarnaast zijn er extensies voor specifieke toepassingen. Die voegen nog eens ruim 150 termen (types) aan het vocabulaire toe. De gedefinieerde termen zijn in een hiërarchische structuur geordend. Die heeft als alles overkoepelende topterm ‘Thing’. De acht hoofdcategorieën uit het kernvocabulaire daaronder zijn: ‘Action’, ‘CreativeWork’, ‘Event’, ‘Intangible’, ‘Organization’, ‘Person’, ‘Place’ en ‘Product’.

Met de extensies wordt daar nog ‘MedicalEntity’ aan toegevoegd. Onder Intangible vallen onder meer begrippen als: ‘Audience’, ‘Brand’, ‘Order’, ‘Quantity’, ‘Rating’, ‘Service’ en ‘Ticket’.



Afb. 13.23 - Visualisatie van de aantallen begrippen die in Schema.org gedefinieerd zijn in de drie niveaus onder de negen hoofdrubrieken.



Afb. 13.24 - Een stukje van de boomstructuur van Schema.org rondom het begrip 'Restaurant'; overige hoofdcategorieën zijn ingeklapt.

Een visualisatie van de aantallen begrippen in de verschillende niveaus onder de hoofdcategorieën wordt getoond in afbeelding 13.23. Afbeelding 13.24 toont een grotendeels ingeklapt beeld van de hiërarchie, met 'Thing' als topterm. De volledige hiërarchie is te vinden op <https://schema.org/docs/full.html>.

Elk van de in Schema.org beschreven 'dingen' kan zelf weer allerlei eigenschappen hebben. Dergelijke eigenschappen kunnen ook geërfd worden van 'dingen' hoger in de hiërarchie. Afbeelding 13.25 geeft een voorbeeld hoe die eigenschappen op de Schema.org site gedefinieerd

Restaurant

Thing > Organization > LocalBusiness > FoodEstablishment > Restaurant

Thing > Place > LocalBusiness > FoodEstablishment > Restaurant

A restaurant.

Usage: Between 10,000 and 50,000 domains

[more...]

Property	Expected Type	Description
Properties from FoodEstablishment		
<u>acceptsReservations</u>	Text or URL or Boolean	Indicates whether a FoodEstablishment accepts reservations. Values can be Boolean, a URL at which reservations can be made or (for backwards compatibility) the strings Yes or No.
<u>menu</u>	Text or URL	Either the actual menu or a URL of the menu.
<u>servesCuisine</u>	Text	The cuisine of the restaurant.
Properties from LocalBusiness		
<u>branchCode</u>	Text	A short textual code (also called "store code") that uniquely identifies a place of business. The code is typically assigned by the parentOrganization and used in structured URLs. For example, in the URL http://www.starbucks.co.uk/store-locator/etc/detail/3047 the code "3047" is a branchCode for a particular branch.
<u>currenciesAccepted</u>	Text	The currency accepted (in ISO 4217 currency format).
<u>openingHours</u>	Text	The opening hours for a business. Opening hours can be specified as a weekly time range, starting with days, then times per day. Multiple days can be listed with commas ',' separating each day. Day or time ranges are specified using a hyphen '-'. - Days are specified using the following two-letter combinations: Mo, Tu, We, Th, Fr, Sa, Su. - Times are specified using 24:00 time. For example, 3pm is specified as 15:00. - Here is an example: <code><time itemprop="openingHours" datetime="Tu,Th 16:00-20:00">Tuesdays and Thursdays 4-8pm</time></code> . - If a business is open 7 days a week, then it can be specified as <code><time itemprop="openingHours" datetime="Mo-Su">Monday through Sunday, all day</time></code> .
<u>paymentAccepted</u>	Text	Cash, credit card, etc.
<u>priceRange</u>	Text	The price range of the business, for example \$\$.
Properties from Place		

Afb. 13.25 - Gedeelte van de definities van de eigenschappen van Restaurant uit Schema.org.

worden. Daar zie je dat 'Restaurant' onder meer de eigenschappen 'menu' en 'servesCuisine' erft van het algemenere 'FoodEstablishment' en de eigenschappen 'openingHours' en 'priceRange' van het nog algemenere 'LocalBusiness'. In het voorbeeld in afbeelding 13.26 zie je hoe dat in de code van een echte webpagina is toegepast.

Veel begrippen en eigenschappen uit andere schema's of ontologieën zijn in de loop van de tijd in Schema.org opgenomen. Een voorbeeld daarvan is de Good Relations ontologie (<http://www.heppnetz.de/ontologies/goodrelations/v1.html>), die was ontwikkeld voor e-commerce toepassingen op internet.

Een voorbeeld van de toepassing van Schema.org, is de manier waarop restaurantbeoordelingssite IENS daar gebruik van maakt. In de broncode van de webpagina met beoordelingsgegevens van een Amsterdamse Pizzeria komen onder meer de drie fragmenten voor die in afbeelding 13.26 getoond worden. Daarbij wordt telkens via een URL verwezen naar de gegevens van een bepaald soort 'ding' op de Schema.org site. Hier is dat respectievelijk Restaurant, AggregateRating en PostalAddress. De eigenschappen van die "dingen" worden telkens gekarakteriseerd via het attribuut *itemprop* bij een HTML-code. Sommige van die gegevens verschijnen ook in zoekresultaten bij Google, als zogenaamde *rich snippets*. In afbeelding 13.27 zien we daarvan priceRange, ratingValue, bestRating en ratingCount terug. Op deze gegevens kan niet gericht gezocht worden, maar ze zijn voor de zoeker wel nuttig om een keuze te maken welk resultaat aan te klikken. Ze dienen dus de evaluatiefunctie. Voordeel voor de aanbieder van de pagina is dat daardoor de kans groter is dat een gebruiker die pagina uit de resultatenlijst aanklikt. Daarmee is toepassing van Schema.org een onderdeel geworden van de search engine optimalisatie die op webpagina's wordt toegepast.

