

Inhoudelijk toegankelijk maken van informatiecollecties in een digitale omgeving

Eric Sieverts

*Instituut voor Media & Informatie Management,
Hogeschool van Amsterdam*

&

Universiteitsbibliotheek Utrecht

[preprint van bijdrage aan:
Handboek informatiewetenschap voor bibliotheek en archief, Kluwer, Alphen aan den Rijn,
februari 2007]

Inhoud

Inleiding	3
1. Methoden van inhoudelijke ontsluiting	4
1.1 Stand van zaken	4
1.1.1 Classificaties	6
1.1.2 Thesauri.....	9
1.1.3 Ontologieën.....	11
1.1.4 Topic maps.....	12
1.1.5 “Tagging” en folksonomies	13
1.2 Interactie tussen gebruiker en ontsluitingssysteem.....	14
1.3 Praktische overwegingen	15
1.4 De toekomst van inhoudelijke ontsluiting	16
2. Geautomatiseerde methoden van inhoudelijke karakterisering, classificatie en verrijking.....	17
2.1 Stand van zaken	17
2.2 Toepassen van geautomatiseerde inhoudelijke ontsluiting.....	20
2.3 Praktische voorbeelden	21
2.4 De toekomst van geautomatiseerde ontsluiting	24
3. Free-text retrieval in plaats van ontsluiting?	25
3.1 Stand van zaken	26
3.2 Free-text retrieval of inhoudelijke ontsluiting	29
3.3 De toekomst van free-text retrieval methoden.....	30
4. Ontsluiten van heterogeen materiaal	31
4.1 Digitaal versus niet-digitaal materiaal	32
4.2 Gespecialiseerde versus algemene publicaties.....	34
4.3 Fictie en non-fictie.	35
5. Interoperabiliteit van systemen	36
5.1 Interoperabiliteit van metadata schema’s.....	37
5.2 Concordantie van ontsluitingssystemen.....	38
5.3 Praktische toepassingen van interoperabiliteit.....	40
5.4 Terminology Services en de SKOS-standaard.....	41
6 Enkele algemene conclusies.....	42
Literatuur	44
Gegevens van relevante projecten, producten en websites	51
Gebruikte afkortingen	55

Inleiding

Inhoudelijke ontsluiting was van oudsher de aangewezen methode om informatiecollecties op onderwerp toegankelijk te maken. Het houdt in dat documenten inhoudelijk worden gekarakteriseerd door daaraan begrippen of woorden toe te kennen of door die documenten aan categorieën of rubrieken toe te delen. Met de beschikbaarheid van steeds meer informatie in digitale vorm, is ook de mogelijkheid ontstaan om informatie door de computer te laten terugzoeken op basis van indexen van de woorden die zijn ontleend aan de inhoud van de documenten zelf, alle woorden, met de Google-experience als ultieme toepassing daarvan. Dat maakt de vraag relevant of ontsluiting - ook wel toegekend indexerend genoemd - nog altijd noodzakelijk is om informatie goed terug te kunnen vinden.

In elk geval blijkt er nog onverminderde belangstelling te zijn voor inhoudelijke ontsluiting, hetgeen zich uit in de ontwikkeling van nieuwe ideeën en toepassingen. Met de term Knowledge Organisation Systems worden alle soorten systemen voor gecontroleerde inhoudelijke ontsluiting aangeduid. Naast klassieke, als classificaties en thesauri, vallen daaronder ook nieuwe soorten als ontologieën en topic maps. Daarbij zijn classificaties vooral een belangrijk hulpmiddel om gebruikers via de systematische indeling van categorieën via opvolgende keuzes bij de juiste rubriek te laten uitkomen. Dankzij het Web en de daar gebruikte navigeer- en link-technieken bestaat hiervoor hernieuwde belangstelling. Een aantal van dergelijke innovatieve toepassingen van classificaties zullen in dit artikel aan de orde komen. Hoewel dergelijke projecten in eerste aanzet vaak gericht zijn op ontsluiting van materiaal op het web, blijken veel ervaringen en conclusies ook van belang voor het toegankelijk maken van ander materiaal, ook bibliotheekcollecties, al dan niet digitaal.

Thesauri zijn vooral een belangrijk hulpmiddel in situaties waarin gebruikers zelf hun zoekvragen kunnen of moeten formuleren. Omdat daarbij vaak een discrepantie bestaat tussen gebruikerstaal en systeemtaal, zal aandacht worden besteed aan methoden om die te overbruggen, bijvoorbeeld door zoektermen van gebruikers automatisch te associëren met corresponderende thesaurustermen. Inhoudelijke ontsluiting is duur, met welk soort systeem het ook wordt gedaan. Goedkoper kan het door dat proces door de computer te laten uitvoeren via analyse van het - daarvoor wel noodzakelijkerwijze digitaal aanwezige - tekstmateriaal. Het lijkt dat de kwaliteit daarvan al heel bevredigend is. Een eveneens goedkoop alternatief is de techniek van "tagging", waarbij gebruikers van informatiesystemen zelf materiaal van trefwoorden kunnen voorzien. Maar dat is dan geen gecontroleerde ontsluiting meer.

Een ander aspect waaraan hier aandacht wordt besteed is heterogeniteit. Enerzijds kan die zich uiten in de heterogeniteit van het materiaal zelf, wat het noodzakelijk kan maken een gevarieerde aanpak van de ontsluiting toe te passen. Anderzijds kan die heterogeniteit het gevolg zijn van het feit dat in de huidige netwerkgeving afzonderlijke collecties met verschillende structuur en wijze van ontsluiting toch geïntegreerd doorzocht kunnen worden. In dat kader is het aspect van de interoperabiliteit van die systemen van belang, waarbij in het algemeen wordt uitgegaan van collecties die wel al van enige vorm van gecontroleerde ontsluiting zijn voorzien.

1. Methoden van inhoudelijke ontsluiting

1.1 Stand van zaken

De laatste jaren bestaat groeiende belangstelling voor standaardisatie van metadata-systemen. Daarbij is het opmerkelijk dat vaak meer nadruk wordt gelegd op standaardisatie van de formats dan van de inhoud van de metadata (Milstead 1999). In dit hoofdstuk wordt gekeken in hoeverre die inhoud inderdaad nog van belang wordt geacht voor inhoudelijke ontsluiting van informatie en welke ontwikkelingen er op dat terrein zijn geweest. Daarbij komen verschillende methoden voor gecontroleerde inhoudelijke ontsluiting aan de orde, met als voornaamste klassieke exponenten classificaties en thesauri.

Grotendeels los van degenen die zich met deze klassieke ontsluitingsmethoden bezig houden, heeft zich een onderzoekstak ontwikkeld die zich in het algemeen richt op "Knowledge Organisation Systems" (KOS). Hill (2002) geeft een helder overzicht van de verschillende soorten KOSsen die kunnen worden onderscheiden:

- Systemen voor classificatie en categorisatie (waaronder classificaties, subject headings en taxonomieën),
- Metadata-achtige modellen (waaronder geografische "gazetteers"),
- Relationele modellen (waaronder thesauri, semantische netwerken en ontologieën),
- Term lijsten (waaronder autorisatielijsten en woordenboeken).

Zij meent dat de KOS, meer dan voorheen, geïntegreerd moet worden in de architectuur van de Digitale Bibliotheek, in nauwe samenhang met de collecties en diensten die daarin aan gebruikers worden geboden. Functies van een KOS daarbij zijn beschrijvend (via gecontroleerde "labels"), definiërend (door betekenis aan die labels te geven), vertalend (via concordanties tussen kennisrepresentaties) en navigerend (via de gestructureerde wijze waarop ze georganiseerd zijn). Dit betekent dat ook concordantie en interoperabiliteit tussen ontsluitingssystemen - hier besproken in hoofdstuk 5 - tot het werkkerrein van de KOS behoren. In het kader van KOS-onderzoek wordt ook aandacht besteed aan methoden om gebruikers eenvoudiger toegang te bieden tot de vaak wat kunstmatige vocabulaires die voor gecontroleerde ontsluiting van informatie worden gebruikt (Binding 2004).

Opvallend is de herleving van op classificaties gebaseerde ontsluitingssystemen. Dit hangt direct samen met de opkomst van web-gebaseerde systemen. Classificaties zijn namelijk veel geschikter om door systematische onderwerpsindelingen te browsen en daaruit rubrieken aan te klikken, dan woordsystemen als thesauri. Classificaties kunnen een passieve zoekwijze bieden, waarbij de gebruiker niet zelf zijn onderwerp hoeft te omschrijven, maar waar hij uit gepresenteerde omschrijvingen van categorieën die keuzes kan aanklikken die het meest in overeenstemming lijken met zijn informatiebehoefte. In de uiteindelijk gekozen eindrubriek vindt de gebruiker in principe meteen alle relevante documenten bij elkaar, zonder dat hij zelf zoekwoorden heeft hoeven te bedenken.

Thesauri op hun beurt, zijn over het algemeen geschikter voor zoeksystemen, waarin de gebruiker zelf actief zijn zoekvraag kan formuleren. Die vraag zal veelal bestaan uit een door de gebruiker geformuleerde combinatie van de concepten die in de gewenste documenten tezamen aan de orde moeten komen. Een hinderpaal bij toepassing van thesauri aan de gebruikerskant is meestal de discrepantie tussen de belevingswereld van de gebruiker en het binnen een thesaurus vastgelegde voorkeursvocabulary. Dit is een van de oorzaken van het vaak gesignaleerde ondergebruik van dergelijke ontsluitingssystemen. Zie daarvoor bijvoorbeeld het overzichtsartikel over thesauri van Aitchison (2004) en een studie van Greenberg (2004). Ontwikkelingen om iets aan deze discrepantie te doen en daarmee de mate van gebruik te stimuleren - soms zelfs ongemerkt voor de gebruiker - worden in §1.2 apart belicht.

Naast het feit dat gecontroleerde ontsluitingssystemen, zonder inzet van speciale hulpmiddelen, als gebruiksonvriendelijk worden beschouwd, heeft het toekennen van categorieën of thesaurustermen als grootste nadeel dat daarin onevenredig veel tijd geïnvesteerd moet worden, als dat geheel handmatig moet gebeuren. Daarom is het interessant dat voor classificaties, en in zekere mate ook voor thesauri, geautomatiseerde methoden bestaan waarmee die toekenning door de computer uitgevoerd of op zijn minst ondersteund kan worden. Gezien het belang van deze methoden, wordt daarop in hoofdstuk 2 afzonderlijk ingegaan.

Naast klassieke ontsluitingssystemen als classificaties en thesauri, wordt ook geëxperimenteerd met nieuwe soorten systemen, zoals ontologieën en topic maps. Die komen in §1.1.3 en §1.1.4 aan de orde. Overigens is het woord ontologie in sommige kringen ook tot een soort generiek begrip geworden, waaronder in de praktijk uiteenlopende soorten ontsluitingsmethoden of KOSsen worden verstaan.

Op internet is, vrijwel los van de klassieke wereld van de informatievoorziening, een heel nieuwe aanpak ontstaan voor het vindbaar maken van informatie - in welke vorm dan ook - die van de activiteit van de gebruikers zelf uitgaat. Die collectieve ontsluitingsactiviteit wordt wel aangeduid als “tagging” en de daarbij spontaan ontstane metadata-schema's - naar analogie van het begrip “taxonomy” - als “folksonomies”. In §1.1.5 wordt daarop verder ingegaan.

1.1.1 Classificaties

Classificaties werden enige tijd beschouwd als achterhaald en van beperkt nut bij het toegankelijk maken van grote hoeveelheden informatie in een geautomatiseerde omgeving. Dit hing vooral samen met de opkomst van retrievalssystemen. Met woordsystemen als thesauri kunnen zoekspecialisten eenvoudiger op inhoudelijke elementen zoeken, dan met classificatiesystemen met hun vaak complex opgebouwde codes. In dit verband waren de hoogste niveaus van classificaties nog wel van nut om resultaten uit woordgebaseerde zoeksystemen in te perken op bepaalde disciplines. Daarnaast zijn classificatiesystemen van oudsher nuttig om eenvoudig overzicht over hele collecties te krijgen.

Met de komst van op hyperlinking gebaseerde web-interfaces op allerlei soorten informatiesystemen is die situatie sterk veranderd. Eenvoudige classificatiesystemen lenen zich namelijk wel goed als leidraad bij het browsen en navigeren door een systematische indeling in categorieën, waarin grote hoeveelheden informatie geordend kunnen zijn, of dat nu fysieke collecties betreft of collecties verzamelde links naar webpagina's. Voor dat doel dienen dan wel voldoende herkenbare benamingen voor rubrieken te zijn gebruikt en dient de rubrieksindeling te voldoen aan minimale kwaliteitseisen ten aanzien van de eenheid van verdelingskarakteristiek en het ontbreken van overlap tussen rubrieken. Met de introductie van het begrip taxonomie voor informatieverzamelingen ten behoeve van de interne kennisinfrastructuur van bedrijven en organisaties, volgens een meestal op classificaties gebaseerde ordening, hebben dergelijke ontsluitingsmethoden ook een meer "sexy" uitstraling gekregen.

Voor dit soort web-gebaseerde systemen zijn vereenvoudigde (versies van) classificaties nodig, omdat complexe notaties uit bijvoorbeeld de Universele Decimale Classificatie (UDC) moeilijk te vertalen zijn in overzichtelijke, in woorden omschreven keuzemenu's en aanklikbare hiërarchieën. Internationaal gaat de belangstelling op dit moment vooral uit naar de Dewey Decimale Classificatie (DDC), die makkelijk op die manier te gebruiken is, met als bijkomend voordeel dat veel - vooral Engelstalig - materiaal bij publicatie al DDC-codes meekrijgt (zie ook Tinker, 1999). Voor de meer complexe UDC en zelfs voor de Library of Congress Classificatie (LCC) bestaat in dit opzicht wat minder belangstelling.

Zowel voor DDC als voor LCC zijn experimenten uitgevoerd met aanpassingen aan hun structuur, om die beter geschikt te maken voor browsen en navigeren in een webomgeving. Deze aanpassingen betreffen zowel beperking van de diepte van toepassing van de oorspronkelijke classificatie, als gewijzigde rubriekvolgordes, onderverdelingen en omschrijvingen. Davis (2002) geeft een interessante beschrijving van een voor Columbia University uitgevoerde aanpassing van LCC ten behoeve van de toegankelijkheid van een collectie digitale bronnen van overigens beperkte omvang. Hij geeft daarin ook voorbeelden op welke punten de originele LCC niet ongewijzigd bruikbaar is voor deze toepassing. Hij gaat uit van twaalf hoofdcategorieën, gaat slechts bij uitzondering meer dan drie niveaus diep, past waar nodig verdubbelingen van rubrieken toe en gebruikt rubrieksomschrijvingen die vaak zijn ontleend aan het trefwoordvocabulaire uit de Library of Congress Subject Headings (LCSH). Hij benadrukt ook dat DDC in principe een beter uitgangspunt had geboden dan LCC, maar dat dat in zijn situatie "politiek onhaalbaar" was.

Voor de Duitse Nationale Bibliografie wordt in het DDC-Deutsch-project ten behoeve van web-presentatie met DDC geëxperimenteerd (Heiner-Freiling 2003). In een overzicht van diverse experimenten en aanpassingen van bestaande systemen komt ook Saeed (2001) tot de conclusie dat DDC het meest geschikt is voor een dergelijke toepassing. Hij concludeert echter ook dat DDC in een webomgeving in de praktijk vaak nog niet optimaal en weinig innovatief gebruikt wordt. Experimenten met toepassingen van automatische classificatie op basis van deze systemen komen in hoofdstuk 2 aan de orde.

Davis (2002) benadrukt het algemeen aanvaarde principe dat niet te veel muisklikken nodig moeten zijn om bij de gewenste informatie te komen. In de gebruikelijke boomstructuren stelt dat beperkingen aan het aantal niveaus van de hiërarchie. Omdat ook het aantal keuzes op elk niveau niet te groot mag worden, kan geen onbeperkt aantal categorieën worden aangeboden. Een maximum van drie niveaus en niet meer dan circa vijftien vervolgkeuzes per categorie, leidt dan tot een maximum aantal mogelijke categorieën van ruim 3000. Anderzijds zien we bij grote onderwerps-taxonomieën op het web (Yahoo, OpenDirectory) veel grotere aantallen categorieën - bij Open Directory zelfs meer dan 500.000 voor ruim vier miljoen items - en uiteraard veel grotere aantallen niveaus - bij Yahoo vaak wel acht of meer niveaus diep. Dat is ook niet verwonderlijk want in zulke grote systemen met enkele miljoenen items, zou het aantal items per categorie anders onhanteerbaar groot worden.

Dergelijke grote aantallen categorieën komen veelal voort uit precoördinatieve combinaties van meer onderwerpsfacetten. In die gevallen wordt de klassieke vraag welke combinatievolgorde moet worden aangehouden weer actueel. Delen we “Scuba-duiken in Australië” in volgens

Recreatie – buitensport – scuba – regionaal – Australië
of volgens

Regionaal – Australië – recreatie – buitensport – scuba ?

In een geautomatiseerde situatie kan gelukkig gezorgd worden dat beide volgordes als navigatiepad worden aangeboden, zodat de gebruiker hoe-dan-ook bij de beoogde klasse terecht komt, zoals Yahoo's webtaxonomie laat zien. Voor dergelijke zeer uitgebreide systemen wordt vaak aangeraden liever van de gewone zoekfunctie gebruik te maken, als methode om snel de juiste rubriek(en) voor een bepaald onderwerp te vinden. In die rubrieken zijn dan alle items over de gezochte onderwerpen al bij elkaar gebracht. Ook Davis (2002) presenteert daarom als antwoord op zoekvragen eerst de overeenkomende rubrieken en pas dan de op basis van de zoektermen gevonden items zelf. Overigens valt het voordeel van de browse-functie van classificaties bij dit soort gebruik dus weer weg.

Vizine-Goetz (2002) heeft op basis van meer dan 500.000 in WorldCat (OCLC) op basis van LCC en DDC gecatalogiseerde internetbronnen een vergelijking gemaakt met de webtaxonomieën van Yahoo en Looksmart. De verdeling van rubrieken over de niveaus bleek vooral tussen DDC en Yahoo heel vergelijkbaar. Datzelfde gold ook voor de verdeling van documenten over de niveaus en de rubrieken. De basis van de speciaal voor het web ontwikkelde classificaties blijkt vaak wel een heel andere te zijn dan die van de klassieke bibliotheekclassificaties. Hudon (2001) vergeleek daartoe Yahoo en een Canadees webportal met DDC en UDC. Anders dan DDC en UDC zijn de beide web-classificaties op hun beginniveau niet gebaseerd op een indeling van de wetenschap, maar veeleer op een pragmatisch mengsel van thema's, disciplines,

informatiesoorten en interessegebieden. Daarmee wordt lang niet meer voldaan aan eisen als "eenheid van verdelingskarakteristiek" of "onderling uitsluitend zijn van categorieën".

Om gebruikers beter de weg te laten vinden in op classificaties gebaseerde systemen, worden ook wel visualisatietechnieken toegepast. Vooral de OpenDirectory web-taxonomie wordt nogal eens gebruikt om dergelijke systemen te illustreren, zoals met Webbrain. Beagle (2003) heeft dat ook gedaan voor een aanpassing van LCC, bij Belmont Abbey College (Canada). Ook hij combineert dit met een zoekmogelijkheid als alternatieve methode om gebruikers bij de juiste rubriek te laten uitkomen.

In het kader van webtoepassingen zijn ook concordanties van belang - in de praktijk met name tussen DDC en LCC. Daarmee wordt het mogelijk om vanuit verschillende indelingsschema's dezelfde informatiecollectie, of vanuit een zelfde indelingsschema verschillende informatiecollecties te raadplegen. Hoewel deze problematiek van interoperabiliteit van ontsluitingssystemen op dit moment vooral aandacht krijgt vanuit de idee dat collecties van meer organisaties in één keer doorzocht moeten kunnen worden, speelt dit uiteraard ook binnen grote bibliotheken waar verschillende deelcollecties volgens verschillende systemen worden ontsloten of waar vanaf een bepaald tijdstip op een ander systeem is/wordt overgegaan. In hoofdstuk 5 wordt hierop uitgebreider teruggekomen.

Nog los van eventuele webtoepassingen, bestaat er voor classificaties ook een tendens facetsystemen toe te passen. Dat is juist in UDC veel makkelijker te realiseren, dan in de sterk pre-coördinatieve LCC. Voor UDC bestaan standaard algoritmes om apart op de verschillende facetcomponenten te kunnen zoeken. Onder meer bij de Universiteit van Leuven wordt met behulp van trefwoorden via een soort verfijnde concordantie op UDC-codes gezocht. Toch beschrijft Pollitt (1998) hoe juist DDC aangepast kan worden tot een facetclassificatie voor toegang tot online informatie. Anderson (2006) geeft aan hoe een volledige gefacetteerde BLISS-classificatie op al toegekende LCSH headings gebaseerd kan worden.

In bedrijfstoepassingen, waar het vaak een veel breder informatieaanbod betreft dan alleen dat van de bibliotheek, komen onder de noemer van taxonomieën eveneens facetachtige toepassingen tot ontwikkeling (Cheung 2005). Voor verschillende invalshoeken waaronder informatie toegankelijk gemaakt kan worden, worden dan afzonderlijke, meestal vrij eenvoudige classificaties of taxonomieën ontwikkeld, waarin onafhankelijk gebrowseed kan worden. Als resultaat wordt telkens de doorsnede van de geselecteerde, als verschillende dimensies te beschouwen categorieën getoond.

Ook vanuit het oogpunt van het ontwerpen van gebruiksvriendelijke zoekinterfaces wordt een facetaanpak aangeraden (Hearst 2002, Yee 2003). In het Flamenco project werd daarbij de nadruk gelegd op het vindbaar maken van beeldmateriaal. Daarbij werd zorg gedragen dat:

- bij het zoeken te combineren facetten echt verschillende invalshoeken vormen,
- voor elk van die facetten een hiërarchische bladerstructuur wordt aangeboden,
- gebruikers tijdens het zoekproces steeds alleen de nog beschikbare keuzes en te verwachten resultaatantallen gepresenteerd krijgen.

In zoeksystemen wordt deze functionaliteit nu vaak "parametric search" genoemd.

1.1.2 Thesauri

Als hulpmiddel voor betere zoekresultaten, hebben thesauri als belangrijke voordelen:

- dat ze te gebruiken zoekwoorden formaliseren, zodat polysemie (homoniemen) geen aanleiding kan geven tot verlies aan precisie, en anderzijds geen gebruik van synonieme zoekwoorden nodig is om toch voldoende hoge recall te verkrijgen,
- dat ze hiërarchische relaties tussen begrippen toestaan, zodat in principe generieke zoekacties kunnen worden uitgevoerd,
- dat ze meestal als postcoördinatief ontsluitingssysteem worden toegepast, hetgeen de flexibiliteit van het zoekproces bevordert (al heeft dat ook bekende nadelige invloed op de precisie - zogenaamde "false coordination").

De mate van postcoördinatie of "facettering" die in thesauri wordt toegepast is meestal minder extreem dan die van het Uniterm-systeem. De gefaceteerde toepassing van LCSH in FAST bij OCLC (O'Neill 2001, 2003, Dean 2004) blijkt zich in de praktijk vooral te beperken tot een uitsplitsing in onderwerps-, geografische, periode- en vorm-attributen die corresponderen met afzonderlijke Dublin Core velden (Chan 2001). Een ander voorbeeld is het FACET-project (Binding 2004, Tudhope 2002).

Voor informatiezoekers hebben thesauri als belangrijkste nadeel dat ze niet erg gebruiksvriendelijk zijn. Hoe weet de zoeker immers welke termen de "juiste" zijn om mee te zoeken? Omdat thesauri vrijwel nooit - zoals classificaties - vanuit een beperkt aantal hoofdcategorieën zijn opgebouwd, lenen ze zich ook minder goed om de gebruiker via een browse-proces naar de juiste termen te leiden. Om de gebruiker niettemin bij de juiste termen te brengen of het systeem ongemerkt automatisch op de juiste termen te laten zoeken, worden in de praktijk diverse methoden toegepast.

Een voor de hand liggende methode is om elke thesaurusterm te associëren met een voldoende groot aantal synonieme of quasi-synonieme woorden, zodat bij het gebruik van enig van die woorden automatisch de juiste thesaurusterm in de zoekactie gesubstitueerd kan worden. Dit associëren kan onder meer gebeuren:

- doordat een groot aantal synoniemen en nauw verwante woorden en begrippen als verwijstermen in de thesaurus is opgenomen; in PubMed worden MeSH-termen die met ingetikte zoekwoorden corresponderen, zo automatisch aan zoekvragen toegevoegd (Jacsó 2004); bij zo'n zogenaamde "user-thesaurus", ook genoemd in Greenberg (2004), kunnen die termen ook verwijzen naar AND-combinaties van descriptoren (bij complexe begrippen) of OR-combinaties (bij homografen);
- door de door gebruikers in de loop der tijd gebezigde zoekwoorden te analyseren, bijvoorbeeld via logfiles, en die woorden zo goed mogelijk af te beelden op de termen uit de thesaurus; zo kan de mapping tussen gebruikerstaal en systeemtaal geleidelijk steeds vollediger worden; in feite leidt dit ook tot zo'n user-thesaurus;
- door termen uit de thesaurus af te beelden op een semantisch netwerk, waarin het gehele woordenboek van de gebruikte taal is opgenomen, al dan niet aangevuld met domeinspecifiek vocabulaire.

Vanuit het oogpunt van gebruiksgemak lijkt het in eerste instantie raadzaam dit soort vervangingen zo automatisch mogelijk te laten plaats vinden en de thesaurus zelf zo veel mogelijk aan het oog van de gebruiker te onttrekken.

In het FACET-project worden deze technieken toegepast voor het matchen van zoekvragen met thesaurustermen op het terrein van kunst en architectuur (Tudhope 2002, Binding 2004). In een project van Buckland (1999) werd dit met diverse

gespecialiseerde thesauri gedaan. Aan al deze methoden kleeft overigens het bezwaar dat woorden en begrippen maar zelden volledig synoniem zijn, met identieke betekenis, gevoelswaarde, breedte enzovoort. Bovendien kan niet makkelijk rekening worden gehouden met verschillende betekenissen die een door een zoeker gebruikte term kan hebben. Als dat toch gepoogd wordt met behulp van OR-relaties tussen descriptorren die met verschillende betekenissen van de oorspronkelijke zoekterm corresponderen, dan kan dat al snel leiden tot een onaanvaardbaar slechte precisie van de zoekresultaten. Geheel automatische vervanging blijkt in de praktijk dus niet altijd gewenst. Methoden van gebruikersinteractie in zulke situaties komen in §1.2 nog aan de orde. In een overigens niet zeer representatief onderzoek door Greenberg (2004), bleken gebruikers van systemen die zoekwoorden met thesaurustermen associëren ook in meerderheid eigen keuzevrijheid te prefereren. Men wilde zelf termen uit keuzelijstjes kunnen selecteren.

Wanneer ook titels, samenvattingen of zelfs nog meer niet-geformaliseerde tekst digitaal beschikbaar is en dus doorzocht kan worden, is ook een andere methode mogelijk. Van zoekresultaten die zijn verkregen op basis van de oorspronkelijke zoekvraag van de zoeker, kan de tekst statistisch geanalyseerd worden, om te zien welke thesaurustermen daarin (toevallig) al aanwezig waren. Daarna kan de zoekactie - al dan niet automatisch - op basis van die termen worden geherformuleerd. Aan deze methode kleeft het bezwaar dat dergelijke statistische analyses lang niet altijd (uitsluitend) de voor de oorspronkelijke informatievraag juiste of meest relevante thesaurustermen opleveren. In dergelijke gevallen lijkt het zeker gewenst de gebruiker zelf te laten kiezen welke van de zo afgeleide thesaurustermen uiteindelijk in de zoekactie gebruikt moeten worden.

In zoeksystemen wordt de hiërarchie van de thesaurus ook steeds vaker gebruikt om een vollediger opbrengst (hogere recall) van zoekacties te bereiken. Veel gebruikers zijn zich namelijk nauwelijks bewust van het feit dat het zoeken op specifiekere begrippen vrijwel altijd veel meer resultaten oplevert dan zoeken op meer algemene begrippen, zeker in collecties met veel gespecialiseerd materiaal, zoals wetenschappelijke artikelen en hoofdstukken uit monografieën. Greenberg (2001a,b) bevestigde dat automatische "query expansion" met specifiekere termen (NT) en synoniemen uit een thesaurus tot betere recall leidde, zonder verslechtering van de precisie. Bij query-expansie met ruimere en verwante begrippen (BT en RT) ging dit wel ten koste van de precisie. In dat geval kan de gebruiker beter zelf de keuze worden gelaten welke termen aan de zoekvraag toe te voegen. Dergelijk automatisch uitgevoerde generieke zoekacties hebben overigens alleen zin als de betreffende thesaurus over het hele dekkingsgebied een voldoende strikte en systematisch hiërarchische opbouw kent.

Een andere toepassing is altijd heel uitgebreid met gerelateerde termen te expanderen en op het zoekresultaat relevance ranking toe te passen, gebaseerd op de zogenaamde "conceptuele afstand" tussen de aan de documenten toegekende termen en de oorspronkelijke zoekwoorden (Tudhope 2006a). Bij een andere toepassing wordt gekeken hoeveel en welke termen in gevonden documenten voorkomen, die in de thesaurus een relatie (NT, BT, RT) hebben met de oorspronkelijke zoekterm. Door hierop relevance ranking te baseren, werd in een specialistisch domein een aanzienlijke verbetering van de precisie van zoekacties bereikt (Silveira 2004).

1.1.3 Ontologieën

Vooraf in het kader van het semantisch web, wordt gepropageerd kennisrepresentaties in de vorm van ontologieën op te zetten. Die beogen op geformaliseerde en door computers interpreteerbare wijze 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 deze 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. Door een zoekstelsel ook gebruik te laten maken van in een ontologie gedefinieerde rollen, eigenschappen en relaties, zou toepassing daarvan vooral tot hogere precisie moeten leiden, al wordt ook wel van hogere recall gesproken. Hard bewijs voor verbeterde zoekresultaten bestaat echter nog niet. De in ontologieën vastgelegde regels staan de computer in principe ook toe logische gevolgtrekkingen te maken.

De ontologieën waar deze definitie eigenlijk betrekking op heeft, worden meestal voor vrij beperkte onderwerpsdomeinen opgezet. Vanuit onderzoek aan kunstmatige intelligentie worden onder de naam van ontologieën echter ook veel algemener kennisrepresentaties opgezet (Patel 2005). 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. Sommigen beschouwen ook het Wordnet semantisch netwerk als een voorbeeld van zo'n "upper ontology". Een al wat meer gespecialiseerde tussenvorm zijn de "core ontologies" die een specifiek, maar breed toepassingsgebied proberen te beschrijven. Een voorbeeld is het CIDOC CRM (Conceptual Reference Model). Dat bevat definities en een formele structuur voor de beschrijving van concepten, hun relaties en de eigenschappen daarvan, op het terrein van cultureel erfgoed (Doerr 2003). Dit soort ontologieën wordt vooral opgezet om het mogelijk te maken op heterogene wijze ontsloten materiaal 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 dus een rol spelen om interoperabiliteit van informatiesystemen te faciliteren. (Zie ook hoofdstuk 5).

Een voorbeeld van het omwerken van een bestaande thesaurus naar een ontologie wordt beschreven door Soergel (2004) voor de thesaurus van de FAO. Hij bespreekt ook tekortkomingen van klassieke thesauri. Over ontologieën die hij "semantically rich knowledge organization systems" noemt, heeft Soergel hooggestemde verwachtingen. Hij verwacht dat die uiteindelijk zullen leiden tot:

- betere gebruikersinteractie en daardoor betere queries,
- automatische intelligente query-expansie,
- intelligente ondersteuning van menselijke indexeerders en verbeterde automatische indexerings,
- ondersteuning van semantisch web toepassingen.

Dat laatste punt heeft dus ook weer betrekking op de rol van ontologieën bij de interoperabiliteit van informatiesystemen.

1.1.4 Topic maps

Vanuit de XML-wereld zijn "topic maps" geïntroduceerd als middel om informatie op een intelligente manier toegankelijk te maken. De "topics" uit de Topic Maps representeren concepten, wat willekeurige onderwerpen kunnen zijn. Voor het beschrijven van die onderwerpen wordt gebruik gemaakt van de volgende vier begrippen: "names", "types", "occurrences" en "associations" (Garshol 2004).

"Names" bieden de mogelijkheid topics te beschrijven met alle woorden die daarvoor maar in aanmerking komen, dus ook alle synoniemen. Anderzijds mogen "names" (gewoon woorden) ook in verschillende betekenissen worden gebruikt. Door de daaraan gerelateerde *types*, *occurrences* en *associations* worden die betekenissen namelijk toch wel weer onderscheiden. Bij elke "name" kan via topics een "scope" worden aangegeven, de context waarbinnen die "name" betekenis heeft. Zo kunnen bijvoorbeeld culturele verschillen worden opgevangen die bestaan in het gebruik van bepaald vocabulaire tussen verschillende bedrijven of regio's. "Types" kunnen worden gebruikt om topics nader te typeren, bijvoorbeeld om aan te geven dat het topic "Engels" *een taal* betreft of dat het topic "Google" *een zoekmachine* is. Daarbij zijn die types zelf ook weer (vrij te definiëren) topics.

"Occurrences" relateren topics aan informatie waarvoor die relevant zijn. Dat is dus andersom dan bij traditionele ontsluiting. Daar wordt bij een document aangegeven *waarover* dat gaat; hier wordt bij een onderwerp aangegeven welke informatie *daarover* gaat. Topics staan dus niet los van de content. Occurrences worden ook weer getypeerd, wat wil zeggen dat de aard van de relatie met de informatie wordt aangegeven. Bijvoorbeeld dat het gaat om een portret of het telefoonnummer van de persoon waar een topic betrekking op heeft, of dat het een tutorial van een betreffend onderwerp is. Deze soorten occurrences - in feite ook types - zijn zelf ook weer topics.

"Associations" specificeren relaties tussen de onderwerpen. Die associations kunnen ook weer getypeerd worden, waardoor ze veel meer semantische inhoud kunnen hebben dan de relaties tussen termen in traditionele ontsluitingssystemen als thesauri en classificaties. Het resultaat van dit alles is uiteindelijk geen boomstructuur, maar een netwerk van "subjects" waartussen scherp gedefinieerde relaties. Afwijkend is ook dat de ontsloten informatie zelf onderdeel uitmaakt van de Topic Map. Sterker nog, de te ontsluiten objecten kunnen zelf weer als topics gedefinieerd worden, zodat ze ook met names, associations, types en occurrences beschreven kunnen worden. In feite is onze eigen verbeelding de enige beperking bij het opzetten van Topic Maps.