Zoals we zien worden lang niet alle volgens Schema.org gecodeerde gegevens als *rich snippets* in het zoekresultaat getoond. Welke dat zijn is een keuze van de betreffende zoekmachine. Wat niet in het zoekresultaat verschijnt, kan (en zal) door de zoekmachine overigens wel gebruikt worden om de te tonen resultaten zo goed mogelijk te laten aansluiten bij de veronderstelde intentie van de zoekvraag. Daarmee draagt de toepassing van Schema.org dus ook bij aan de in paragraaf 11.8 genoemde semantische zoektechnieken.

```

<section data-restaurant-id="1501" class="restaurantHeader restaurantHeader_hasPhotoSlider restaurantHeader_hasBookingWidget"
itemscope itemtype="http://schema.org/Restaurant">
  <div class="wrapper restaurantHeader_wrapper">
    <div class="restaurantHeader_header absolute">
      <div class="photoSlider" data-plugin="photoSlider">
        <span class="photoItem"></span>
      </div>
      <div class="small-show restaurantHeader_overlay hide-on-swipe">
        <h1 class="titlePrimary restaurantHeader_title" data-restaurant-name=YamYam Trattoria Pizzeria</h1>
        <ol class="labels restaurantHeader_labels fontSmall listStyle-none">
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#keuken" data-restaurant-navigation>Restaurant</a>
</li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#keuken" data-restaurant-navigation
itemprop="servesCuisine" content="Pizzeria">Pizzeria</a></li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/duurzaamheid#restaurantNav" data-restaurant-navigation>69%
Duurzaam</a></li>
          <li><a href="/restaurant/1501/amsterdam-yamyam-trattoria-pizzeria/details#prijzen" data-restaurant-navigation
title="Menuprijs" data-plugin="tipsy" itemprop="priceRange" content="22.50"><strong>€</strong> 22.50</a></li>
        </ol><ol class="restaurantHeader_grades scoreList listStyle-none">
          <ol class="restaurantHeader_grades scoreList listStyle-none" itemprop="aggregateRating" itemscope
itemtype="http://schema.org/AggregateRating">
            <li>
              <div class="scoreMedium scoreMediumBetter fontStandOut rating">
                <span class="average" itemprop="ratingValue">8.2</span>
                <span class="isInvisible" itemprop="bestRating">10</span>
                <span class="isInvisible" itemprop="worstRating">1</span>
                <span class="isInvisible" itemprop="ratingCount">34</span>
              </div>
              <p class="fontSmall">Eten</p>
            </li>
          </ol>
          <div class="restaurantWidget_locationBox medium-6 large-12 large-last small-hide" itemprop="address"
itemscope itemtype="http://schema.org/PostalAddress">
            <div class="restaurantWidget_location">
              <div class="mapCanvas" data-map data-marker="[52.376095,4.873566]" data-icon="restaurant" data-zoom="15"
data-mouse-zoom="false" data-re-centre="true"></div>
              <a class="restaurantWidget_locationLargeMap fontSmall" href="/restaurant/1501/amsterdam-yamyam-trattoria-
pizzeria/contact#kaart" data-restaurant-navigation>Grote kaart</a>
              <div class="restaurantWidget_locationText fontSmall">
                <div class="restaurantWidget_locationAddress">
                  <i class="icon mrs left icon-location icon-light"></i>
                  <address><span itemprop="streetAddress">Frederik Hendrikstraat 90</span>,
                    <span class="restaurantWidget_locationCity">
                      <span itemprop="postalCode">1052 HZ</span>,
                      <span itemprop="addressLocality">Amsterdam</span>
                    </span>
                  </address>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>

```

Afb. 13.26 - Fragmenten uit de HTML broncode van een webpagina van restaurantbeoordelingssite iens.nl, waarin Schema.org kenmerken volgens de microdata-standaard zijn toegevoegd.

In de webpagina in afbeelding 13.26 zijn de Schema.org metadata in HTML gecodeerd volgens de microdata standaard. De daar gebruikte attributen *itemtype*, *itemscope* en *itemprop* zijn uit die standaard afkomstig. Daarnaast worden ook wel twee andere standaarden toegepast om Schema.org metadata in webpagina's te coderen. Dat zijn RDFa en JSON-LD. RDFa (RDF in attributen) is gebaseerd op het in paragraaf 9.4 behandelde RDF. Op de Schema.org site worden bij de beschrijvingen van elk 'ding' voorbeelden gegeven hoe metadata volgens elk van deze drie methoden gecodeerd moeten worden. RDFa verschilt vrijwel alleen van microdata in de benaming van de in HTML te gebruiken attributen. JSON-LD

✓ Restaurant YamYam Trattoria Pizzeria in Amsterdam ... - iens.nl

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 8.2/10 - 34 votes - Price range: 22.50

Bij **Pizzeria** restaurant YamYam Trattoria **Pizzeria** in **Amsterdam** eten? ... Noord-Holland; >; **Amsterdam**; >; **West**; >; YamYam Trattoria **Pizzeria** Gemakkelijk een van de betere Italiaanse restaurants in **Amsterdam** vanuit dat oogpunt.

✓ Restaurant Forno in Amsterdam - Recensies | iens.nl - Waar gaan we ...

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 7.7/10 - 11 votes - Price range: 25.00

Word nooit teleurgesteld, een van de lekkerste **pizza** die er in **Amsterdam** te ie wordt opgediend door het mooiste blonde meisje van **Amsterdam-West** zou ...

✓ Restaurant Pazzi II in Amsterdam - Recensies | iens.nl - Waar gaan we ...

[www.iens.nl](#) > ... > Provincie Noord-Holland > Amsterdam > West ▼ [Translate this page](#)

★★★★★ Rating: 8.1/10 - 20 votes - Price range: 25.00

Een leuke **pizzeria** voor een snelle hap met de lekkerste pizzabodem tot nog toe! ... Dit is toch zeker wel de lekkerst **pizza** die je in **Amsterdam** kan eten!