Garshol (2004) geeft een vergelijking tussen topic maps en thesauri, taxonomieën en ontologieën. Hij laat daarin ook zien hoe je traditionele ontsluitingssystemen in de vorm van Topic Maps kunt representeren. Toepassing van Topic Maps lijkt zich nog te beperken tot specialistische domeinen. Hoewel Topic Maps voor verschillende onderwerpen in principe koppelbaar zouden moeten zijn, is niet te verwachten dat dit al op korte termijn tot toepassingen voor zeer brede informatiecollecties zal leiden.

Een experiment met Topic Maps bij de University of Michigan (Rothman 2002), om een op "user context" gebaseerde ontsluiting van met LCC ontsloten materiaal - wel breed - te realiseren, bleek veel te arbeidsintensief te zijn. Bij OCLC is al in 1999 geëxperimenteerd om voor web-bronnen automatisch topic-maps te laten genereren (Godby 2002), maar ook dit lijkt nog niet tot concrete resultaten geleid te hebben.

1.1.5 “Tagging” en folksonomies

Op internet zijn de afgelopen jaren onder de benaming “web 2.0” talloze nieuwe diensten en technieken ontwikkeld, met als belangrijk gemeenschappelijk kenmerk dat gebruikers het op allerlei terreinen zelf voor het zeggen krijgen. Inhoudelijke ontsluiting is één van die terreinen. Vrijwel elke nieuwe dienst biedt gebruikers namelijk de mogelijkheid zogenaamde “tags” toe te voegen. Dat kan bij eigen materiaal dat op internet wordt aangeboden, zoals afleveringen van de eigen weblog, op Flickr neergezette foto’s of bij YouTube aangeboden videomateriaal. Toevoegen van tags kan ook bij al op internet aanwezig materiaal, bijvoorbeeld bij (openbare) bookmarks die op Del.icio.us, Furl, CiteULike of Connotea worden toegevoegd of bij nieuwsberichten die op Digg worden aangemeld (Hammond 2005, Lund 2005).

In feite zijn deze tags gewoon ongecontroleerde trefwoorden. Ze hebben dan ook last van alle nadelen die aan ongecontroleerde ontsluitingsmethoden kleven (Macgregor 2006). Ze hebben echter het grote voordeel dat ze berusten op het woordgebruik van de informatiegebruikers zelf. Ontsluitingssystemen met tags worden - naar analogie van het woord “taxonomy” - ook wel aangeduid als “folksonomies”, door mensen zelf (“folks”) opgezette taxonomieën. In het Engelse taalgebied wordt ook wel van “grass-roots” taxonomieën gesproken. De gebruiker wordt daarbij geacht zelf het best te weten waarop hij voor zijn onderwerp moet zoeken. Bovendien wordt tagging gezien als hulpmiddel voor het samenwerken van mensen, ook een belangrijk web-2.0 kenmerk. Men spreekt in dit verband daarom ook wel van “social software”. Het taggen van materiaal kan dan helpen virtuele gemeenschappen te laten ontstaan van mensen met gelijke interesse en daardoor hopelijk gelijksoortig woordgebruik.

Dit wijst er al op dat tagging in feite voor een ander soort toegankelijkheid bedoeld is, dan klassieke ontsluiting. Dat ze tot slechte recall en precisie leiden, speelt hierbij een veel minder belangrijke rol dan bij “echt” zoeken. Het is meer bedoeld voor browsen en men gaat meer af op suggesties die door collega’s en vrienden zijn gedaan door middel van tags, dan dat men op eigen initiatief diepgravend naar informatie zoekt (Sterling 2005). Tag-clouds laten bij dergelijke systemen ook grafisch zien welke tags het meest frequent en het meest recent gebruikt worden, hoe groter de letters hoe populairder de woorden zijn.

Onderzoek aan tagging beperkt zich op dit moment nog vooral tot kwaliteitsanalyses. In hoeverre eenduidige termen worden gebruikt, in hoeverre er problemen zijn met synoniemie, in hoeverre er voldoende tags worden gebruikt om alle aspecten van een document, een foto of een video te beschrijven, in hoeverre de tags vooral bedoeld zijn voor persoonlijk (her)gebruik en niet gericht op publiekelijke bruikbaarheid, et cetera (Macgregor 2006, Golder 2006). Vaak wordt geconcludeerd dat tagging meer een aanvulling is op klassieke ontsluiting, dan een vervanging daarvan (Guy 2006). Ook als voor gecontroleerde ontsluiting al helemaal geen menskracht beschikbaar is, kan dit een zinnig alternatief zijn. Er is op dit moment nog weinig onderzoek waarbij men tracht een brug te slaan naar gecontroleerde ontsluitingssystemen. Zoals de in §1.1.2 genoemde “user-thesaurus” de zoekwoorden van gebruikers kan koppelen aan officiële thesaurustermen waarmee documenten zijn ontsloten, zo zouden voor ontsluiting gebruikte ongecontroleerde tags andersom ook weer aan thesaurustermen gekoppeld kunnen worden. Het is ook alleszins interessant te zien in hoeverre tagging in klassieke situaties van bibliotheekcatalogi een toegevoegde waarde kan hebben.

1.2 Interactie tussen gebruiker en ontsluitingssysteem

Zoals eerder aangegeven, is de discrepantie tussen het door gebruikers in hun zoekvragen gehanteerde vocabulaire en dat van het systeem een belangrijke hinderpaal bij het zoeken met behulp van gecontroleerde ontsluitingssystemen. Dat is niet alleen bij woordsysteem zoals thesauri het geval, maar ook bij classificaties / taxonomieën. Bij die laatste wordt het echter minder gevoeld, door de meer passieve wijze van gebruik van de meeste van dergelijke systemen - de gebruiker kiest uit de gepresenteerde rubrieksomschrijvingen die welke het best lijkt te corresponderen met het gewenste onderwerp. Daarvoor is het natuurlijk wel essentieel dat voor de presentatie van classificaties niet van codes gebruik wordt gemaakt, maar van zo duidelijk mogelijke omschrijvingen in woorden. En ook dan nog zo goed mogelijk aansluitend op het woordgebruik van de gebruikers van het systeem. Daarnaast spelen klassieke kwaliteitseisen voor classificaties, zoals eenheid van verdelingskarakteristiek, co-extensie van een rubriek met haar subrubrieken en orthogonaliteit van aangeboden (sub)rubrieken, nog altijd een belangrijke rol om gebruikers makkelijk de juiste categorie te laten vinden. Dat specifieke web-classificaties daar overigens lang niet altijd aan voldoen, liet Hudon (2001) zien. Anderzijds kunnen zoeksystemen ook behulpzaam zijn om - zeker in systemen met zeer veel rubrieken - naar de juiste (sub)rubriek te leiden.

Bij de bespreking van thesauri werd al aangegeven dat er verschillende methoden zijn om gebruikers zonder kennis van het specifieke vocabulaire van een thesaurus, op basis van de door hen ingetikte zoektermen, toch automatisch bij de juiste thesaurustermen te laten uitkomen. Dit kan ook wat minder automatisch. Op basis van de aanvankelijke zoekvraag, kan een zoekstelsel met eenvoudige statistiek de meest in het antwoord voorkomende thesaurustermen ter keuze aan de gebruiker voorleggen. Het blijkt dat veel gebruikers dan geneigd zijn om daarbij gepresenteerde narrower en related terms ook aan hun zoekvraag toe te voegen om de opbrengst te verbeteren (Shiri 2006).

De wereld van de free-text retrieval kent een onderzoekstak die zich richt op dialoogsystemen, om de computer er achter te laten komen wat de gebruiker precies bedoelt. Dergelijke dialogen kunnen ook gebruikt worden om meer zekerheid te krijgen dat de zoektermen van de gebruiker inderdaad aan de juiste thesaurustermen gekoppeld worden. Dat kan zeker nuttig zijn om in gevallen van ambiguïteit uitsluitel te krijgen, welke van de door het systeem geassocieerde thesaurustermen voor deze gebruiker de juiste is. In geval van het ontbreken van een voldoende waarschijnlijke kandidaat, kan echter ook om aanvullende informatie gevraagd worden opdat het systeem alsnog geschikte termen kan vinden (Kelly 2007).

Anderzijds worden ook experimenten gedaan om vragen van gebruikers op dezelfde manier automatisch te classificeren - aan een trefwoord of klasse uit een taxonomie te koppelen - als dat met documenten gebeurt (zie hoofdstuk 2). Omdat een zoekvraag over het algemeen veel minder tekst bevat dan een document, zal de zoeker ook hierbij meestal via een dialoogstelsel om aanvullende informatie gevraagd moeten worden.

1.3 Praktische overwegingen

Hoewel veel van de in dit hoofdstuk genoemde voorbeelden toepassingen zijn waarin vooral web-informatie wordt ontsloten, zijn de beschreven methoden veel algemener toepasbaar. De voornaamste aanpassingen ten opzichte van klassieke versies van de betreffende classificaties hebben namelijk betrekking op de browse-baarheid van classificaties in een web-interface en niet specifiek op de aard van het ontsloten materiaal. Die aard van het materiaal bepaalt echter wel welke methode van ontsluiting het meest geschikt is voor het inhoudelijk toegankelijk maken van dat materiaal.

Bij de classificaties blijken in internationaal verband DDC en LCC, in al dan niet aangepaste vorm, nog steeds de toonaangevende systemen te zijn. Voornamelijk redenen daarvoor zijn dat ze al zo lang door zoveel bibliotheken worden toegepast, zeker in het Engelse taalgebied, en dat bovendien zoveel boekmateriaal al bij uitgave, dan wel kort daarna, van de betreffende codes wordt voorzien, wat “ontlenen” van die ontsluiting aantrekkelijk maakt. Er is echter ook wel degelijk een beperking. Gespecialiseerde wetenschappelijke artikelen kunnen onvoldoende gedetailleerd inhoudelijk worden gekarakteriseerd met alleen dergelijke systemen, als ze zijn vereenvoudigd ten behoeve van browse-toepassing in een webomgeving. Ook de hoeveelheid te ontsluiten materiaal speelt daarbij een beperkende factor. Bij een collectie van 1 miljoen documenten, zal in een classificatie met 2000 klassen, elke klasse gemiddeld nog 500 documenten bevatten. Dat is uiteraard veel te veel voor browse-toepassingen. Voor dergelijk materiaal kunnen classificaties - en dat geldt niet alleen beide genoemde voorbeelden - dus eigenlijk alleen dienen voor globale voorselectie op de onderwerpsdomeinen waartoe de publicaties behoren.

Voor gedetailleerde inhoudelijke ontsluiting van publicaties over specialistische onderwerpen zijn thesauri wel geschikt. Als daarbij wordt gezorgd voor de in §1.1.2 besproken methoden om eindgebruikers automatisch van die thesauri gebruik te laten maken, biedt dat in principe goede mogelijkheden tot verbeterde toegankelijkheid van materiaal. Probleem daarbij is dat de best uitgewerkte thesauri vrijwel allemaal voor beperkte onderwerpsdomeinen zijn ontwikkeld. In het kader van interoperabiliteit (zie hoofdstuk 5) worden wel verschillende vocabulaires op elkaar gemapt, maar dat betreft meestal algemene vocabulaires, waarvan de onderwerpsgebieden elkaar vrijwel geheel overlappen.

Voor het probleem dat gecontroleerd ontsluiten van informatie duur is, en nog eens te meer voor de grote hoeveelheden vooral digitale documenten waartoe veel organisaties toegang hebben, bestaan meer mogelijke oplossingen. In de eerste plaats kunnen geautomatiseerde methoden van ontsluiting worden toegepast, zoals besproken in hoofdstuk 2. Daarnaast kan samenwerking met andere organisaties een oplossing zijn, een reden te meer om aandacht te besteden aan interoperabiliteit van toegepaste ontsluitingssystemen. Tot slot kan ontsluiting worden uitbesteed aan eindgebruikers via technieken van “social tagging”, al heeft dat als nadeel dat de kwaliteit daarvan voor sommige doeleinden nog te wensen over laat.

1.4 De toekomst van inhoudelijke ontsluiting

Voor klassieke methoden van inhoudelijke ontsluiting lijken geen baanbrekende nieuwe ontwikkelingen op komst te zijn. Wat dat betreft moeten we overigens onderscheid maken tussen enerzijds de technieken voor het opzetten van dergelijke kennisrepresentaties, in de vorm van rubrieksindelingen, taxonomieën of thesauri, en anderzijds het praktische gebruik aan de “voor-“ en “achterkant” van systemen.

Aan de “achterkant” is dat het ontsluitingsproces zelf, waarbij materiaal wordt ingedeeld of ontsloten met behulp van dergelijke systemen. Daar staat wel veel te gebeuren, waarbij een belangrijke rol lijkt weggelegd voor geautomatiseerde methoden om materiaal op basis van taxonomieën of thesauri in te delen of te ontsluiten. In het volgende hoofdstuk wordt daar verder op ingegaan.

Aan de “voorkant”, de kant van de gebruiker waar het zoekproces plaats vindt, worden steeds meer technieken ontwikkeld om voor "gewone" zoekvragen, op de achtergrond, vrijwel onzichtbaar voor de gebruiker, toch gecontroleerde ontsluitings-systemen toe te passen. Dat maakt automatisch een einde aan het alom gesignaleerde ondergebruik van de huidige ontsluitingsystemen. Toepassing van bepaalde vormen van kunstmatige intelligentie (AI) in technieken om te komen tot een betere interpretatie van de informatiebehoefte en ook van de bedoeling van de gebruiker (user intent) kunnen daarbij een rol gaan spelen.

Over acceptatie en toepassing van betrekkelijk nieuwe methoden als ontologieën en topic maps valt nog weinig met zekerheid te zeggen. Er is nog maar een beperkt aantal experimenten op dit terrein. De indruk bestaat dat opzetten en toepassen van dergelijke systemen zeker niet minder arbeidsintensief zal zijn dan bij classificaties en thesauri. Dit aspect zal dus zeker niet van doorslaggevende betekenis zijn voor een zonnige toekomst voor deze systemen.

Het inzetten van het publiek voor het ontsluiten van informatie, gebruikmakend van de “wisdom of the crowds” via tagging of folksonomies, lijkt een interessante ontwikkeling. In hoeverre dat alleen voor informatie op internet een waardevolle bijdrage kan leveren, of dat het ook voor fysieke collecties bij bibliotheken nuttig kan worden, moet nog worden afgewacht. Voor die laatste toepassing zal dan ook nog moeten worden onderzocht in hoeverre controle en eventuele verbetering van zo toegekende trefwoorden nodig, nuttig en haalbaar is.

Ten behoeve van JISC (het Joint Information Systems Committee dat informatie en communicatietechnologie ten behoeve van onderwijs en onderzoek in het Verenigd Koninkrijk ondersteunt) is recent een verkenning uitgevoerd van ontwikkelingen op het gebied van inhoudelijke ontsluiting (Tudhope 2006b). Daarbij komen dezelfde aspecten van inhoudelijke ontsluiting aan de orde als in dit artikel. Nadruk ligt daar echter vooral op de ontwikkeling van zogenaamde “terminology services” die het mogelijk maken om in een netwerkomgeving gebruik te maken van al door andere organisaties ontwikkeld en beschikbaar gesteld gecontroleerd vocabulair. Dit wordt als een belangrijke toekomstige ontwikkeling gezien. Ook in hoofdstuk 5 van dit artikel komen enkele aspecten daarvan onder het kopje “interoperabiliteit” nog nader aan de orde.

2. Geautomatiseerde methoden van inhoudelijke karakterisering, classificatie en verrijking

2.1 Stand van zaken

Er is al enige jaren ervaring met systemen voor het automatisch classificeren van digitaal beschikbare documenten en voor het verrijken van digitale documenten met trefwoorden, thesaurustermen of andere manieren van karakterisering. Anderson (2001, part I) geeft aan dat relatief weinig bekend is van menselijke analyseprocessen die aan handmatige ontsluiting ten grondslag liggen. Al is er een ISO-standaard hoe indexeerders de onderwerpen van een document moeten bepalen, toch spelen in de praktijk veel meer subjectieve elementen mee, dan bij geautomatiseerd classificeren. Dat geautomatiseerde ontsluiting vaak even goed wordt beoordeeld als handmatige, hoeft dus misschien niet eens veel verwondering te wekken.

Geautomatiseerde methoden berusten meestal op het genereren van een zogenaamde vingerafdruk uit een digitaal beschikbaar tekstdocument. Voor het maken van die vingerafdrucken staat een heel arsenaal aan methoden ter beschikking. Een overzicht daarvan wordt onder meer gegeven door Anderson (2001, part II). In feite zijn dat voor een belangrijk deel dezelfde soort technieken die worden toegepast voor verbetering van full-text retrieval, zoals in het volgende hoofdstuk besproken. Globaal kan gezegd worden dat die technieken uiteenvallen in statistische methoden, kennisgebaseerde methoden en linguïstische methoden.

Met statistische methoden - vooral relatieve woordfrequenties - kan worden bepaald wat de inhoudelijk belangrijkste en meest karakteristieke woorden uit een document zijn. Met - meestal handmatig opgestelde - kennisregels kan het belang van bepaalde woorden worden afgeleid, op basis van logische regels, op basis van hun locatie in het document, op basis van aanwezige speciale markering e.d. Met linguïstische methoden kan onder meer een soort normalisatie van aanwezige termen worden bewerkstelligd, door

- reductie van woorden tot hun morfologische woordstam, zodat voor de statistiek geen onderscheid meer wordt gemaakt tussen enkel- en meervoud, vervoegingen, verbuigingen e.d.,
- zogenaamde "decompounding" van samengestelde woorden in hun losse componenten, zodat die in de woordstatistiek kunnen worden meegenomen,
- syntactische analyse, zodat equivalentie tussen zinsdelen en losse woorden kan worden vastgesteld (bijvoorbeeld: "besparing van energie" = "energiebesparing"),
- beschikbaarheid van semantische kennis over mogelijke betekenisrelaties tussen woorden, bijvoorbeeld welke woorden (vrijwel) synoniem zijn.

In de praktijk blijken deze linguïstische technieken niet altijd allemaal te hoeven worden toegepast om al redelijk betrouwbare resultaten te verkrijgen. Welke van het meeste belang zijn, hangt vaak af van de taal van de te karakteriseren documenten. In de praktijk wordt vrijwel altijd een combinatie toegepast van ten minste statistische + regelgebaseerde of statistische + linguïstische techniek.

De zo afgeleide vingerafdrukken, in de vorm van een collectie losse, eventueel van gewichtsfactoren voorziene termen, of taalkundig bewerkt tot lopende zinnen in een computergegenereerde samenvatting, kunnen het eindpunt zijn van de inhoudelijke karakterisering of verrijking van documenten. Gebruik van deze vingerafdrukken voor free-text retrieval, in plaats van de volledige teksten van de documenten, kan namelijk al leiden tot betere recall en precisie van zoekacties. De vingerafdrukken kunnen echter ook worden gebruikt om de tekstdocumenten te matchen met de best gelijkende representaties van beschikbare klassen uit een classificatie/taxonomie of van termen uit een thesaurus. Voor dat doel moet het systeem kunnen beschikken over soortgelijke vingerafdrukken voor alle klassen uit de taxonomie of alle termen uit de thesaurus. Elke klasse of thesaurusterm is dan als het ware verrijkt met een verzameling termen die, elk met hun eigen gewicht, gezamenlijk karakteristiek zijn voor die klasse of thesaurusterm. Men spreekt daarbij wel van equivalentieklassen.

Die aan de klassen of termen gehechte vingerafdrukken kunnen de vorm hebben van een soort kennisregels die geheel door menselijke tussenkomst zijn opgesteld. Daarin kunnen bijvoorbeeld Booleaanse combinaties van voor het betreffende onderwerp karakteristieke woorden en woordstammen verwerkt zijn, waarbij aan die woorden vaak ook gewichtsfactoren gehecht kunnen worden. De meeste moderne systemen bieden echter de mogelijkheid deze vingerafdrukken automatisch tot stand te laten komen door training van het systeem. Dit zal zeker aantrekkelijk zijn, wanneer een collectie digitale documenten beschikbaar is, die al handmatig is ontsloten op basis van de te gebruiken taxonomie of thesaurus. In een dergelijk geval kan het trainingsproces grotendeels geautomatiseerd worden.

Hierbij zullen meestal minimaal tien documenten per klasse als trainingsmateriaal nodig zijn. Elk daarvan dient echter wel voldoende karakteristiek te zijn voor de daarmee te trainen klassen, zodat het trainingsmateriaal enigszins selectief uit de al ontsloten documenten gekozen zal moeten worden. Als 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 1000 klassen vereist al dat 10.000 voldoende karakteristieke documenten worden geselecteerd en handmatig worden ingedeeld of ontsloten.

Daarnaast is het aan te raden met testdocumenten te onderzoeken of het systeem voldoende goed getraind is. Ook die documenten moeten of al ontsloten zijn, of alsnog door mensen worden beoordeeld. De ervaring leert overigens dat bij automatische testruns geconstateerde "fouten" vaker het gevolg zijn van het feit dat de betreffende testdocumenten destijds door menselijke indexeerders onjuist waren ontsloten, dan dat ze nu door het systeem onjuist geclassificeerd werden. Deze observaties zijn overigens vooral gebaseerd op ervaringen met krantenartikelen en bedrijfsinformatie en nog nauwelijks op ervaringen met wetenschappelijke documenten (Van Gent 2002).

De indruk bestaat dat voor wetenschappelijke artikelen classificatie op basis van samenvattingen iets beter werkt dan op basis van de volledige tekst van artikelen. Er worden zelfs experimenten gedaan om te zien of in bepaalde vakgebieden de titels van wetenschappelijke artikelen misschien al zodanig concreet beschrijvend zijn, dat alleen op grond daarvan indeling al mogelijk is.

Over het algemeen zullen voor het classificatieproces drempelwaarden kunnen worden ingesteld voor de waarschijnlijkheid dat een toekenning ook werkelijk betrouwbaar is. Bij een hoge betrouwbaarheidsdrempel zal een hoog percentage van de geclassificeerde documenten - vaak ruim meer dan 90% - in de correcte klasse(n) terecht komen, maar zal ook een flink percentage - soms meer dan 20% - wegens onzekerheid door het systeem terzijde gelegd worden, zodat het alsnog handmatig verwerkt kan worden. Als het automatisch classificeren alleen maar dient als hulpmiddel voor menselijke indexeerders kan de betrouwbaarheidsdrempel lager, en het aantal toe te kennen klassen hoger ingesteld worden. In die situatie dient een dergelijk systeem namelijk alleen om de meest in aanmerking komende klassen aan de menselijke indexeerder te presenteren, zodat die sneller en in de praktijk ook beter, vollediger en consistentere kan ontsluiten, zeker ook in situaties waar het om niet-professionele ontsluiters gaat.

Interessant is ook de mogelijkheid is om automatische classificatie op zoekvragen toe te passen. Daarmee kunnen zoekvragen met thesaurustermen geassocieerd worden. Een vereiste daarvoor is wel dat de gebruiker een voldoende uitgebreide omschrijving van zijn zoekvraag geeft. Een voorbeeld van een dergelijk vraagsysteem voor het zoeken naar milieu-gerelateerde informatie wordt beschreven door Quarles van Ufford (2004). Daarbij kunnen dialoogsystemen nuttig zijn, die de gebruiker stimuleren zoveel mogelijk informatie te verschaffen. In principe kun je de gebruikte thesaurus voor deze toepassing ook trainen met zoekvragen in plaats van met documenten. Deze toepassing beoogt dus ook een oplossing te bieden voor de in §1.1.2 al genoemde discrepantie tussen de belevingswereld van de gebruiker en het binnen een thesaurus vastgelegde voorkeursvocabulaire. Voor al te uitgebreide thesauri lijkt dit echter nog niet te werken.

Wanneer (nog) geen classificatie of thesaurus beschikbaar is, kunnen de hier besproken methoden om “vingerafdrukken” 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 vector-model, kan dan getracht worden de gehele documentcollectie in clusters uiteen te laten vallen. Die clusters kunnen dan beschouwd worden als uit de collectie zelf voortgekomen onderwerpscategorieën. Vaak zal het systeem, op grond van woorden uit de vingerafdrukken van de documenten uit zo'n cluster, ook al een omschrijving voor die categorie kunnen genereren. Experimenten met deze techniek zijn uitgevoerd op zelfs zeer grote informatiecollecties van wel 20 miljoen webpagina's (Golub 2006).

Een recent overzicht van toegepaste technieken en voorbeelden is samengesteld door Golub (2006). Een uitgebreide bibliografie op het terrein van geautomatiseerde ontsluiting is samengesteld door Fabrizio Sebastiani en wordt nu onderhouden door Evgeniy Gabrilovich (Bibliography on Automated Text Categorization; <http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html>). In december 2006 bevatte deze bibliografie verwijzingen naar 577 artikelen, congresbijdragen en andere soorten publicaties.

2.2 Toepassen van geautomatiseerde inhoudelijke ontsluiting

In deze paragraaf beperken we ons tot geautomatiseerde inhoudelijke ontsluiting in de vorm van hetzij classificatie op basis van een bestaand indelingsschema, hetzij op het toekennen van termen die afkomstig zijn uit een gecontroleerd vocabulaire. Om in deze situaties te kunnen ontsluiten, is het een noodzakelijke voorwaarde dat al een classificatie of taxonomie, dan wel een thesaurus of andere collectie gecontroleerde termen, beschikbaar is. Voorts dient het systeem, zoals in de vorige paragraaf beschreven, op alle daarin voorkomende klassen of termen getraind te zijn.

Ervaring van een leverancier leert dat ontsluiting met dergelijke systemen nog goed kan werken voor schema's met maximaal circa 5000 klassen. Nog afgezien van de al in de vorige paragraaf genoemde problematiek van het trainen van zeer grote aantallen klassen, kan het bij grotere aantallen klassen namelijk een probleem worden dat de computer steeds moeilijker onderscheid kan maken tussen verschillende in de praktijk nauw verwante begrippen. Als illustratie hierbij kan ervaring dienen uit een bij de UB Utrecht uitgevoerd experiment. 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 vingerafdrukken voor deze termen - misschien ook niet echt verwonderlijk - onvoldoende onderscheidend. Dit illustreert dat "orthogonaliteit" - niet te veel inhoudelijke overlap - tussen klassen uit de taxonomie of thesaurus ook een eis is voor succesvol automatisch classificeren.

Anderzijds bleek uit jarenlange ervaring met automatisch indexeren ten behoeve van de PHYS database (Engelstalige bibliografische database met fysica-artikelen) 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 aan te passen aan geleidelijk veranderend vocabulaire en zwaartepunten. Ook voor het classificeren van "health resources" op internet met behulp van de MeSH, een zeer uitgebreide medische thesaurus, worden bevredigende resultaten gemeld (Névél 2006). Een techniek die een combinatie vormt tussen echte training van het systeem en het afleiden van termen uit de tekst zelf, is recent ontwikkeld door Medelyan (2006).

In § 1.1.1 was ook sprake van facet taxonomieën. Vooral in bedrijfsomgevingen kunnen daarmee verschillende aspecten van documenten, ook van ongestructureerde informatie als bijvoorbeeld e-mail, als het ware meerdimensionaal gekarakteriseerd worden. Ook daarbij kan automatische classificatie worden toegepast (Cheung 2005).

Van de te ontsluiten informatie dient uiteraard een voldoende hoeveelheid tekst digitaal beschikbaar te zijn. 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 - of waar tekst met het karakter van een samenvatting ontbreekt - zal meer tekst digitaal beschikbaar moeten zijn.

2.3 Praktische voorbeelden

Op het terrein van bedrijfsinformatie en kennissystemen worden vormen van geautomatiseerde inhoudelijke ontsluiting al enige tijd toegepast. Dat is ook het geval voor het classificeren van webbronnen. Bij de voormalige zoekmachine NorthernLight werd bijvoorbeeld gebruik gemaakt van een door bibliothecarissen opgezette classificatie. Die omvatte 16000 onderwerpscategorieën in maximaal negen niveaus. Dit werd nog aangevuld met circa 150 categorieën voor meer formele kenmerken van webpagina's, zoals land-, taal- en genre-aanduidingen. Op basis hiervan werden geïndexeerde webpagina's automatisch geïnclassificeerd, zodat zoekresultaten op basis van deze categorieën uitgesorteerd konden worden (Dumais 2002). Een ander tamelijk grootschalig web-voorbeeld is de Thunderstone webgids. De gespecialiseerde onderwerpsgids CORA (Computer Science Research Paper Search Engine), waarin de inhoud op soortgelijke wijze werd ingedeeld, is helaas niet meer actief.

Toepassingen in een meer klassieke bibliotheekomgeving zijn er ook. Vrij veel daarvan zijn overigens uitgevoerd in de vorm van projecten, waarvan niet altijd duidelijk is of ze uiteindelijk geleid hebben tot in de praktijk toegepaste systemen. Veel van die projecten hebben ook betrekking op de ontsluiting van webpagina's, omdat de te karakteriseren informatie in dat geval gegarandeerd digitaal beschikbaar is. Zo zijn bij het Amerikaanse OCLC zelfs diverse projecten op dit terrein uitgevoerd. Een overzicht van toepassingen waarbij voor de ontsluiting vooral wordt uitgegaan van bestaande bibliotheekclassificaties wordt gegeven door Toth (2002). De belangrijkste projecten die in dat artikel genoemd worden, komen ook in onderstaand overzicht aan de orde. (URL's van de hier vermelde projecten zijn in een bijlage bij dit artikel te vinden).

ARION

Advanced Lightweight Architecture for a Digital Library of Scientific Collections. Het project richt zich op zowel de productie als de toegankelijkheid van wetenschappelijke informatie in een gedistribueerde omgeving. Een kennisbank voor het automatisch toekennen van metadata vormt maar een klein onderdeel van het totale project.

BINDEX

Bilingual Automatic Parallel Indexing and Classification. Hierin zorgde de module AUTINDEX voor automatische indexering en classificatie, tijdens het productieproces van documenten, in zowel Engels als Duits. Hierin werd gebruik gemaakt van bestaande thesauri op de terreinen van techniek en fysica. Looptijd van het project: van 2000 tot 2002.

CARMEN

Content Analysis, Retrieval and MetaData: Effective Networking. Dit project richtte zich op de inhoudsanalyse van heterogeen wetenschappelijk materiaal in een gedistribueerde omgeving. In dat kader werd ook aandacht besteed aan de interoperabiliteit van verschillende systemen (zie hoofdstuk 5). Looptijd van het project: van 1999 tot 2002.

DESIRE

Development of a European Service for Information on Research and Education. Richtte zich op verbeterde toegang en uitwisseling van onderzoeksinformatie in een netwerkomgeving. Automatische classificatie maakte daar deel van uit. Hierin werd samengewerkt met OCLC. Via de harvester COMBINE was er ook een relatie met EEL(S). Het project liep van 1998 tot 2000.

EEL(S)

Engineering Electronic Library, Sweden. Een informatiesysteem voor op kwaliteit beoordeelde internetbronnen op het terrein van technische wetenschappen. Het systeem maakt gebruik van de Engineering Index classificatie. COMBINE was de harvester om op te nemen informatie te oogsten (OAI-PMH bestond toen nog niet) en volgens de EI-classificatie in te delen (Ardö 1999). Het project liep van 1994 tot 1999. Het systeem is nog in productie (Lindholm 2003).

GERHARD

Project van de Carl von Ossietzky Universität Oldenburg voor het opzetten van een Duitse zoekmachine voor wetenschappelijke internetbronnen, waarbij ook automatische classificatie op basis van UDC (3-talig; 60.000 categorieën) wordt toegepast (Wätjen 1998). De eerste fase van het project werd in 1998 afgesloten. Op grond van de daarin opgedane ervaringen is in 2001-2003 een tweede fase uitgevoerd. Daarin werden onder meer verbeterde linguïstische analyses toegepast, hetgeen tot een betrouwbaarder toekenning van classificatiecodes moest leiden (Nigg 2004).

MEANING

In het MEANING-project wordt specifiek gekeken naar de toegevoegde waarde van linguïstische technieken voor allerlei soorten meertalige toepassingen met nadruk op zowel automatische classificatie als gewone retrieval. Onderscheiden van woordbetekenissen is daar een belangrijk aspect van. In één van de deelprojecten worden documenten uit een corpus van Reuters-artikelen verrijkt met relevante termen uit een meertalig semantisch netwerk. Bij de analyse van de behaalde resultaten wordt ook gebruik gemaakt van de begrippen recall en precisie om aan te geven hoeveel van de voor het document relevante termen ook echt zijn toegekend en hoeveel van de eraan toegekende termen inhoudelijk correct (relevant) zijn. Vossen (2004) rapporteerde dat door toepassing van linguïstische technieken de zo gedefinieerde recall toenam van 67% tot 80%.

OCLC (CORC, FAST, Scorpion, Wordsmith)

Bij OCLC hebben verschillende onderling gerelateerde projecten gelopen die intussen deels in praktische toepassingen verwerkt zijn. Hoewel OCLC deze projecten in de praktijk vooral richtte op het catalogiseren en ontsluiten van webbronnen, zijn ze als methodiek van algemener belang. Zolang documenten maar digitaal beschikbaar zijn, kunnen dergelijke technieken immers op elk soort materiaal worden toegepast.

Het CORC-project was gericht op het realiseren van een "Cooperative Online Resource Catalog", een collectief opgebouwde catalogus van vooral internetbronnen. Dit in 1999 afgesloten project is overgegaan in het reguliere OCLC-product Connexion. Voor ondersteuning van het catalogiseerproces in Connexion zijn projecten voor automatische indexering uitgevoerd.

Scorpion beoogde hulpmiddelen - ook software - te ontwikkelen voor automatische onderwerpsherkenning op basis van bekende classificaties als DDC (Godby 1998, Shafer 2001). Daarbij werd een combinatie van statistische en linguïstische technieken toegepast. Nieuwe documenten vormden zoekacties in een vector-representatie tegen een database met beschrijvingen van Dewey-categorieën. Scorpion als project werd in 2000 afgesloten.

Wordsmith was een project om uit digitale teksten een beperkt aantal voor de inhoud belangrijke woorden en onderwerpszinnetjes te destilleren (Godby 2001b). Ook Wordsmith werd in 2000 afgesloten. Resultaten van beide projecten lijken intussen te worden toegepast in Connexion.

In FAST, "Faceted Application of Subject Terminology", wordt getracht een vereenvoudigde syntax en regelgeving te ontwikkelen voor toepassing van het tamelijk complexe LCSH-vocabulaire. Daarbij richt men zich onder meer op het ontleden van LCSH-vocabulaire in de samenstellende facetten. De hierbij opgebouwde FAST-database wordt op zijn beurt weer gebruikt als kennisbank voor het ondersteunen van het proces van automatisch classificeren (Godby 2001a,c, Chan 2001, Dean 2004).

PEKING

Deels door de Universiteit Nijmegen uitgevoerd project voor "supervised and unsupervised classification and (cross-lingual) matching of documents in organizations". Op de website (<http://www.cs.kun.nl/peking/>) zijn diverse (technische) artikelen over het project te vinden, o.a. Bel (2003) en Koster (2003a,b). Het project is afgesloten in 2003.