Afb. 13.27 - Voorbeeld van een zoekresultaat uit Google waarin enkele van de in afbeelding 13.26 getoonde gegevens als rich snippets getoond worden.

(JavaScript Object Notation for Linked Data) is meer een programmeertaal waarin de metadata beschreven worden. Deze programmaatjes worden als scripts aan de broncode van webpagina's toegevoegd. Zoals de benaming al aangeeft is het gebruik hiervan er speciaal op gericht om de metadata (ook) voor *linked data* te kunnen gebruiken. In het kader van *linked data* wordt Schema.org markup ook toegepast door OCLC bij de webpresentatie van catalogusrecords in WorldCat.

Vooral in het kader van *search engine optimalisatie* (zie kader) wordt het voor aanbieders van websites steeds belangrijker om semantische codering, zoals die uit het Schema.org vocabulaire, in hun webpagina's te verwerken. Om die gegevens (vaak aangeduid als *structured data*) op eenvoudige wijze te kunnen controleren en analyseren bestaan intussen allerlei tools. Een veelgebruikt algemeen hulpmiddel hiervoor is Google's Structured Data Testing Tool (<https://search.google.com/structured-data/testing-tool/u/0/>). Daarin kun je het URL van een webpagina invoeren, waarna een rapportage verschijnt, zoals in afbeelding 13.28. Een overzicht van nog een groot aantal andere tools voor het genereren, analyseren en valideren van dit soort *structured data* is te vinden op

 https://www.iens.nl/restaurant/yamyam-trattoria-pizzeria/220297

NEW TEST



```

bediening zijn heel vriendelijk en oplettend. En
misschien nog wel belangrijker: het eten is uitzonderlijk!
De specialiteit is de pizza Yam Yam met truffelcreme,
beensham en rucola. Maar eigenlijk is alles van de kaart
verrassend lekker. Kom het snel proeven!", "telephone": "31-
206815097", "image":
["https://u.tfstatic.com/restaurant_photos/297/22029
7/169/612/yamyam-trattoria-pizzeria-restaurantzaal-
718a1.jpg"], "address":
{"@type": "PostalAddress", "streetAddress": "Frederik
Hendrikstraat 88-90", "postalCode": "1052
HZ", "addressLocality": "Amsterdam", "addressCountry": "Neder
land"}, "aggregateRating":
{"@type": "AggregateRating", "ratingValue": "8.1", "bestRating":
"10", "ratingCount": "70"}, "geo":
{"@type": "GeoCoordinates", "latitude": "52.376149", "longitude":
"4.873466", "review":
[{"@type": "Review", "datePublished": "2015-12-
17", "reviewBody": "ik had de pizza Yam Yam die heerlijk
was. De bediening was prima, heel aardig. De inrichting
is sober maar daar hou ik juist wel van. Een aanrader
dus.", "author": {"@type": "Person", "name": "K.kerkhof9
."}, "reviewRating":
{"@type": "Rating", "ratingValue": "8.5", "bestRating": "10"}},
{"@type": "Review", "datePublished": "2015-12-
11", "reviewBody": "Heerlijk! Vooral de citroentaart met
aardbeien, super. Kom hier al jaren en het blijft van
prima kwaliteit. Echt Yam Yam!", "author":
{"@type": "Person", "name": "Erna.bijldevroe
."}, "reviewRating":

```

address	
@type	PostalAddress
streetAddress	Frederik Hendrikstraat 88-90
postalCode	1052 HZ
addressLocality	Amsterdam
addressCountry	
@type	Country
name	Nederland
aggregateRating	
@type	AggregateRating
ratingValue	8.1
bestRating	10
ratingCount	70
geo	
@type	GeoCoordinates
latitude	52.376149
longitude	4.873466
review	
@type	Review
datePublished	2015-12-17

Afb. 13.28 - Rapportage door Google's Structured Data Testing Tool van in een webpagina gedetecteerde structured data. Dit voorbeeld betreft Schema.org markup in een JSON-LD script.

<https://www.schemaapp.com/60-structured-data-tools-create-test-plugins-more/>.

Gebruikte literatuur

DEN(z.j.) De projectenbank: innovatieve projecten van Nederlandse erfgoedinstellingen. Geraadpleegd op: 19 augustus 2016, van: <http://www.den.nl/projecten/?pl=A&sort=projecten.titel&asc=ASC>

Getty institute (z.j.) Geraadpleegd op 19 oktober 2016, van: <http://www.getty.edu/research/tools/vocabularies/lod/>

NIBG (z.j.) Gemeenschappelijke Thesaurus Audiovisuele Archieven. Geraadpleegd op 19 augustus 2016, van: <http://gtaa.beeldengeluid.nl>

RKD (z.j.) AAT Nederland. Geraadpleegd op: 21 augustus 2016, van: <http://website.rkd.nl/Databases/aat-ned>

Definities

Aboutness - Uitdrukking van het onderwerp waar een document over gaat.

Administratieve metadata - Gegevens die van belang zijn bij het beheer van documenten en gegevensobjecten; te denken valt aan datum van aanmaak en eigenaar.

Allocatie - Het fysiek of digitaal verspreid staan van objecten met een gezamenlijk kenmerk, omdat dit kenmerk niet het primaire indelingscriterium is.

Analytisch-synthetisch schema - Classificatieschema waarbij alleen afzonderlijke klassen getoond worden die gebaseerd zijn op dezelfde homogene groep. De combinatie van afzonderlijke elementen tot een nieuwe klasse wordt gedaan tijdens de indexeerfase, waarbij in principe elke combinatie nog mogelijk is. Veel classificaties hebben analytisch-synthetische elementen doordat gebruik wordt gemaakt van lijsten met landen, talen, vormen.