Thunderstone

Een automatisch gevulde webgids van het type Yahoo of OpenDirectory, met tien hoofdrubrieken en daaronder tot vier niveaus met specifiekere categorieën. Sinds 1998 analyseert het systeem individuele webpagina's uit de .COM, .NET en .ORG domeinen, op grond waarvan de betreffende websites als geheel aan onderwerps-categorieën worden toegekend. De laatste jaren lijken geen vernieuwingen meer te zijn doorgevoerd in de werking van het systeem.

2.4 De toekomst van geautomatiseerde ontsluiting

Bij de afweging tussen handmatige en automatische ontsluiting speelt een aantal factoren mee:

- de stijgende kosten van handmatige indexering,
- de groeiende hoeveelheid te ontsluiten materiaal,
- de - in de toekomst naar verwachting nog verbeterende - kwaliteit van automatische indexering, en
- de vraag of 100% correcte ontsluiting voor al het ontsloten materiaal even belangrijk is.

In relatie tot het laatste punt wil ik graag citeren uit een artikel van Anderson (2001, part II):

It is clear from research and the experience of users that automatic machine-based indexing and human intellectual analysis-based indexing both make important, but very different, contributions to successful information retrieval. At the same time, expert human indexing keeps getting more expensive, while automatic indexing becomes, comparatively, less and less expensive and more effective. Therefore, it seems likely that future IR databases will seek to maximize benefits by allocating human analysis and indexing to situations where the benefits of human expertise are most apparent and immediate.

In order to improve the effectiveness and efficiency of the information retrieval enterprise, librarians, database producers, and other information professionals need to stop treating every document as if all documents, all texts, and all messages were equally important. We know this is not the case. We need to be more judgmental and discriminating, in the best sense of these terms. We all learn about the so-called "80-20 rule" that suggests that in any large collection of documents, 20% will get 80% of the use, or, to put it differently, 20% of the documents will answer 80% of the questions, or respond to 80% of the needs or desires of users. To allocate human analysis expertise in a rational, cost-effective manner, we need to develop methods for predicting the more important documents and devoting human analysis to them. All documents can receive inexpensive, relatively effective automatic, machine-based analysis and indexing. For important documents, automatic indexing can be augmented by human indexing, to make these documents even more accessible to a broader clientele.

Dergelijke afwegingen, of alle documenten altijd 100% correct ontsloten moeten zijn, of dat een 80-20 regel acceptabel is, zullen 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 wel vereist zijn. Dit zijn wel zaken die moeten worden afgewogen voordat aan geautomatiseerde ontsluiting wordt begonnen.

Ook moet worden bedacht dat, wanneer uitsluitend geautomatiseerde technieken worden toegepast, alleen nog digitaal beschikbaar materiaal wordt ontsloten en niet-digitaal materiaal inhoudelijk onvindbaar wordt. Anderzijds is de verwachting dat van steeds meer fysieke documenten een voldoende hoeveelheid inhoudelijke tekst - bij boeken: inhoudsopgave, samenvatting, flaptekst - digitaal beschikbaar komt, om geautomatiseerde ontsluiting daarop te baseren.

3. Free-text retrieval in plaats van ontsluiting?

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. Nu een steeds groter deel van de inhoud van documenten, artikelen, rapporten en boeken op zijn minst gedeeltelijk digitaal beschikbaar is en dus met zoekmachine- of retrievalsoftware doorzocht kan worden, bestaat bij veel mensen de verwachting dat zo in principe alles terug te vinden is. Men zou dit het Google-effect kunnen noemen. De vraag is dan ook gerechtvaardigd of de zo geboden zoekfunctionaliteit het al mogelijk maakt om in elke situatie de gewenste informatie met voldoende recall en precisie terug te vinden.

Met de eerste generatie free-text 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 recall en precisie.

Enkele vaak voorkomende oorzaken dat op deze manier relevante informatie wordt gemist:

- in documenten kunnen allerlei andere woordvormen (enkel-/meervoud, vervoegingen, verbuigingen) voorkomen dan het exacte woord uit de zoekvraag,
- in documenten kunnen woorden in verschillende spelling voorkomen,
- in documenten kunnen synoniemen van gebruikte zoekwoorden voorkomen,
- in documenten kunnen specifiekere woorden voorkomen dan waarop gezocht is.

Enkele vaak voorkomende oorzaken dat op deze manier niet-relevante informatie wordt gevonden:

- gebruikte zoekwoorden kunnen ook in andere betekenissen voorkomen,
- gebruikte zoekwoorden kunnen in verschillende contexten voorkomen,
- dat twee woorden samen in een tekst voorkomen, is nog geen garantie dat ze daarin ook de inhoudelijke relatie hebben die de zoekvraag impliceerde,
- in lange teksten kunnen veel woorden voorkomen die weinig te maken hebben met het werkelijke onderwerp van het document.

Veel van die problemen spelen in principe veel minder als bij het zoeken gebruik gemaakt kan worden van gecontroleerd vocabulair. Als voornaamste doel hebben die immers om terminologie eenduidig te maken, te uniformeren, te standaardiseren en relaties te leggen tussen begrippen en tussen onderwerpen.

In nieuwe generaties free-text retrievalsoftware worden echter technieken toegepast die tot doel hebben op zijn minst een gedeeltelijke oplossing voor dit soort problemen te bieden. Dat rechtvaardigt een afweging tussen free-text retrieval en gecontroleerde ontsluiting.

3.1 Stand van zaken

Mede onder invloed van de TREC-competitie (Text REtrieval Conference) en CLEF (Cross-Language Evaluation Forum) voor multilinguale retrieval, is de kwaliteit van free-text retrieval, gemeten in termen van recall en precisie, het laatste decennium aanzienlijk verbeterd. Diverse technieken die veelal gezamenlijk worden toegepast, hebben daaraan bijgedragen. Veel van die technieken blijken taalafhankelijk en soms ook contextafhankelijk, zowel in hun precieze wijze van toepassing, als in de mate waarin ze van belang zijn voor verbetering van retrieval performance (Savoy 2004). Dat is waarschijnlijk een van de redenen dat in veel commerciële zoeksoftware deze technieken nog maar beperkt worden toegepast - meestal alleen statistiek en word-stemming en maar zelden verdere linguïstische technieken. De belangrijkste van de technieken waarop hier bedoeld wordt zijn (zie bijvoorbeeld Hiemstra 2001):

Word-stemming

Door bij het indexeren van tekst de daarin voorkomende woorden te reduceren tot hun morfologische woordstam, maakt het niet meer uit op welke woordvorm (enkelvoud, meervoud, verbuiging, vervoeging, enzovoort) wordt gezocht en evenmin welke woordvormen toevallig in de te doorzoeken teksten voorkomen. Als zodanig heeft het een positieve invloed op de recall van zoekacties. De meest toegepaste technieken voor word-stemming zijn regelgebaseerd. Deze methode is sterk taalafhankelijk en niet voor alle talen even eenvoudig mogelijk. Probleem met deze methode is dat er altijd uitzonderingen kunnen voorkomen, waarbij de algemene stemmingregels niet tot correcte resultaten leiden (bijvoorbeeld: communism - community - communication). Stemming kan ook op basis van woordenlijsten en zelfs bestaan er trainbare, deels statistisch afgeleide morfologieën (Bacchin 2005). 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 afgehakt om de juiste woordstam over te houden? En is dat bij het woord "chemical" net zo? Stemming geeft meer verbetering van de recall, naarmate de hoeveelheid digitaal doorzoekbare tekst van documenten kleiner is. In lange documenten zullen namelijk toch al vaak verschillende woordvormen naast elkaar aanwezig zijn.

Fuzzy zoeken

Er zijn verschillende technieken om te kunnen zoeken op woorden die wat betreft spelling of uitspraak sterk lijken op de woorden in de zoekvraag. Hiermee kan gecompenseerd worden voor spelfouten in zowel de documenten als de zoekvragen, alsmede voor spellingsvarianties en morfologische varianten. Hiervoor zijn diverse technieken in gebruik. Hierbij gaat de verbetering van de recall overigens vaak ten koste van de precisie.

Compound-splitting

Het splitsen van samengestelde woorden in hun afzonderlijke bestanddelen heeft veel invloed op retrieval performance voor talen als Duits, Deens en Nederlands (Braschler 2004b). Wel ligt het vaak heel subtiel voor welke woorden het zinnig is ze wel of niet te splitsen. Zo zal het Duitse woord "Frühstück" zeker niet gesplitst moeten worden. Naast echte samenstellingen als "vrachtschip", moeten zogenaamde head-modifier constructies (zeil-makerij) ook gesplitst worden omdat het zinvol is dat documenten ook op die losse woorddelen vindbaar zijn.

Disambiguering

Op basis van woordrelaties in semantische netwerken kunnen verschillende betekenissen van woorden in teksten worden onderscheiden, hetgeen de precisie van zoekacties kan verbeteren. Bij meertalige retrieval zijn hiervoor diverse methoden in gebruik (Kishida 2007).

Vraagexpansie

Met behulp van een semantisch netwerk kunnen zoekvragen worden uitgebreid met woorden die in het netwerk op korte semantische afstand van de zoekwoorden voorkomen, als zijnde inhoudelijk meer of minder gerelateerd of zelfs synoniem met die zoekwoorden. Toch is het niet altijd even duidelijk of de hieruit resulterende verbetering van de recall niet te veel ten koste gaat van de precisie.

Lexical phrases

Software kan trachten diverse soorten zogenaamde "lexical phrases" in teksten te herkennen. Dat kunnen in het Engels veelvuldig voorkomende "noun phrases" zijn (uit twee losse zelfstandig naamwoorden opgebouwde begrippen), maar ook complexere stukken zin met voorzetsels, bijvoeglijke naamwoorden enzovoort, die als geheel een concept of begrip representeren. Door die als mogelijke zoekbegrippen te gebruiken kan de precisie van zoekresultaten worden verhoogd. Doordat dergelijke phrases in zoveel varianten kunnen voorkomen, is het echter lastig om phrases met dezelfde betekenis als zodanig te herkennen om ze allemaal in een zoekvraag te gebruiken. Daardoor is deze techniek meestal nadelig voor de recall van zoekacties.

Een aantal van de bovengenoemde technieken dient in feite als hulpmiddel om te komen tot normalisatie van de gebruikte woorden, begrippen en uitdrukkingen, tot één standaardvorm. Andere technieken daarbij zijn ook syntactische analyse om verschillende syntactische vormen van samengestelde begrippen als gelijkwaardig te herkennen en semantische waarbij met behulp van bijvoorbeeld een semantisch netwerk synonieme begrippen gemapt kunnen worden. Vooral meerwoordsbegrippen leveren echter meestal nog problemen op. In feite zijn dit voor een groot deel dezelfde technieken die ook al in §2.1 werden toegepast bij het karakteriseren van documenten ten behoeve van geautomatiseerde inhoudelijke ontsluiting.

In het al genoemde artikel van Braschler (2004b) wordt de effectiviteit van diverse bestaande technieken vergeleken, zowel voor decompounding als voor stemming. Bij gebruik van de best presterende technieken rapporteert hij verbetering van de precisie met wel 23% en van de recall met 12%. Probleem bij de meeste van deze technieken is dat iedere taalkundige intelligentie ook fouten introduceert. Hoewel ze vaak meer winst dan verlies opleveren, reageren gebruikers toch veelal zeer afwijzend als daardoor - al is het maar af en toe - onzinnige woordvarianten (uitzonderingen bij stemming, te grote fuzziness) of onzinnige "synoniemen" (te grote semantische afstand, synoniemen in verkeerde betekenis) in zoekvragen worden meegenomen.

In zijn algemeenheid lijkt de kwaliteit van zoekresultaten te verminderen met de omvang van de verwerkte teksten. Retrieval op basis van goede samenvattingen geeft vaak betere resultaten dan gebruik van volledige teksten. Als toch volledige teksten van lange documenten full-text doorzoekbaar worden gemaakt, dan kunnen die beter apart per pagina verwerkt worden dan als één geheel. Dat is ook beter dan alleen de

eerste 1000 woorden te nemen. Soortgelijke resultaten rapporteerde Williams (1998) al, op basis van experimenten met het opdelen van documenten door middel van overlappende "windows" van 250 - 1000 woorden.

In dit kader kan ook worden gekeken welke onderdelen van documenten het belangrijkste zijn voor hun terugvindbaarheid, in termen van precisie en van recall (Trotman 2005). Zoekacties kunnen dan tot die onderdelen beperkt blijven. In de eerste plaats zijn dat samenvattingen en koppen. Daarnaast blijkt dat de eerste en laatste zinnen van paragrafen een grotere informatiedichtheid hebben. Dat geldt in wellicht nog sterkere mate voor specifiek wetenschappelijke paragrafen als "Methoden" en "Conclusies". Bij digitaal beschikbare handboeken is vooral de inhoudsopgave een belangrijk onderdeel om doorzoekbaar te maken.

Veel van het onderzoek aan de kwaliteit van retrieval-technieken is gebaseerd op Engelstalige documenten en corpora. Hollink (2004) heeft onderzocht welke van de eerder genoemde technieken voor andere Europese talen tot belangrijke verbetering van retrieval-resultaten leiden. Zij onderscheidt daarbij taalafhankelijke en taal-onafhankelijke technieken. Uiteindelijk blijkt het sterk taalafhankelijk te zijn welke combinaties van technieken uiteindelijk de meeste verbetering veroorzaken. Om ook in meertalige collecties goede zoekresultaten te geven, zonder dat de gebruiker aparte zoekvragen in verschillende talen hoeft te formuleren, wordt in het kader van CLEF en CLIR (Cross-Language Information Retrieval) onderzoek gedaan aan twee- of meertalige zoektechnieken. Ook daarbij worden combinaties van allerlei technieken toegepast. Dat zijn deels dezelfde als voor enkeltalige retrieval, maar die worden onder meer aangevuld met digitale woordenboeken, gebruik van parallelle corpora en terugkoppeling door de gebruiker bij dubbelzinnigheden bij het vertalen. Niettemin zijn de resultaten nog niet zo goed als bij enkeltalige retrieval. (Zie onder meer Braschler 2004a, Chen 2004, Hedlund 2004, Lehtokangas 2004, Kishida 2007).

Een techniek die zowel op het web als in commerciële retrieval-software veel wordt toegepast is het clusteren van verkregen zoekresultaten in groepen documenten die onderlinge gelijkheid vertonen. Dat kan zijn op basis van statistiek of op basis van al toegekende trefwoorden of categorieën. Daarbij worden vaak dezelfde technieken toegepast als die eerder werden beschreven bij het automatisch classificeren van documenten. Zulke clustering disambigueert vaak automatisch verschillende betekenissen van door zoekers gebruikte zoekwoorden. In het Scorpion-project van OCLC werd al een dergelijke techniek toegepast. Op dit moment hoeven we hiervoor maar te kijken naar (meta-)zoekmachines en zoek-frontends op het web, zoals Clusty, Kartoo, Ask, Quintura of Collarity, of naar software van leveranciers als Autonomy.

Het meeste praktisch gerichte retrieval-onderzoek - zeker ook dat in TREC-verband - vindt plaats op grote tekst-corpora en veelal tekstrijke documenten. Toch worden hier beschreven methoden een enkele keer ook wel toegepast op veel tekstarmere corpora uit bibliotheekcatalogi. Een voorbeeld daarvan is onderzoek van Grumann (2000) ten behoeve van openbare bibliotheekcollecties. Dit onderzoek bouwde voort op de eerdere MILOS-projecten voor wetenschappelijke literatuur, uit de jaren 1994-1996 bij de Universitäts- und Landesbibliothek Düsseldorf. Recall en precisie bleken bij die OB-collecties veel minder te verbeteren dan in het MILOS-project gerapporteerd was, waarschijnlijk omdat titels van wetenschappelijke publicaties vaak veel preciezer beschrijvingen van het behandelde onderwerp geven.

3.2 Free-text retrieval of inhoudelijke ontsluiting

Voor free-text retrieval zijn, vooral op basis van omvangrijke gecontroleerde corpora in TREC, veel vergelijkende gegevens over de kwaliteit van allerlei retrieval-systemen beschikbaar. Vergelijkingen tussen zoekresultaten op basis van free-text retrieval en op basis van gecontroleerde handmatige ontsluiting zijn er daarentegen maar weinig - en zeker nauwelijks grootschalig. De paar kleinschalige onderzoeken laten de indruk achter dat beide methoden ongeveer even goed uit de bus komen. Het enige wat grootschaliger onderzoek (Savoy 2005) is uitgevoerd op een bibliografische database met bijna 150.000 franstalige artikelen waarvan titels en samenvattingen digitaal beschikbaar waren. Ook daaruit kwam geen significant verschil tussen de resultaten van gebruik van handmatige toegekende ontsluiting en free-text retrieval. Voor free-text retrieval moesten in al die gevallen wel alle technieken worden ingezet die het meest geëigend zijn voor dat materiaal en die taal (zoals in de voorgaande paragraaf werden besproken). De uitkomsten van wat in allerlei situaties de meest geëigende technieken zijn, blijkt echter nog niet altijd erg eenduidig.

Dat er nog veel te verbeteren valt aan huidige manieren van toegang tot zowel fysieke als digitale bibliotheekcollecties wordt indringend geformuleerd door Davis (2002):

Those who follow the progress of library-based information access and retrieval technologies will, if pressed, be obliged to admit that libraries and the automated system vendors that serve them have done little in the last decade to improve subject access to our print and, now, online collections. Much has of course been written and proposed in the library and information science literature about possible new strategies for access and retrieval, but few new approaches have actually been developed, tested and implemented in recent generations of library OPACs. Some would attribute this variously to: the marginal economics of library automation's niche marketplace; the timid approach vendors have taken to their feature enhancement processes; the enormous technical infrastructure changes libraries and vendors have had to absorb over the last ten years in order to stay even minimally current with new technologies; the aging systems of classification and subject analysis that continue to serve as our cataloging standards; the difficulty of innovating in OPACs when developers are constrained by the heavy hand of Z39.50 and fear the loss of interoperability with consortia and other cooperative systems; and the rise of the Web and the seemingly universal appeal of know-nothing, shot-in-the-dark keyword-Booleanism.