Associatieve relaties - Semantische relaties tussen termen die niet als gelijkwaardigheidsrelatie of hiërarchische relatie worden aangemerkt. Termen met een associatieve relatie worden vaak aangeduid als gerelateerde termen.

Automatisch verrijken - Het automatisch door software laten toekennen van metadata aan documenten of het laten indelen van deze documenten in een bepaalde categorie.

Autorisatie - Controle op toe te kennen metadata met behulp van voorgedefinieerde lijsten met termen.

Best match - Zoektechnologie die bepaalt welke documenten uit een collectie het meest overeenkomen (matchen) met een gestelde zoekvraag. Op basis van deze bepaling kan een ordening op relevantie worden aangebracht.

Betekenisrelatie - zie semantische relatie.

Booleaanse operatoren - Wiskundige operatoren genoemd naar George Boole (1815-1864) die ze introduceerde als onderdeel van de Booleaanse logica en de verzamelingenleer. Bij zoekacties in informatiesystemen worden met name de AND, OR en NOT operator ingezet om twee of meer termen in een zoekactie te betrekken.

Classificatie - Systeem waarin begrippen (onderwerpen) naar hun inhoudelijke samenhang op basis van een indelingscriterium geordend zijn in zo homogeen mogelijke groepen (klassen) en in hun hiërarchische relatie worden weergegeven.

Classificatieschema - Een vastgelegde classificatie, voorzien van labels, codes of notaties.

Classificeren - Het bouwen van een classificatie.

Clusters - Groepen documenten die op basis van bepaalde onderlinge overeenkomsten bij elkaar zijn geplaatst. Daartoe worden veelal automatische, statistische analysemethoden toegepast.

Co-extensie - Regel binnen de classificatietheorie die aangeeft dat de gezamenlijke omvang van het totaal van de subklassen even groot moet zijn als de overkoepelende klasse, met andere woorden: alles wat onder een overkoepelende klasse thuishoort moet in één van de subklassen ondergebracht kunnen worden.

Collocatie - Het fysiek of digitaal bij elkaar plaatsen van objecten op basis van een gezamenlijk kenmerk.

Collocatievolgorde - Volgorde van begrippen binnen een klasse.

Combinatievolgorde - Volgorde van de aspecten die achtereenvolgens worden ingezet als indelingscriterium voor het onderverdelen van begrippen. Binnen de classificatietheorie wordt daarbij gekozen voor een volgorde waarbij het meest specifieke aspect als belangrijkste indelingscriterium geldt.

Concept - Een begrip of cognitieve eenheid, waarmee een bepaalde klasse of idee als object aangeduid kan worden.

Conceptuele analyse - Inhoudelijke analyse van een document waarbij wordt gekeken naar concepten, facetten en/of invalshoeken die daarin aan de orde zijn.

Content - Datgene wat als data of informatie in een informatiesysteem of in een onderdeel daarvan is geplaatst.

Cross-classification - Het plaatsen van gelijksoortig materiaal in verschillende klassen. Dit is meestal het gevolg van het niet consequent aanhouden van de combinatievolgorde bij het indelen van klassen of mappen.

Data - Kale gegevens. Het hangt van de context en het gebruik af in hoeverre deze betekenis hebben.

Decompounding - Opsplitsen van samengestelde woorden in hun losse componenten. Voorbeeld: Kabinetcrisis opsplitsen in de woorden Kabinet en Crisis.

Descriptive taxonomy - Hierarchische lijst met begrippen die als metadata aan documenten kunnen worden toegekend.

Diepte-indexering - Het toekennen van voldoende termen om het onderwerp van een document compleet te beschrijven, waarbij aandacht is voor deelonderwerpen en onderdelen van een document. De tegenhanger is globaalindexering.

Directe toegankelijkheid - Vindbaarheid van documenten op basis van een gekozen ordening. Het gebruik van een catalogus of zoekmachine is daarbij niet nodig.

Document - Vastgelegde en reproduceerbare tekst, geluid of beeld in analoge of digitale vorm.

Eenheid van verdelingskarakteristiek - De onderverdeling van een klasse op basis van één en hetzelfde criterium.

Entiteiten - Met een naam benoembare dingen of objecten, zowel concreet als abstract, zoals personen producten, plaatsen of gebeurtenissen ; in het Engels wel named entities genoemd.

Enumeratieve classificatie - Classificatie waarin de zinvolle combinaties van verschillende elementen door de bouwer in het schema zijn vastgelegd in de klassen. Binnen een klasse dienen de relevante (andere) homogene groepen als een herkenbare verdere onderverdeling.

Evaluatiefunctie - Functionaliteit die het de gebruiker mogelijk maakt om vast te stellen of het resultaat van een zoekactie de gezochte informatie bevat.

Facetclassificatie - Classificatie waarin de homogene groepen afzonderlijk te herkennen zijn.

Facetzoeken - Zoekmogelijkheid waarbij de gebruiker resultaten kan filteren op basis van verschillende aspecten (facetten). Een voorbeeld zijn sites met vakantiehuizen waarbij gefilterd kan worden op aantal personen, ligging, aanwezigheid zwembad, wifi enzovoort.

False coördination - Verkeerd verband tussen de gezochte woorden in een zoekresultaat. Een zoekactie met de termen 'bonus' en 'directeur' kan zowel documenten opleveren over bonussen die de directeur ontvangt als over bonussen die de directeur aan zijn medewerkers toekent.

Folksonomy - Samentrekking van de Engelse termen Folks en Taxonomy. Verzameling van termen die door de gebruikers van een informatiesysteem aan documenten zijn toegekend, bedoeld om informatie makkelijker te kunnen vinden en met elkaar te delen. De termen zijn in principe niet afkomstig uit een gecontroleerde lijst en kunnen uiteenlopende aspecten van een document betreffen.