De observatie dat in OPAC's nog weinig geavanceerde technieken voor onderwerps-toegang worden toegepast, kan onmiddellijk worden beaamd. Dat daarin nog weinig van de in dit hoofdstuk besproken retrieval-technieken worden toegepast, is niet zo verwonderlijk, omdat in onze huidige catalogi erg weinig digitale tekst beschikbaar is waarop die taaltechnieken zouden kunnen worden losgelaten. Op die problematiek van de tekstarmoede van catalogusrecords wordt in het volgende hoofdstuk nog nader ingegaan. Inderdaad zal het "schot in het duister", met een combinatie van wat willekeurige zoektermen, die op het web met zijn 40 miljard webpagina's altijd nog wel enig zinnig resultaat oplevert, voor de veel kleinere collecties van bibliotheken meestal bedroevend slechte resultaten geven, ook als de tekstarmoede van catalogusrecords genezen zou zijn. Waarschijnlijk zal een combinatie nodig zijn van moderne retrieval-technieken, automatische karakterisering van documenten en met moderne middelen opgewaardeerde klassieke ontsluitingssystemen, om Davis uiteindelijk tevreden te kunnen stellen.

3.3 De toekomst van free-text retrieval methoden

De verbetering van taaltechnologische technieken die zich manifesteert in geleidelijk steeds wat betere resultaten in TREC zal nog wel enige tijd doorgaan. Of de best haalbare recall- en precisie-cijfers uiteindelijk asymptotisch naderen naar een eigenlijk nog niet acceptabele waarde of dat ze werkelijk aanmerkelijk beter blijven worden, valt echter moeilijk te voorspellen. In dat verband moeten we ons realiseren dat niet iedereen even tevreden is als de mensen uit de information-retrieval-gemeenschap zelf. Van daaruit wordt ons al meer dan een decennium voorgehouden dat we nog maar een kleine stap verwijderd zijn van de uiteindelijke oplossing van de problemen met de niet-eenduidigheid van taal, die ons daarbij vooral opbreekt. Ook de komende tien jaar zou het best nog steeds diezelfde kleine stap kunnen blijven die ons van "de" oplossing scheidt.

4. Ontsluiten van heterogeen materiaal

In de huidige informatiecollecties van veel organisaties kan nogal heterogeen materiaal aanwezig zijn. Die heterogeniteit kan betrekking hebben op verschillende aspecten. Bijvoorbeeld:

- het al dan niet digitaal zijn van het materiaal, zoals bij volledig digitaal aanwezige tijdschriftartikelen, proefschriften of zelfs E-books tegenover boekbeschrijvingen in de catalogus,
- de mate van specialisatie van het materiaal, zoals algemene leerboeken tegenover tijdschriftartikelen over een heel specifiek specialistisch onderwerp (al kunnen ook in een leerboek veel verschillende, tamelijk specialistische onderwerpen aan de orde komen),
- de inhoudelijke aard van het materiaal, zoals bij fictie tegenover non-fictie, waarbij de laatste soort altijd “over” iets gaat - het heeft “aboutness” - maar de eerste lang niet altijd.

In principe kunnen die verschillende soorten materiaal verschillende eisen stellen aan de methoden en technieken die moeten worden ingezet voor hun inhoudelijke ontsluiting. In sommige gevallen gaat het zelfs om de afweging of ontsluiting voor dat materiaal hoe dan ook nodig, nuttig en mogelijk is. Het is hierdoor zeker niet vanzelfsprekend dat voor ontsluiting van die verschillende soorten materiaal een volledig uniforme aanpak gerealiseerd kan worden, met gebruikmaking van hetzelfde vocabulaire. In dit hoofdstuk worden voor een aantal situaties de consequenties en mogelijkheden besproken.

Hoewel een uniforme methode van ontsluiting in elk geval voldoende is om ook geïntegreerd in alle materiaal te kunnen zoeken, is het geen noodzakelijke voorwaarde. Er is ook een andere mogelijkheid. Door het bewerkstelligen van zogenaamde interoperabiliteit tussen systemen waarin het materiaal op verschillende manieren is ontsloten, kan ook geïntegreerd gezocht worden. In het volgende hoofdstuk zullen we daarom ook nog kijken naar de mogelijkheden die er op dat gebied zijn.

Op verschillen tussen tekst- en beeld- of multimediaal materiaal zullen we hier niet ingaan. De ontsluiting van beeldmateriaal - stilstaand of bewegend - is een specialistisch onderwerp met eigen problemen en oplossingen, dat het kader van dit artikel te buiten gaat.

4.1 Digitaal versus niet-digitaal materiaal

Om geautomatiseerde technieken te kunnen toepassen, of dat nu state-of-the-art text-retrieval-technieken betreft, of methoden voor automatische inhoudelijke ontsluiting, zal altijd een voldoende hoeveelheid informatie over of uit de publicaties digitaal beschikbaar moeten zijn. Met een toenemend deel van het materiaal waartoe de hybride bibliotheek toegang verleent, is dat gelukkig wel al het geval. Voor een ander deel echter nog niet. Hoewel voor nieuw uitgegeven boeken in sommige gevallen wel inhoudsopgaven en korte beschrijvingen digitaal beschikbaar worden gesteld, blijft er toch een grote hoeveelheid, vooral ouder materiaal waarvan niet te verwachten is dat op korte termijn dat soort informatie beschikbaar komt.

Een mogelijke oplossing voor dat materiaal is uiteraard scannen en OCR-en. Hoewel klassieke catalogiseerprocessen ook al tamelijk arbeidsintensief zijn, blijken deze technische stappen in het verwerkingsproces van papieren documenten vaak als te arbeidsintensief te worden ervaren. In principe kunnen verbeterde logistiek en workflow en verbeteringen in de techniek zelf tot versnelling van het proces leiden, zoals het Google Book project laat zien. Maar belangrijker is nog dat het besef toeneemt van het belang van deze extra stap in het catalogiseerproces.

Daarbij moet wel een keuze worden gemaakt welke onderdelen van documenten zo gedigitaliseerd moeten worden. In §3.1 werd al besproken wat de meest nuttige en representatieve onderdelen van documenten zijn om in te zoeken. Zo zal men zich bij veel boekmateriaal kunnen beperken tot inhoudsopgaven, omdat de in het boek behandelde onderwerpen daarin meestal in de vorm van hoofdstuk- en paragraaf-titels worden beschreven. Als een samenvatting of een scanbare flap-tekst beschikbaar is, kan ook die voor inhoudelijke toegang nuttige informatie bevatten. Men moet zich wel realiseren dat dergelijke onderdelen niet voor alle boeken beschikbaar zijn. Hoe waardevol ze zijn, kan ook per vakgebied en van boek tot boek verschillen. Daarom zal per individueel boek een verwerkingsstrategie moeten worden bepaald.

In plaats van scannen kan ook geprobeerd worden van andere organisaties al digitaal beschikbare informatie over of uit de boeken te betrekken. Bij Amazon.com en bij een organisatie als Syndetics zijn op dit moment van veel Engelstalige boeken al nuttige onderdelen - samenvattingen, inhoudsopgaven, besprekingen - in digitale vorm op te halen. In de toekomst is zeker te verwachten dat meer van dergelijke diensten beschikbaar komen, hetzij van uitgeverij zelf, hetzij van gespecialiseerde bedrijven. Bij gebruik hiervan zal het zelf scannen tot een geleidelijk afnemend deel van het bij de bibliotheek te verwerken materiaal beperkt kunnen blijven.

Als uiteindelijk een voldoende hoeveelheid van de meest representatieve onderdelen van een publicatie digitaal beschikbaar is, dan kan die gebruikt worden, zowel voor slimme retrieval-methoden, zoals besproken in hoofdstuk 3, als voor automatische indexing zoals besproken in hoofdstuk 2.

In de hybride bibliotheek komen de digitale en de fysieke collectie vaak als nogal ongelijksoortige deelcollecties samen. Bij de vraag of digitaal en niet-digitaal beschikbaar materiaal op dezelfde wijze ontsloten moet en kan worden, staat de vraag centraal of en hoe al geautomatiseerde methoden van ontsluiting worden toegepast. De twee meest gebruikelijke situaties die zich daarbij kunnen voordoen, zijn:

1. Digitaal materiaal niet gecontroleerd ontsloten

Er is besloten dat gecontroleerde ontsluiting van het digitaal beschikbare materiaal niet noodzakelijk is, omdat moderne retrieval-technieken voldoende goede zoekmogelijkheden bieden. In dat geval kan niet meer op dezelfde wijze naar het niet-digitale materiaal gezocht worden, omdat dat relatief veel minder zoekingen biedt. Alleen als op de hierboven beschreven wijze ook nog voldoende digitale informatie over dit materiaal verkregen kan worden, vervalt dit onderscheid. Anders zal op het niet-digitale materiaal een liefst gecontroleerde vorm van handmatige inhoudelijke ontsluiting moeten worden toegepast. Om dan toch in een enkele zoekactie beide deelcollecties met succes te kunnen doorzoeken, zal op zijn minst gezorgd moeten worden dat een voldoende goede mapping beschikbaar is tussen dit gecontroleerde vocabulaire en de natuurlijke taal waarmee de gebruiker in het digitale deel van de collectie zoekt. Op zich correspondeert dat aardig met de in §1.1.2 besproken methoden om thesauri beter te laten aansluiten op het door zoekers gebezigde zoekvocabulaire.

Wanneer het geavanceerde zoekstelsel echter methoden van relevantieordening en relevantierugkoppeling toepast die voortvloeien uit de gebruikte free-text retrieval-technieken (hetgeen zeer waarschijnlijk is), kunnen er problemen ontstaan. Omdat bij het handmatig ontsloten materiaal alleen een beperkt aantal losse concepten zoekbaar is (of dat nu thesaurustermen of categorieën uit een classificatie zijn), zullen die methoden daarop niet (of op zijn best heel anders) toepasbaar zijn. Zoekresultaten uit beide deelcollecties zullen dan niet op zinvolle wijze samengevoegd kunnen worden, zodat de gebruiker toch nog met twee losse lijsten zoekresultaten geconfronteerd zal worden.

2. Digitaal materiaal wel gecontroleerd ontsloten

Er is besloten het digitaal beschikbare materiaal wel van gecontroleerde inhoudelijke ontsluiting te voorzien, maar dit door een geautomatiseerd systeem te laten toekennen. In dit geval kan in elk geval hetzelfde vocabulaire worden gebruikt dat ook voor de handmatige ontsluiting van het niet-digitale materiaal wordt toegepast. De in §1.1.2 en §1.2 besproken methoden om de gebruiker - misschien ongemerkt - bij de juiste gecontroleerde zoektermen te laten uitkomen, kunnen hier dus op de gehele collectie worden toegepast. Toch dreigt hier ook nog een probleem. Op dit moment lijkt geautomatiseerde ontsluiting met een zeer specifiek en daardoor ook zeer omvangrijk vocabulaire nog niet goed mogelijk te zijn. Dit stelt beperkingen aan de mate van specificiteit van de te gebruiken gecontroleerde ontsluiting. Bij het digitaal beschikbare deel van de collectie kan dit opgevangen worden door aanvullend ook full-text retrieval toe te passen. Daarbij wordt het niet-digitale deel van de collectie toch weer van die zoekactie uitgesloten.

In beide situaties kan voor gebruikers van de collectie(s) in elk geval enige mate van uniformiteit gegarandeerd worden. Aan de ontsluitingskant vraagt dat – als men zich voor het digitale materiaal niet helemaal op geavanceerde full-text retrieval-technieken wil verlaten – echter wel een voldoende mate van uniformiteit voor de handmatige en de geautomatiseerde ontsluiting. Men hoeft zich dan niet de extra inspanningen te getroosten die nodig zijn voor het opzetten en onderhouden van een concordantie of andere methoden voor interoperabiliteit tussen verschillende gecontroleerde ontsluitingssystemen.

4.2 Gespecialiseerde versus algemene publicaties

Een scheidslijn waarlangs collecties van veel bibliotheken ook in ongelijksoortige delen uiteen kunnen vallen is die tussen algemene en zeer specialistische publicaties. Hoewel collecties van klassieke papieren bibliotheken naast algemene werken ook specialistische monografieën kunnen bevatten, is dat bij een hybride bibliotheek vaak nog veel sterker het geval, omdat daar ook individuele tijdschriftartikelen tot de collectie kunnen behoren. In de eerste plaats is hier sprake van een schaalverschil, doordat digitale artikelen meestal meteen in zeer grote aantallen aanwezig zullen zijn. In de tweede plaats zal de inhoud van veel artikelen zich beperken tot specialistischer deelonderwerpen dan met monografieën al het geval was. Dat vereist een afweging tussen zeer specifieke ontsluiting van hoog-specialistische (wetenschappelijke) artikelen, en de veel minder specifieke concepten waarmee bijvoorbeeld leerboeken, algemene werken en verzamelwerken worden ontsloten.

Franklin (2003) geeft aan dat het voor wetenschappelijk onderzoekers beslist nodig is op zeer specifiek niveau te kunnen zoeken. Voor bibliotheken met een collectie op vrijwel alle onderwerpsgebieden levert dat het probleem dat niet voor alle disciplines voldoende specifieke ontsluitingssystemen bestaan, die voor gebruikers acceptabel zijn, en dat die zeker nog niet geïntegreerd beschikbaar zijn. Ook als in het kader van interoperabiliteit oplossingen voor integratie en acceptatie worden aangedragen, hebben die vaak nog als beperking dat ze in de opbouwfase te bewerkelijk zijn om ze op voldoende specifiek niveau voor alle onderwerpsdomeinen te kunnen uitwerken.

Een ander mogelijk probleem vormen digitale objecten, zoals bijvoorbeeld databases, die veelal als geheel gecatalogiseerd worden, maar die informatie over een veelheid aan heel specifieke deelonderwerpen kunnen bevatten. Daar zal ergens een praktische grens liggen aan het aantal zeer specifieke termen dat aan een dergelijk object kan worden toegekend. Hoewel dit ook in fysieke collecties van bibliotheken regelmatig voorkwam, onder meer bij verzamelbundels, is er wel sprake van een schaalverschil ten opzichte van sommige digitale objecten.

Anderzijds moet ook de vraag gesteld worden hoe zinvol het is om verschillende materiaalsoorten, als tijdschriftartikelen en boeken, geïntegreerd doorzoekbaar aan te bieden. Voor het zoeken van artikelen staan vaak al gespecialiseerde niet collectiegebonden bibliografische hulpmiddelen ter beschikking. De online beschikbaarheid van de volledige digitale tekst van artikelen kan echter een sterk argument zijn om die wel als integraal onderdeel van de lokale collectie te presenteren en te doorzoeken.

Zoals al aangegeven door Mai (2003), zal er in elk geval altijd behoefte bestaan om ook op globaal niveau in collecties te kunnen zoeken. Hiervoor zijn gecontroleerde ontsluitingssystemen wel beschikbaar. In voorgaande hoofdstukken is al aangegeven hoe die kunnen worden aangepast aan browse-gemak in web-omgevingen en/of aan gebruikersvocabulary. Anderzijds zagen we al hoe onder die omstandigheden digitaal beschikbaar materiaal automatisch gecategoriseerd of van trefwoorden voorzien kan worden. In aanvulling daarop kan voor het specifieke niveau van onderwerpstoegang dan toch eenvoudigweg van full-text retrieval technieken gebruik gemaakt worden. De kwaliteit daarvan is in elk geval goed genoeg om die op dit aanvullende niveau toe te passen. Door een dergelijke combinatie kan zowel de algemene zoeker als de specialistische wetenschappelijke zoeker aan zijn trekken komen.

4.3 Fictie en non-fictie.

Er zijn meer criteria op basis waarvan collecties van bibliotheken in ongelijksoortige deelcollecties uiteenvallen. Zo kan men ook een opdeling maken in twee soorten materiaal die ik gemakshalve met fictie en non-fictie aanduid. Vanuit het oogpunt van inhoudelijke ontsluiting is dit waarschijnlijk een zinvoller onderscheid dan dat tussen wetenschappelijk en niet-wetenschappelijk. Van alle non-fictie, wetenschappelijk of niet, kan immers in principe worden bepaald, waar het inhoudelijk over gaat. Hoogstens kan er verschil zijn in mate van specialisatie of mate van complexiteit van de behandelde en voor ontsluitingsdoeleinden te representeren onderwerpen. In feite moet ervan worden uitgegaan dat alles wat tot dusverre in dit rapport geschreven is, in de praktijk eigenlijk betrekking had op dit soort materiaal.

Alle fictie anderzijds, of het nu "hoge cultuur" of "lage cultuur", doktersroman of Nobelprijsliteratuur betreft, heeft als gezamenlijk kenmerk dat we hierbij meer moeite hebben met de vraag of inhoudelijke ontsluiting hiervan mogelijk en zinvol is. In Nederland wordt bij openbare bibliotheken eigenlijk nooit verder gegaan dan genre-aanduidingen. Elders - en vooral in het Angelsaksische taalgebied - bestaat echter een veel sterkere traditie om aandacht te besteden aan het inhoudelijke aspect van fictie. Miller (2003) bespreekt dit onderwerp bijvoorbeeld vanuit de *Guidelines on Subject Access to Individual Works of Fiction Drama, Etc.* van de ALA (American Library Association), waarin sterk nadruk wordt gelegd op het belang van *aboutness* en *whatness*, ook voor fictie. Vooral voor historisch onderzoek acht hij beschikbaarheid van inhoudelijke ontsluiting van fictie van belang. In de praktijk echter, zullen voor het overgrote deel van het materiaal toch niet veel meer dan geografische, historische en genreaspecten als inhoudelijke ontsluiting in aanmerking komen, in de trant van "historische roman - Frankrijk - 14de eeuw".