Formele metadata - Metadata van een document die je kunt vaststellen zonder inhoudelijke interpretatie: auteur en titel zijn hiervan voorbeelden, evenals de lengte van een film of de resolutie van een foto.

Free-text retrieval - Zoekmethode waarbij de digitaal beschikbare tekst van een document zelf wordt doorzocht.

Fuzzy zoeken - Automatisch laten zoeken naar woorden die wat betreft spelling of uitspraak sterk lijken op de woorden in de zoekvraag.

Gelijkwaardigheidsrelatie - Relatie tussen termen met gelijke of gelijkwaardig toegekende betekenis.

Generiek zoeken - Zoekactie waarbij ook op de hiërarchisch onderliggende termen (narrower terms) van een zoekterm gezocht wordt.

Generieke lijst - Lijst met termen waarbij de hiërarchische verdeling van termen zichtbaar is.

Generieke relatie - Hierarchische relatie op basis van geslacht-soort.

Globaalindexering - Indexering waarbij documenten worden gekarakteriseerd met termen die het hele document betreffen. De tegenhanger is diepte-indexering.

Grafen - Netwerken die (in wiskundige termen) bestaan uit "knooppunten" die onderling verbonden zijn door "takken" of "pijlen"; op het terrein van kennisorganisatie worden de knooppunten gevormd door concepten of entiteiten, die onderling zijn verbonden door relaties, zoals de "predicaten" in RDF-tripels.

Hierarchische relatie - Relatie tussen termen die boven- en ondergeschied aan elkaar zijn, op basis van geslacht-soort of geheel-deel.

Hyperfictie - Literair genre waarbij de loop van een verhaal door de lezer wordt bepaald op basis van de hyperlinks die worden gekozen.

Hyperlink - Verbinding tussen informatie-eenheden in een hypertext, die door een gebruiker geactiveerd kan worden om een gerelateerde informatie-eenheid op te vragen.

Hypertext - Informatiesysteem dat is opgebouwd uit individuele brokstukjes informatie in allerlei verschijningsvormen, waartussen relaties zijn vastgelegd waardoor het een systeem vormt dat niet sequentieel georganiseerd is, maar via systematische en/of functionele en/of associatieve relaties tussen die stukjes informatie.

Indexeerdiepte - De mate van detail waarin een document geanalyseerd en geïndexeerd wordt. De indexeerdiepte is uit te drukken in een getal: het aantal termen dat aan een document is toegekend.

Indexeren - Termen toevoegen aan een document.

Indirecte toegankelijkheid - Toegankelijkheid van informatie, waarbij vanuit een catalogus of ander systeem wordt doorverwezen naar de plaats van een document (fysiek of digitaal).

Informatie - Gegevens die zodanig bijeen zijn gebracht en worden voorgesteld dat er betekenis aan kan worden toegekend.

informatietaal - Kunstmatige taal voor het ontsluiten van de inhoud en bepaalde formele aspecten van documenten.

Information retrieval - Algemeen: het vindbaar maken en gericht zoeken naar informatie. Specifiek (met name internationaal): het automatisch toegankelijk maken van informatie met behulp van zoekmachines.

Inhoudelijke metadata - Metadata die iets zeggen over de inhoud van een document. Dit kunnen trefwoorden, thesaurustermen, tags of een samenvatting zijn.

Inhoudstypen - Metadata-optie in SharePoint waarmee het type document kan worden aangeduid. Tevens kan hiermee het gedrag van documenten worden gestuurd zoals bewaartermijnen en de toekenning van metadata.

Interoperabiliteit - Mogelijkheid om informatiesystemen aan elkaar te koppelen door gebruik van dezelfde termenset of door middel van een overkoepelende, vertalende laag (zoals een KOS).

Isness - Uitdrukking van wat voor type document (boek, rapport, artikel) sprake is.

Kennis - Informatie die door ervaring, vaardigheid en attitude zodanig wordt verwerkt dat zij in een later stadium kan worden toegepast.

Klasseren - Het onderbrengen van informatie in een bepaalde klasse of informatie voorzien van metadata van een in een schema vastgelegde structuur.

KOS - Knowledge organisation system: Systeem waarin op geformaliseerde wijze kennis over een afgebakend deel van de wereld en op afgebakend niveau en diepgang is vastgelegd. Voorbeelden zijn taxonomieën, thesauri, ontologieën, woordenboeken en grafen.

Known item - Informatie-eenheid waarvan de gebruiker weet dat hij bestaat en die op basis van bij de gebruiker bekende belangrijke kenmerken teruggevonden kan worden.

Lexical phrases - Drie of meer woorden die gezamenlijk één concept uitdrukken. Een voorbeeld is: 'Opwarming van de aarde'.

Linked open data - Data die openbaar beschikbaar is en via RDF-achtige technieken gestructureerd opgeslagen en met elkaar verbonden is.

Locatiefunctie - Functionaliteit die ervoor zorgt dat vanuit een retrievalssysteem wordt verwezen naar de plek waar een document gevonden kan worden.

Machine learning - Techniek uit het domein van de kunstmatige intelligentie, waarbij een computerprogramma op basis van een aantal gegeven voorbeelden leert om bepaalde taken uit te voeren, zoals bijvoorbeeld het klasseren van tekstdocumenten of van foto's.

Maskeren - Interne truncatie: het vervangen van één of meer letters binnen een woord door 'wildcards' om aan te geven dat op die plaats willekeurige letters mogen voorkomen.

Metadata - "Gegevens over gegevens": gegevens die aan objecten worden toegevoegd om ze daarmee te karakteriseren, teneinde ze daarmee vindbaar, identificeerbaar en interpreteerbaar te maken en/of op grond daarvan te kunnen ordenen.

Metadata-architectuur - Algemeen: Overzicht van metadata en eventuele bijbehorende informatietalen die ingezet worden binnen één of meer systemen. Binnen SharePoint: Het vaststellen en inzetten van kolommen, termensets, inhoudstypen en mappen binnen een site of site-verzameling.