In principe is het daarom weinig zinvol en zelfs wat kunstmatig om te proberen dit materiaal helemaal op dezelfde wijze, met hetzelfde vocabulaire, inhoudelijk toegankelijk te maken als non-fictie. Als bepaalde inhoudelijke elementen van fictie toch worden ontsloten, verdient het aanbeveling dat desgewenst wel in alles tegelijk gezocht kan worden. In bijvoorbeeld de catalogus van de nationale bibliotheek van Slovenië levert gebruik van onderwerpsingangen inderdaad zowel non-fictie als fictie op. Ook in de Duitse nationale bibliotheek worden trefwoorden uit de SWD/RSWK tevens voor fictie toegepast. Onderzoek hoe nuttig het zoeken van fictie op inhoudelijke elementen door onderzoekers wordt gevonden, ontbreekt echter nog. Anderszijds moet in zoekacties fictie natuurlijk ook altijd kunnen worden uitgesloten via materiaaltipe of vormrubriek.

5. Interoperabiliteit van systemen

Door de opkomst van het web is een netwerkomgeving ontstaan waarin sterke nadruk ligt op de interoperabiliteit van informatiecollecties. Het biedt de mogelijkheid om zelfs geografisch gescheiden collecties gelijktijdig te doorzoeken. Dat kan zowel gebeuren via metasearch-oplossingen, als via het “oogsten” (harvesting) van door deelnemende organisaties in repositories beschikbaar gesteld materiaal. Het uit verschillende bronnen geogoste materiaal kan dan bijvoorbeeld in één zoekstelsel doorzoekbaar worden gemaakt. Het oogsten zelf beperkt zich vaak tot de metadata die materiaal of objecten beschrijven, maar soms betreft het ook het materiaal zelf om daarvan de volledige tekst doorzoekbaar te kunnen maken. Met de komst van XML als universele coderings- en uitwisselingstaal is speciaal die oogsttechniek populair geworden, met als bekendste voorbeeld het hiervoor ontwikkelde OAI-PMH protocol.

Het belangrijkste probleem dat werkelijke interoperabiliteit in de weg kan staan, is dat collecties vaak op heel verschillende wijze zijn ontsloten. Zowel syntactische als semantische aspecten spelen daarbij een rol. Niet helemaal parallel daaraan, zijn uit praktisch oogpunt ook de volgende twee niveaus van verschil te onderscheiden:

- verschil in metadata schema (welke “velden” gebruikt worden voor ontsluiting),
- verschil in gebruikte termen (het vocabulaire dat in de velden wordt ingevuld).

Bij onderzoek en projecten op het terrein van interoperabiliteit ligt vaak meer nadruk op dat tweede aspect, de mogelijkheden van mapping of concordanties tussen de vocabulaires van verschillende inhoudelijke ontsluitingssystemen. Dat aspect sluit ook nauw aan bij de scope van het huidige artikel. Toch is het zeker nodig ook het eerstgenoemde aspect consequent bij deze problematiek te betrekken. Bij web-toepassingen komen daar bovendien nog zaken bij als technische protocollen en record-syntax (Koch 2006). In het kader van het Delos-project (A Network of Excellence on Digital Libraries) is een gedetailleerd rapport verschenen over vooral de semantische aspecten van de problematiek van interoperabiliteit (Patel 2005).

Voor de in het vorige hoofdstuk gestelde vraag naar de noodzaak van een uniforme aanpak van de inhoudelijke ontsluiting bij heterogene collecties, is interoperabiliteit ook een relevante invalshoek. Als die namelijk makkelijk te bewerkstelligen is, bestaat er weinig noodzaak in elke situatie een uniforme aanpak na te streven voor het ontsluiten van verschillendsoortig materiaal en afzonderlijke deelcollecties binnen eenzelfde organisatie. Anderzijds kunnen de redenen waarom deelcollecties verschillend zijn ontsloten, juist oorzaken zijn dat ook interoperabiliteit middels concordanties op problemen stuit. Voor interoperabiliteit tussen organisaties in verschillende landen kan de problematiek van meertaligheid nog een belangrijke extra complicatie vormen bij het opzetten van concordanties.

In het kader van het semantisch web zijn intussen standaarden ontwikkeld die het voor computersystemen makkelijker moeten maken om, via verwijzingen naar computerleesbare beschrijvingen van verschillende ontsluitingssystemen, automatisch concordanties af te leiden, zowel tussen gebruikte metadata schema's (de velden) als tussen de toegekende termen. Een standaard die daarbij een eerste aanzet tot interoperabiliteit lijkt te kunnen geven, is SKOS (Simple Knowledge Organisation System). Aan het eind van dit hoofdstuk komt die daarom ook kort aan de orde.

5.1 Interoperabiliteit van metadata schema's

Voor de zaken die spelen bij de interoperabiliteit tussen schema's die worden gebruikt om objecten formeel en inhoudelijk te beschrijven, geven Chan (2006) en Zeng (2006) een gedetailleerd overzicht. Het gaat daarbij vooral om de manieren waarop je kunt specificeren welke metadata velden in een bepaald systeem worden gebruikt, welke betekenis die velden hebben (hun semantiek) en volgens welke formele en inhoudelijke afspraken die velden gevuld worden. Dat kan dan als basis dienen om een mapping tussen die velden voor verschillende systemen te realiseren.

Voor de beschrijving van toepassingen van metadata-schema's en van de relaties daartussen, onderscheiden zij de volgende situaties en aanpakken:

- Derivation: de situatie waarin het ene schema is afgeleid uit het andere, meestal als vereenvoudiging, maar soms ook als aanpassing aan een specifieke situatie. Voorbeelden zijn het Metadata Object Description Schema (MODS) dat is afgeleid van de volledige MARC-21 set voor catalogi en MARC-XML dat daar een XML-versie van is.
- Application profiles: specificaties welke elementen uit welke (eventueel verschillende) metadata standaarden in een specifieke situatie worden toegepast (Duval 2002). Zo kan bijvoorbeeld worden aangegeven welke Dublin Core (DC) elementen en welke Learning Object Metadata (LOM) tezamen worden gebruikt om verschillende typen objecten in een bepaald systeem te karakteriseren. Via het XML-mechanisme van "namespaces" kan daarbij naar computerleesbare definities van die verschillende element sets worden verwezen.
- Metadata registry: op het web aanwezige beschrijving van een specifiek metadata schema, waarin alle gegevens daarover gevonden kunnen worden en waarnaar verwezen kan worden ten behoeve van correct en consistent hergebruik in andere toepassingen (Heery 2002, Duval 2002). Er bestaan registries van een enkel schema, bijvoorbeeld het Dublin Core Metadata Registry op de DCMI website, maar ook registries waarin gegevens van meer schema's ten behoeve van een specifieke toepassing worden gecombineerd, zoals beschreven in application profiles. Dit is bijvoorbeeld hoe het bij The European Library wordt toegepast (van Veen 2004).
- Crosswalks: concordanties tussen de elementen van verschillende metadata schema's waarin hun betekenis en syntax worden gespecificeerd (Godby 2004). Zo bestaan er bijvoorbeeld crosswalks tussen MARC en Dublin Core of tussen VRA Core (Visual Resources Association) en Dublin Core. Hierbij kan ook worden aangegeven in hoeverre de betreffende elementen werkelijk equivalent zijn.
- Switching: concordanties tussen grotere aantallen metadata schema's tegelijk, via één gemeenschappelijke tussenstap. Bij meer dan drie verschillende schema's is dit economischer dan alle afzonderlijke 1:1 crosswalks te specificeren. Bij de Getty-trust worden bijvoorbeeld zeven schema's onderling gekoppeld via de CDWA (Categories for the Description of Works of Art). Zelf noemt men ook dit overigens gewoon een crosswalk.

In de huidige netwerk omgeving speelt het begrip "namespaces" in bijna al deze gevallen een belangrijke rol, voor het verwijzen naar de plaatsen (op internet) waar de naamgeving van metadata-schema's, standaarden en definities, liefst computer-interpreteerbaar, te vinden zijn.

5.2 Concordantie van ontsluitingssystemen

Een uitgebreid overzicht van voor concordantie toe te passen technieken en van voorwaarden en beperkingen in verschillende situaties is te vinden in een artikel van Doerr (2001). Daarbij legt hij de nadruk vooral op het interoperabel maken van thesauri. Chan (2002) en meer recent Zeng (2004) geven overzichten, zowel van de verschillende types en methoden van interoperabiliteit, als van de stand van zaken met betrekking tot een aantal projecten op dit terrein.

Speciaal fundamentele verschillen tussen verschillende systemen kunnen het erg lastig maken tot een voldoende exacte mapping van termen of categorieën te komen. Bij deze problematiek gaat het onder meer om:

- verschil in algehele structuur en uitgangspunten,
- verschil in mate van specificiteit,
- verschil in mate van pre- of post-coördinatie,
- verschil in gelaagdheid en uitgebreidheid van de hiërarchische relaties,
- taalverschillen (bij meertaligheid),
- culturele verschillen (als meer landen betrokken zijn).

Het koppelen van verschillende vocabulaires kan op verschillende manieren plaats vinden:

- met klassieke concordanties die termen rechtstreeks aan elkaar koppelen,
- via een centrale "tussentaal",
- door linking tussen zogenaamde "subject authority files", die de volledige beschrijvingen van termen bevatten, en waarin kruisverwijzingen naar andere systemen kunnen zijn opgenomen,
- door statistische methoden, waarbij het samen voorkomen van termen of codes wordt geanalyseerd, bijvoorbeeld op basis van materiaal dat in het verleden al met verschillende systemen is ontsloten.

Zeng verwacht dat hierbij voorlopig nog veel menselijke intellectuele inzet nodig blijft, maar dat geautomatiseerde methodes wel steeds belangrijker worden. De komende tijd zullen die twee aanpakken elkaar in de praktijk dus nog wel blijven aanvullen.

Op dit moment zijn het vooral ontsluitingssystemen voor beperkte onderwerps-domeinen waarvoor al met succes interoperabiliteit bereikt wordt of waarvoor concordanties haalbaar geacht worden. Aders (2005) heeft bijvoorbeeld onderzoek gedaan naar de mogelijkheid om de Thesaurus Gezondheidszorg en de thesaurus van het Nederlands Instituut voor Alcohol en Drugs te integreren. Dit was nodig in verband met een fusie tussen twee instituten die deze thesauri gebruikten. Bij haar onderzoek baseerde zij zich sterk op de aanpak van Hudon (2004) die voor de compatibiliteit van thesauri een viertal niveaus onderscheidt, waarop de gebruikte vocabulaires moeten worden vergeleken, te weten:

- lexicale compatibiliteit: wordt dezelfde soort woordvorm en dergelijke gebruikt?
- conceptuele compatibiliteit: worden in de thesauri overeenkomstige concepten beschreven?
- structurele compatibiliteit: vertonen de thesauri vergelijkbare structuur en relaties tussen de concepten?
- onderwerpscompatibiliteit: in hoeverre dekken de thesauri helemaal hetzelfde onderwerpsgebied?

Bij concordanties in het medische domein dient de UMLS, het Unified Medical Language System, vrijwel altijd als de centrale spil waaromheen ontsluitings-vocabulaires met elkaar in relatie gebracht worden. Bij enkele projecten in Duitsland (CARMEN, Crosswalk STW-SWD zijn concordanties ontwikkeld tussen vakthesauri enerzijds en het algemene Duitse trefwoordensysteem SWD anderzijds. Vizine-Goetz (2004) beschrijft een case study waarin de ERIC-thesaurus wordt gekoppeld aan LCSH. Zij introduceert hierbij het concept van een via het mechanisme van web-services werkende "terminology service". Via het OAI-PMH protocol moet dat toegang verschaffen tot vocabulairesystemen en concordanties. Zij besteedt daarbij eveneens aandacht aan de factoren op basis waarvan de kwaliteit van de mapping tussen twee systemen geëvalueerd kan worden.

Toch zijn er ook wel projecten die een zeer breed onderwerpsterrein - in feite zelfs "alles" - beslaan. In het MACS-project is in Europees verband een grote concordantie opgezet tussen verschillende woordsystemen (Landry 2004, Kunz 2002). Het ging daarbij om de systemen die in gebruik zijn bij de nationale bibliotheken van Frankrijk, Duitsland en Engeland, respectievelijk RAMEAU, SWD/RSWK en LCSH. Ervaring met dergelijke grootschalige concordanties (het kan om 100.000-en termen gaan) leert overigens dat automatische matching op basis van al met verschillende systemen ontsloten publicaties vaak niet erg goed gaat. Dat blijkt dan vooral te komen door onvoldoende kwaliteit van de eerder handmatig toegekende ontsluiting, en dat in het bijzonder voor technische en bètawetenschappen. Overigens zal men altijd moeten accepteren dat enig verlies aan kwaliteit optreedt, doordat verschillende systemen nooit exact 1-op-1 op elkaar zijn af te beelden. Hoeveel verlies optreedt, blijkt vaak sterk domeinafhankelijk te zijn.

Voor interoperabiliteit op brede terreinen ziet Mai (2003) een goede toekomst voor algemene classificaties - eigenlijk acht hij dit nog het enige nut daarvan - omdat die een goede spilfunctie kunnen vervullen en een globale toegang tot een collectie kunnen garanderen voor gebruikers die niet met het lokale systeem bekend zijn. Een aanvulling met meer gespecialiseerde ontsluitingssystemen kan dan veel preciezer en gedetailleerder toegang bieden, vooral voor gebruikers die wel vertrouwd zijn met dergelijke systemen. Juist voor wetenschappelijk onderzoek is die heel precieze en gecontroleerde toegang, ook via het web, namelijk essentieel (Franklin 2003).

Behalve voor bibliografische toepassingen staan ideeën over interoperabiliteit de laatste tijd ook in de belangstelling voor kennisnetwerken van bedrijven en overheidsorganisaties. Het door Gilchrist (2004) beschreven concept van een "Master Authority File" is bedoeld om in dergelijke situaties te dienen als concordantie, zowel om oorspronkelijk verschillende taxonomieën tot één taxonomie samen te voegen, als om een link naar het gebruikersvocabulaire te leggen.

5.3 Praktische toepassingen van interoperabiliteit

Beknopte gegevens van een aantal projecten op het terrein van interoperabiliteit (URL's van vermelde projecten zijn in een bijlage bij dit artikel te vinden).

CARMEN

In dit al in hoofdstuk 2 genoemde project, werd ook aandacht besteed aan interoperabiliteit door een sociaal-wetenschappelijke thesaurus te koppelen aan de Duitse SWD trefwoorden. (Zie ook Kunz 2002). Project is in 2002 afgesloten.

CERES

Project waarin een milieuthesaurus wordt opgezet via integratie van verschillende al bestaande thesauri op de terreinen van biologie en milieu. Tevens ontwikkeling van tools voor navigatie, metadata toekennen en zoeken. Project is in 2003 afgesloten.

CROSSWALK STW-SWD

Project waarbij een standaard economie-thesaurus (STW) werd gekoppeld aan de Duitse SWD trefwoorden. (Zie ook Kunz 2002). Project is in 2003 afgesloten.

DARPA Unfamiliar Metadata Project

Project waarbij vocabulaire van zoekers wordt omgezet in (voor de zoeker in principe onbekend) gecontroleerd vocabulaire van diverse gespecialiseerde databases (Buckland 1999). Het was geïntegreerd met een op probabilistische zoektechnieken gebaseerd retrievalsysteem. Project is in 2001 afgesloten.

HEREIN

European Heritage Information Network. Project voor meertalige toegang tot Europese collecties op het terrein van cultureel erfgoed. Voor geïntegreerde toegang is een nieuwe thesaurus ontwikkeld op basis van al bestaand vocabulaire.

HILT

High Level Thesaurus Project. Koppeling van ontsluitingssystemen van archieven, onderwijsinstellingen, bibliotheken, musea en het Resource Discovery Network vanuit diverse onderwerpsdomeinen in het Verenigd Koninkrijk. Hiertoe moeten LCSH, UNESCO-thesaurus, DDC, UDC en AAT gekoppeld worden via DDC als centrale spil. Daarbij wordt het concept van een "terminology route map" (TeRM) geïntroduceerd (Nicholson 2001, 2002). Die biedt interactie met gebruikers om betekenissen van begrippen te definiëren en te onderscheiden, en met systemen om die begrippen te vertalen in de termen of combinaties van termen die daarvoor in de te doorzoeken systemen worden gebruikt. Fase III van dit project loopt af in 2007.

MACS

Concordantieproject tussen de ontsluitingssystemen van de nationale bibliotheken van het Verenigd Koninkrijk, Frankrijk en Duitsland, die gebruik maken van LCSH, RAMEAU en SWD/RSWK (Landry 2004, Kunz 2002). Project is in 2002 afgesloten.

RENARDUS

Europees project om "subject gateways" van verschillende Europese partners te integreren. Middels DDC werden de in de diverse systemen gebruikte classificaties gekoppeld. Project is in 2003 afgesloten.

5.4 Terminology Services en de SKOS-standaard

Binding (2004) zag nog als voornaamste hinderpaal voor de interoperabiliteit tussen ontsluitingssystemen dat er nog geen algemeen geaccepteerde standaarden bestonden voor toegang tot "knowledge organisation systems", zoals classificaties en thesauri, en evenmin voor de uitwisseling van gegevens daartussen. Toch bestaan al enige tijd zogenaamde "terminology services" met gestandaardiseerde protocollen, waarmee terminologie van thesauri kan worden bevraagd. Zo bestaat het Zthes-protocol, ten behoeve van representatie, navigatie en bevraging van thesauri en autorisatiefiles.