Metagegevens - Zie metadata.

Modulatie/gradatie - Geleidelijke overgang bij het onderverdelen van een klasse in specifiekere subklassen.

Namespace - Computerleesbare beschrijving van een gebruikte standaard waar bijvoorbeeld in een RDF-tripel naar verwezen kan worden.

Navigational taxonomy - Onderverdeling van begrippen die rubrieken oplevert waaronder documenten kunnen worden geplaatst.

Niveau-indexering - Aan een document toekennen van twee termen uit dezelfde hiërarchische structuur maar uit verschillende niveaus.

Noun phrases - Combinatie van twee of meer losse zelfstandige naamwoorden, die gezamenlijk één begrip representeren (vooral in Engelse tekst; bijvoorbeeld "information retrieval").

Ofness - Uitdrukking van hetgeen op een afbeelding te zien is, in algemene termen als 'vogel', 'vliegtuig' zonder dat wordt aangegeven welke vogel of vliegtuig het betreft.

Ontleend indexeren - Het creëren van metadata door middel van het automatisch betrekken van termen uit digitale documenten teneinde deze documenten vindbaar te maken.

Ontlenen - Het betrekken van termen uit de aanwezige tekst van documenten, om als metadata te fungeren.

Ontologie - Een strikt en uitputtend schema voor een bepaald onderwerpsdomein, meestal in een hiërarchische structuur, dat alle relevante grootheden en hun relaties bevat en ook de regels waaraan die grootheden en relaties binnen dat domein voldoen. Binnen het semantisch web ook een containerbegrip waarmee elk soort kennisorganisatiesysteem kan worden aangeduid, waaronder thesauri, taxonomieën en semantische netwerken.

Ontsluiting - Binnen informatie-omgevingen: methoden en technieken om informatie toegankelijk te maken.

Opbrengst - Het totale aantal treffers dat een zoekactie oplevert
Partitieve (partiele) relatie Hiërarchische relatie op basis van deel-geheel.

Polyhiërarchische relatie - Hierarchische relatie waarbij een term meer dan één algemenere (bovenliggende) term heeft.

Postcoördinatie - Systeem waarbij een complex onderwerp gevonden moet worden door in de zoekfase termen als samenstellende enkelvoudige begrippen te combineren.

Precisie - Dat deel van de bij een zoekactie gevonden treffers dat ook daadwerkelijk relevant is; wordt meestal uitgedrukt in het percentage ten opzichte van de opbrengst.

Precoördinatie - Het in vaste volgorde in één indexterm vastleggen van de elementen uit een complex onderwerp tijdens de bouwfase of de indexeerfase (dus voor de zoekfase).

RDF - Resource Description Framework: algemene methode om de relatie tussen te karakteriseren objecten, alsmede hun eigenschappen en de daarvoor te gebruiken metadatastandaarden, geformaliseerd vast te leggen.

Relevantie - Mate waarin een document voldoet aan de informatievraag van de zoeker.

Relevantie-ordening - Ordening van zoekresultaten op basis van waarschijnlijke mate van relevantie voor de zoeker. De ordening vindt plaats op basis van door de computer berekende factoren.

Ruis - In de context van zoeksystemen: dat deel van de opbrengst van een zoekactie dat niet relevant is.

Schemavolgorde - Volgorde waarin de onderwerpen in een classificatieschema worden aangeboden.

Scope note - Toelichting bij een thesaurusterm.

SEA - Search Engine Advertising: Het koppelen van betaalde advertenties aan zoekvragen in een web-zoekmachine.

Semantisch netwerk - Een netwerk van woorden waartussen getypeerde relaties zijn gelegd. De aard van die relaties is anders dan bij een thesaurus; het aantal mogelijke relaties kan ook groter zijn.

Semantisch web - Ook wel web 3.0. Een beoogde volgende generatie van het World Wide Web waarin veel van de gegevens zijn voorzien van semantische kenmerken, zodat computers die zelfstandig als informatie kunnen interpreteren, betekenis kunnen selecteren en informatie uit verschillende bronnen aan elkaar kunnen relateren.

Semantisch zoeken - Verzameling technieken die bij het zoeken gebruik maakt van de betekenis en de intentie van woorden en zinnen, zowel

in zoekvragen als in gevonden informatie; meer in het bijzonder: zoektechniek die - in plaats van tien links - concrete antwoorden geeft, omdat zoekvragen begrepen worden en antwoorden afgeleid kunnen worden.

Semantische relatie - Relatie tussen twee woorden op basis van betekenis.

SEO - Search Engine Optimization: Activiteiten om webpagina's door middel van goede content, links en metadata zo hoog mogelijk in de zoekresultaten van een web-zoekmachine te laten komen en om gebruikers te verleiden de links naar die pagina's aan te klikken. Dat laatste door gebruik van aansprekende titels en gestructureerde metadata die zorgen dat 'rich snippets' getoond worden.

Site-architectuur (SharePoint) - Vaststelling van site-onderdelen per SharePoint-site en plaatsing van sites binnen een site verzameling.

Sitemap - Bij voorkeur visuele weergave van de structuur van een website: de pagina's en hun onderlinge verbanden.

SKOS - Simple Knowledge Organisation System: Een standaard om thesauri en andere eenvoudige kennisorganisatiesystemen computerleesbaar op internet aan te bieden en zo mogelijk te koppelen.

Sleutelfilterveld - Venster in SharePoint waarin de gebruiker kan zoeken op toegekende termen.

SPARQL - SPARQL Protocol And RDF Query Language: gestandaardiseerde computertaal om RDF-gecodeerde linked open data collecties te bevragen.

Specificiteit - De mate waarin de inhoud van een document exact wordt omschreven in termen van een informatietaal.

Stemming - Zie Word-stemming.

Syntactische analyse - Analyse van woorden en termen zodat equivalentie tussen zinsdelen (lexical phrases) en losse woorden kan worden vastgesteld. Voorbeeld: besparing van energie = energiebesparing.

Syntactische relatie - Relatie tussen woorden die niet is gebaseerd op betekenis, maar op een causale, temporele of andere relatie tussen begrippen. Bijvoorbeeld: 'bijwerkingen van aspirine', 'schilderkunst in de 19de eeuw'.

Tag - Term die wordt toegekend aan een digitaal object of document, die verschillende soorten kenmerken daarvan kan beschrijven, die veelal vanuit het perspectief van een eindgebruiker wordt toegekend en die als

metadata gebruikt kan worden. Een tag wordt in het algemeen niet aan een informatietaal ontleend. (zie ook: folksonomy)

Taxonomie - Een hiërarchische onderverdeling van begrippen die tot rubrieken leidt waarin materiaal kan worden geplaatst of die als termen kunnen worden gebruikt om aan documenten toe te kennen. In het Angelsaksisch taalgebied wordt Taxonomy vaak gebruikt als containerbegrip voor alle typen informatietalen of ordeningssystemen, zelfs als daar geen hiërarchie aan te pas komt.

Term - Woord, woordgroep of code die deel uitmaakt van een informatietaal.

Termenset - Benaming voor de gebruikte informatietaal binnen SharePoint.

TF*IDF - Statistische methode om te bepalen hoe kenmerkend een woord is voor een document. Hierbij wordt het aantal keren dat een woord in het document voorkomt (de termfrequentie TF) vermenigvuldigd met de inverse documentfrequentie (IDF). Oftewel: de termfrequentie wordt gedeeld door het aantal documenten waarin dat woord voorkomt.

Thesaurus - Een geordende verzameling van termen die bestaat uit zoveel mogelijk enkelvoudige begrippen op basis van natuurlijk taalgebruik, waarvan de onderlinge semantische relaties zijn vastgelegd en welke gebruikt wordt voor indexeren en postcoördinatie zoeken.

Thesonomie - Samentrekking van de termen Thesaurus en Taxonomie. Aanduiding van de informatietaal zoals die in SharePoint gebruikt wordt, waarbij wel hiërarchische en equivalente relaties mogelijk zijn, maar geen associatieve. Zie ook termenset.

Toegekend indexeren - Het indexeren van documenten waarbij de toe te kennen termen voortkomen uit een analyse van de inhoud van dat document en niet als zodanig aan het document zelf ontleend hoeven te zijn. Toegekend indexeren kan zowel door mensen als door software worden uitgevoerd.

Trunceren - Zoeken op varianten en samenstellingen van een woord, waarbij het variabele deel van het woord is vervangen door een wildcard; meestal alleen gebruikt voor afkapping aan het eind of eventueel het begin van een woord (zie ook maskeren).

Vangst - Dat deel van de in een systeem aanwezige relevante documenten dat bij een zoekactie daadwerkelijk is gevonden.

Vectormodel - Techniek waarbij documenten worden gerepresenteerd door wiskundige vectoren, die worden berekend op basis van de in de documenten voorkomende woorden.

Vervangend zoeken - Functionaliteit in een informatiesysteem die zorgt dat bij een zoekactie op een niet-voorkeursterm het systeem automatisch een zoekactie op de voorkeursterm uitvoert.

Verwijsterm - Niet voorkeursterm in een thesaurus die doorverwijst naar de voorkeursterm.

Voorkeursterm - Thesaurusterm die gebruikt wordt voor het indexeren van en zoeken naar informatie.

Word-stemming - Reductie van woorden tot hun morfologische woordstam, zodat bij de verwerking van die woorden geen onderscheid meer wordt gemaakt tussen enkel- en meervoud, vervoegingen, verbuigingen e.d.,

Rubenianum (z.j.) Collectie. Geraadpleegd op 19 oktober 2016, van:
<http://anet.ua.ac.be/desktop/rub/core/index.phtml?language=N&eu-ser=&session=&service=opacrub&robot=&deskservice=desktop&desktop=rub&workstation=RUB&extra>

Begrippenlijst

Begrip	Paragraaf
Aboutness	1.3
Abstracte termen	4.5
Administratieve metadata	1.4
Alle/sommige toets	4.5
Allocatie	6.1.2
Analytisch-synthetisch schema	6.2.3
AND-operator	2.4.2
Associatieve relaties	4.5; 12.3
Automatisch verrijken	10.1.2
Automatische beeldherkenning	Kader
Automatische technieken	11.1
Autorisatie	5.1
Beeldmateriaal	Kader
Beheerde metadata	8.2.2
Best match	2.4.3
Betekenisrelaties	9.1
Betrouwbaarheidsdrempel	10.2.2
Bibliotheek	8.1.2
Booleaanse operatoren	2.4.2
Boomstructuur	3.2
Cardsorting	Kader
Categorieën	11.5
Checklist opstellen randvoorwaarden thesaurus	Kader
Cidoc CRM	9.2
Classificatie	6.2
Classificatieproces	10.2.2
Classificaties	1.5
Classificatieschema	6.2
Clusters	11.5
Co-extensie	3.3.3
Collocatie	6.1.2
Collocatiefunctie	1.4
Combinatievolgorde	6.1.2
Concept	4.4

Conceptuele analyse	7.1
Concrete termen	4.5
Content	1.1
Contentsilo	1.1
Core ontologies	9.2
Cross-classification	6.1.2; 6.2.5
Data	1.1
Dbpedia	9.6.2
Decompounding	10.3; 11.2
Descriptive taxonomy	3.2
Diepte-indexering	7.2.2
Directe toegankelijkheid	1.3
Document gerichte benadering	7, inleiding
Dynamische inhoud	3.3.4
Eenheid van verdelingskarakteristiek	3.3.3
Eigenschappen	8.2.1
Enumeratief classificatieschema	6.2.3
Enumeratief schema met analytisch-synthetische mogelijkheden	6.2.3
Evaluatiefunctie	1.4
Facetclassificatie	6.2.3
Faceted search	11.6
Facetten	3.2
Facetzoeken	2.5.4
False Coördination	1.7
Filteren	2.5.4
Folksonomies	7.4
Formele metadata	1.4
FRBR	12.1
Freetekst retrieval	11, inleiding
Fuzzy zoeken	11.2
Gebruikersgerichte benadering	7, inleiding
Gelijkwaardigheidsrelaties	4.5
Generiek zoeken	2.5.3; 4.5
Generieke	4.5
Generieke lijst	4.5
Geonames	9.6.2
Gevolgtrekkingen	9.2
Globaalindexering	7.2.2
Google knowledge graph	9.6.3
Grafen	9.5
Hashtag	7.4
Hiërarchie	3.1
Hiërarchieveld	8.2.6
HTML	12.2
HTTP	12.2
Hyperbolische browsers	12.4
Hypercard	12.2

Hyperfictie	12.6
Hyperlink	12, inleiding
Hypertext	12.2
Indexeerdiepte	7.2.1
Indexeren	1.6
Indirecte toegankelijkheid	1.3
Inference Engine	9.2
Informatie	1.1
Informatietaal	1.5
Informatieverrijking	7, inleiding
Information retrieval	2, inleiding
Inhoudelijke metadata	1.4
Inhoudsanalyse	7.1
Inhoudskarakteristiek	7.1
Inhoudstypen	8.2.4
Interoperabiliteit	9, inleiding
Isness	1.3
ISO 25964	Kader
Kennis	1.1
Kennisregels	10.2.1
Klasseren	7.3
Knowledge cards	9.6.3
Knowledge organisation systems	9, inleiding
Knowledge vault	9.6.3
Known item	2.3
Kolommen	8.2.1
Lexical phrases	11.6
Lijst	8.1.2
Linguistische methoden	10.3
Linked data	9.6.2
Linked open data	9.6.2
Literal	9.4
Locatiefunctie	1.4
Machine learning	Kader
Managed metadata store	8.2.2
Maskeren	11.2
Meervoudige concepten	1.7
Memex	12.2
Metadata	1.4
Metadata-architectuur	8.2
Metagegevens	8.2.6
Methoden	10.2
Modulatie/gradatie	3.3.3
More like this	11.7
Namespaces	9.4
Navigatie-instellingen	8.2.6
Navigatiepad	12.4

Navigational taxonomy	3.2
Near-operator	2.4.4
Niveau-indexering	7.3
NOT-operator	2.4.2
Noun phrases	11.6
Object	9.4
Ofness	1.3
Ondernemingstrefwoorden	8.2.3
Onderwerpsindeling	2.2
Onderwerpsvraag	2.3
Ongecontroleerde trefwoorden	7.4
Ontleend indexeren	11, inleiding
Ontlenen	1.3
Ontologie	9.2
Ontsluiting	1, inleiding
Opbrengst	2.6.2
Ordering	1.3
OR-operator	2.4.2
OWL	9.2
Pagina	8.1.2
Pagina architectuur	8.3
Partitieve	4.5
Polyhiërarchische relaties	4.5
Postcoördinatie	1.7
Postcoördinatief	4.3
Precisie	2.6.2; 11.1
Precoördinatie	1.7
Predicaat	9.4
Qualifier	4.5
RDF Tripels	9.4
Relaties	9.2; 12.1
Relationele database	12.1
Relevantie	2.6.1
Relevantie-ordering	11.4
Relevantievolgorde	2.4.3
Rich snippets	9.6.4
Ruis	4.5
Schema.org	9.6.4
Schemavolgorde	6.2.2
Scope Note	4.5
Semantiek	4.3
Semantisch netwerk	9.1
Semantisch web	9.2
Semantisch zoeken	11.8
Semantische afstand	9.1
Semantische kennis	10.3
Semantische relatie	3.2

SEO en SEA	Kader
Sharepoint mappenstructuur	8.2.7
Sharepoint site	8.1.1
Signatuur	6.2.3
Sitemap	12.4
Site-onderdelen	8.1.2
Sitetemplate	8.1.2
Siteverzameling	8.1.1
SKOS	9.6.1
Sleutelfilterveld	8.2.6
SPARQL	9.6.2
Specificiteit	6.2.3
Spellingsvarianties	2.4.5
Splitsen	11.2
Statistishce methoden	10.3
Subject	9.4
Syntactische analyse	10.3
Syntactische relatie	3.2
Systematische basisstructuur	12.4
Systematische navigatielinks	12.4
Tags	7.4
Taxonomie	3.2
Term	4.4
Termdefinitie	4.5
Termenset	8.2.2
Termkeuze	4.5
Termredactie	4.5
Terugkoppeling	11.7
TF*IDF	10.3; kader
Thesauri	1.5
Thesaurus	4.3
Thesaurussoftware	4.5
Thesonomie	8.2.2
Tijdperk	5.6
Tijdvak	5.6
Toegekend indexeren	1.4
Toekennen	1.3
Trainen	10.2.2
Trefwoordenlijsten	1.5
Triple stores	9.4
Truncatie	2.4.5
Trunceren	11.2
Upper ontologies	9.2
Vangst	2.6.2; 11.1
Vaste Inhoud	3.3.4
Vectormodel	Kader
Vectorrepresentatie	10.2.2

Velden	2.5
Verfijnen	11.6
Vervangend zoeken	2.5.3; 4.1
Visualisatie	12.4
Wijsheid	1.1
Woord	4.4
Woordstam	2.4.5 ; 10.3
Woordsystemen	1.5
Wordnet	9.1
Wordpress	Kader
Word-stemming	11.2
XML	Kader
Zoekhulpmiddelen	2.2
Zoeksysteem	2.1
Zuiverheid	4.5