Oorspronkelijk opgezet voor het traditionele Z39.50 protocol, is het nu ook beschikbaar als een specifiek profiel voor het op XML gebaseerde SRU/SRW protocol. Dit laatste wordt onder andere toegepast door OCLC (Vizine-Goetz 2004, 2006) en binnen het HILT-project. Andere technieken zijn onder meer ontwikkeld in:

- het Alexandria Digital Library Project (ADL) voor het bevragen, navigeren en downloaden van thesauri, en daar specifiek toegepast op de eigen thesaurus van geografische verschijnselen,
- het CERES-project voor milieu-informatie.

Beide technieken zijn ook gebaseerd op XML (Patel 2005).

Daarnaast zijn standaarden voor dit doel tot ontwikkeling gekomen vanuit het concept van het semantische web, wat in feite ultieme interoperabiliteit nastreeft. In dat kader gaat het dan wel over ontologieën en over door computers interpreteerbare geformaliseerde beschrijvingen van betekenissen van daarin gedefinieerde concepten en van hun onderlinge relaties. Door het World Wide Web Consortium (W3C) wordt hiervoor een familie van (technische) standaarden gepromoot, die dit gezamenlijk mogelijk moeten maken. Tot die standaarden behoren:

- RDF (Resource Description Framework), een algemene methode om in XML de relatie vast te leggen tussen te karakteriseren objecten, hun eigenschappen en de daarvoor te gebruiken metadata standaarden; daarin wordt via het mechanisme van "namespaces" verwezen naar de beschrijvingen van die metadata standaarden (van Schie 2006);
- OWL (Web Ontology Language), een formele taal voor het computer-interpreteerbaar beschrijven van ontologieën; deze moeten in dit verband meestal worden opgevat in de algemenere betekenis van "kennis-organisatie-systemen" (Isaac 2006a);
- SKOS (Simple Knowledge Organisation System), een conceptueel model waarin je, met RDF als beschrijvingstaal, kunt aangeven uit welk kennis-organisatie-systeem termen afkomstig zijn die als ontsluiting zijn gebruikt, en hoe die zich verhouden tot andere (Isaac 2006b); in feite zou je SKOS misschien beter een standaard kunnen noemen om "schema's voor kennis-organisatie-systemen" op te zetten.

Op dit moment beperkt SKOS zich nog tot de grootste gemene deler van de diverse soorten kennis-organisatie-systemen, de zogenaamde SKOS-core (Miles 2005). Als voorbeeld uit de klassieke ontsluitingswereld, is intussen LCSH in SKOS gecodeerd (Harper 2006). Verder worden ook aanzetten gedaan om SKOS zodanig uit te breiden dat het ook werkelijk via ontologieën tot interoperabiliteit van ontsluitingssystemen kan leiden (Sanchez-Alonso 2006). SKOS is dus een standaard die nog volop in ontwikkeling is. Hoe belangrijk hij uiteindelijk wordt voor de huis-tuin-en-keuken ontsluitingsproblematiek is daarom nog moeilijk in te schatten.

6 Enkele algemene conclusies

Dankzij combinaties van allerlei taalkundige en statistische technieken is de kwaliteit van zoekresultaten bij free-text/full-text retrieval sterk verbeterd. Die verbetering is echter veel sterker op het terrein van de precisie van de resultaten, dan voor de recall. Zeker als tamelijk generieke concepten onderdeel uitmaken van zoekvragen zullen zoekresultaten meestal erg onvolledig blijven. Van documenten die niet zelf al digitaal zijn, zal bovendien altijd nog een voldoende hoeveelheid karakteristieke tekst gedigitaliseerd moeten worden. De beste full-text retrieval biedt dus zeker nog geen panacee voor alle soorten zoekgebruik, zodat er ruimte blijft voor gecontroleerde ontsluiting. Voor het semantisch web, waarin “begrip” van de betekenis van informatie centraal staat, blijft inhoudelijke ontsluiting, met wat daar in het algemeen ontologieën worden genoemd, zelfs een noodzakelijke voorwaarde.

Classificaties zijn in het algemeen geschikt voor gebruikersvriendelijke toegang tot collecties. Zij laten gebruikers via navigatie bij de juiste onderwerpsrubriek uitkomen. Daarmee kan in principe alle materiaal, zelfs over niet-specialistische onderwerpen, tamelijk volledig bij elkaar gevonden worden. Koppeling met gebruikersvocabulaire en ontwikkelingen op het terrein van interoperabiliteit kunnen de toegankelijkheid nog verder verbeteren. Klassieke bibliotheekclassificaties moeten sterk worden aangepast om daarbij tot gemakkelijke bruikbaarheid en voldoende overzichtelijke schermrepresentaties te komen. Voor de wereldwijd veel gebruikte Dewey (DDC) en Library of Congress (LCC) classificaties bestaan intussen dergelijke aanpassingen. Ook belangstelling voor facetclassificaties neemt toe. In het bedrijfsleven worden zulke meerdimensionale systemen wel toegepast onder de noemer van taxonomieën.

Het is moeilijk om op deze wijze ook heel specialistisch materiaal met voldoende precisie toegankelijk te maken. Bij grote collecties is er bovendien een gerelateerd probleem van schaalgrootte. Voor browse-toegang dienen aantallen documenten per categorie niet te groot te zijn. Voor een collectie van 5 miljoen documenten zijn dan tenminste 100.000 à 200.000 categorieën nodig, maar dat heeft weer als bezwaren:

1. dat verantwoord beheer van het ontsluitingssysteem bij dergelijke aantallen een probleem wordt,
2. dat de hiervoor noodzakelijke diepte van tenminste vijf niveaus veelal te groot wordt geacht voor gebruiksvriendelijk browsen,
3. dat methoden voor automatische categorisatie van documenten nog niet goed zijn toegesneden op zulke grote aantallen categorieën.

Thesauri en soortgelijke woordsystemen zijn in principe goed voor de precisie van zoekacties, mits het vocabulaire voldoende specifieke termen bevat. Om discrepantie tussen gebruikerstaal en systeemtaal te overbruggen, moet daarbij gebruik worden gemaakt van methoden om zoektermen van gebruikers automatisch te associëren met corresponderende thesaurustermen. Bij aanwezigheid van goede hiërarchische relaties kan automatische expansie van zoekvragen met specifiekere termen uit de thesaurus een belangrijk middel zijn, om de recall van zoekacties te verbeteren. Bij brede algemene collecties speelt echter het probleem dat waarschijnlijk nog geen voldoende specifiek vocabulaire beschikbaar is op alle vakgebieden die daarin aanwezig zijn. Als dat wel beschikbaar zou zijn, speelt nog de problematiek van de beheersbaarheid van een dergelijk uitgebreid vocabulaire. Bovendien is automatisch toekennen van termen uit zo'n omvangrijk vocabulaire nog niet altijd goed mogelijk.

In zowel ontologieën als topic maps kunnen op geformaliseerde wijze allerlei soorten relaties worden gelegd tussen begrippen. Daardoor moet informatie in principe beter toegankelijk gemaakt kunnen worden, dan met de klassieke ontsluitingsmethoden. In de praktijk bestaan echter nog geen toepassingen voor grote algemene informatiecollecties. Bovendien lijkt toepassing voorsnog erg arbeidsintensief te zijn.

Computeranalyse van digitaal beschikbare teksten - meestal in een combinatie van statistische, linguïstische en regelgebaseerde technieken - wordt steeds meer toegepast voor automatische karakterisering van hun inhoud. Na training met voorbeeldmateriaal op basis van een bestaande classificatie of thesaurus, kunnen deze vingerafdrukken worden gebruikt om nieuwe documenten met redelijke betrouwbaarheid te categoriseren of van trefwoorden te voorzien. Door instellen van betrouwbaarheidsdrempels kan handmatige ontsluiting dan beperkt blijven tot een klein percentage voor het computersysteem moeilijke gevallen. Deze technieken kunnen ook worden toegepast als ondersteuning bij puur handmatige ontsluiting, opdat menselijke indexeerders consistentere en vollediger termen of klassen toekennen. Voor succesvolle automatische toepassing dient het aantal klassen of thesaurustermen niet te groot te zijn en dienen ze onderling voldoende "orthogonaal" te zijn. Dit stelt een beperking aan de mogelijke mate van specificiteit, in het bijzonder bij algemene collecties. Het trainen van een systeem kan - zeker als nog geen handmatig ontsloten voorbeeldmateriaal beschikbaar is - zeer tijdrovend zijn.

Als niet al te strenge eisen gesteld worden aan zoekkwaliteit, is het via allerlei technieken van vraagvertaling en concordantie mogelijk om gelijktijdig te zoeken in verschillend ontsloten systemen of (deel)collecties. In het kader van interoperabiliteit van systemen wordt daar veel onderzoek naar verricht. Collecties waarin alleen op basis van full-text retrieval wordt gezocht, worden daar niet bij betrokken. Andersom zal het, bij toepassing van full-text retrieval als overkoepelend zoekstelsel, nodig zijn om van elke publicatie uit het niet-digitale deel van de collectie een voldoende hoeveelheid tekst in digitale vorm te verkrijgen, hetzij door OCR, hetzij via een hierin gespecialiseerde leverancier.

Omdat de in de digitale omgeving toegepaste vormen van classificaties een beperking opleggen aan de mate van specificiteit van de ontsluiting, zijn ze voor zeer grote algemene documentcollecties en voor specialistische publicaties eigenlijk alleen geschikt om in zoeksystemen voorselectie (of naselectie) op bepaalde onderwerpsdomeinen mogelijk te maken. Voor precieze zoekacties zal toch van thesauri of van free-text/full-text zoeken gebruik gemaakt moeten worden. De beste oplossing om zowel de algemene als de specialistische (wetenschappelijke) zoeker voldoende aan zijn trekken te laten komen, lijkt op dit moment een combinatie van

- enerzijds een classificatie of een betrekkelijk globale thesaurus om op globale onderwerps-elementen te zoeken of te beperken, waarbij geautomatiseerde categorisatie of trefwoordtoekenning wordt toegepast, en
- anderzijds geavanceerde free-text retrieval-mogelijkheden, met inzet van zoveel mogelijk technologische hoogstandjes, voor het precies (en zelfs redelijk volledig) kunnen zoeken op specialistische concepten en onderwerps-elementen.

Daarnaast vereist interoperabiliteit van ontsluitingsystemen in een netwerkgeving verdere ontwikkeling van technieken en het praktisch realiseren van concordanties.

Literatuur

(URL's gecontroleerd op 6-1-2007)

- Nelleke Aders (2005) - De mogelijkheden tot het integreren van thesauri - *Informatie Professional* 9, nr 9, 31-36
- Jean Aitchison, Stella Dextre Clarke (2004) - The thesaurus: a historical viewpoint, with a look to the future - *Cataloging & Classification Quarterly* 37, nr 3/4, 5-21
- James D. Anderson, José Pérez-Carballo (2001) - The nature of indexing: how humans and machines analyze messages and texts for retrieval;
Part I: Research, and the nature of human indexing - *Information Processing and Management* 37, 231-254
Part II: Machine indexing, and the allocation of human versus machine effort - *Information Processing and Management* 37, 255-277
- James D. Anderson and Melissa A. Hofmann (2006) - A Fully Faceted Syntax for Library of Congress Subject Headings - *Cataloging & Classification Quarterly* 43, nr 1, 7-38
- Anders Ardö, Traugott Koch (1999) - Automatic classification applied to the full-text Internet documents in a robot-generated subject index - *Proceedings of the Online Information Conference 1999*, London, <http://www.it.lth.se/anders/online99/>
- Michela Bacchin, Nicola Ferro, Massimo Melucci (2005) - A probabilistic model for stemmer generation - *Information Processing and Management* 41, 121-137
- Donald Beagle (2003) - Visualizing Keyword Distribution Across Multidisciplinary C-Space – *D-Lib Magazine* 9, nr 6, <http://www.dlib.org/dlib/june03/beagle/06beagle.html>
- Nuria Bel, Cornelis H.A. Koster, Marta Villegas (2003)- Cross-Lingual Text Categorization - *Proceedings ECDL 2003, Trondheim, 2003*, pp 126-139, <http://www.cs.ru.nl/peking/ecdl03.pdf>
- P. Biebricher, N. Fuhr, G. Knorz, G. Lustig, M. Schwantner (1988) - The Automatic Indexing System AIR/PHYS; from Research to Application - 11th International Conference on Research and Development in Information Retrieval, 333-342
- C. Binding, D. Tudhope (2004) - KOS at your Service: Programmatic Access to Knowledge Organisation Systems Example - *Journal of Digital Information* 4, nr 4, <http://jodi.tamu.edu/Articles/v04/i04/Binding/>
- Martin Braschler (2004a) - Combination Approaches for Multilingual Text Retrieval - *Information Retrieval* 7, nr 1, 183-204
- Martin Braschler, Bärbel Ripplinger (2004b) - How Effective is Stemming and Decomposing for German Text Retrieval? - *Information Retrieval* 7, nr 3, 291-316
- Michael Buckland (1999) - Mapping entry vocabulary to unfamiliar metadata vocabularies – *D-Lib Magazine* 5, nr 1, <http://www.dlib.org/dlib/january99/buckland/01buckland.html>
- L.M. Chan, E. Childress, R. Dean, E.T. O'Neill, D. Vizine-Goetz (2001) - A faceted approach to subject data in the Dublin Core metadata record - *Journal of Internet Cataloging* 4, nr 1/2, 35-47
- L.M. Chan, M.L. Zeng (2002) - Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis - 68th IFLA Council and General Conference 2002, *IFLA Journal* 28, nr 5/6, 323-27, <http://www.ifla.org/IV/ifla68/papers/008-122e.pdf>

- L.M. Chan, M.L. Zeng (2006) - Metadata interoperability and standardization: a study of methodology; Part I. Achieving interoperability at the schema level - D-Lib Magazine 12, nr 6, <http://dlib.org/dlib/june06/chan/06chan.html>
- A. Chen, F.C. Gey (2004) - Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding - Information Retrieval 7, nr 1, 149-182
- C.F. Cheung, W.B. Lee, Y. Wang (2005) - A multi-facet taxonomy system with applications in unstructured knowledge management - Journal Of Knowledge Management 9, nr 6, 76-91,
- Stephen Paul Davis (2002) - HILCC: A Hierarchical Interface to Library of Congress Classification - Journal of Internet Cataloging 5, nr 4, 19-49
- Rebecca J Dean (2004) - FAST: development of simplified headings for metadata - Cataloging & Classification Quarterly 39, nr 1/2, 331-352
- Martin Doerr (2001) - Semantic problems of thesaurus mapping - Journal of Digital Information 1, nr 8, <http://jodi.tamu.edu/Articles/v01/i08/Doerr/>
- Martin Doerr, Jane Hunter, Carl Lagoze (2003) - Towards a Core Ontology for Information Integration - Journal of Digital Information 4, nr 1, <http://jodi.tamu.edu/Articles/v04/i01/Doerr/>
- S.T. Dumais, D.D. Lewis, F. Sebastiani (2002) - Report on the workshop on operational text classification systems (OTC-02) - ACM SIGIR Forum 35, nr 2, 8-11, <http://www.acm.org/sigs/sigir/forum/F2002/sebastiani.pdf>
- Erik Duval, Wayne Hodgins, Stuart Sutton, Stuart L. Weibel (2002) - Metadata Principles and Practicalities - D-Lib Magazine 8, nr 4, <http://www.dlib.org/dlib/april02/weibel/04weibel.html>
- Rosemary A. Franklin (2003) - Re-inventing subject access for the semantic Web - Online Information Review 27, nr 2, 94-101
- Lars Marius Garshol (2004) - Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all - Journal of Information Science 30, nr 4, 378-391
- Joop van Gent, Onno Makor (2002) - Automatische verrijking in de praktijk - Informatie Professional 6, nr 7/8, 28-31
- Alan Gilchrist (2004) - Thesauri, taxonomieën en ontologieën: overeenkomsten en verschillen - Informatie Professional 6, nr 10, 24-27
- Carol Jean Godby, Ray Reighart (1998) - Using Machine-Readable Text as a Source of Novel Vocabulary to Update the Dewey Decimal Classification - 1998 ASIS Classification Workshop
- Jean Godby, Jay Stuler (2001a) - The Library of Congress Classification as a knowledge base for automatic subject categorization - IFLA Preconference, "Subject Retrieval in a Networked Environment", Dublin, Ohio, August 2001, http://staff.oclc.org/~godby/auto_class/godby-ifla.html
- C. Jean Godby, Ray R. Reighart (2001b) - The Wordsmith indexing system - Journal of Library Administration 34, nr 3/4, 375-384
- Carol Jean Godby, Ray Reighart (2001c) - Terminology Identification in a Collection of Web Resources - Journal of Internet Cataloging 4, nr 1/2, 49-65
- Jean Godby, Devon Smith (2002) - Strategies for Subject Navigation Using RDF Topic Maps - Knowledge Technologies 2002 Conference. Seattle, Washington, March 2002, http://staff.oclc.org/~godby/auto_class/godby_kt2002.ppt

- Carol Jean Godby, Jeffrey A. Young, Eric Childress (2004) - A Repository of Metadata Crosswalks - D-Lib Magazine 10, nr 12, <http://www.dlib.org/dlib/december04/godby/12godby.html>
- Scott Golder, Bernardo A. Huberman (2006) - Usage Patterns of Collaborative Tagging Systems - Journal of Information Science 32, nr 2, 198-208, <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>
- Koraljka Golub (2006) - Automated subject classification of textual web documents - Journal of Documentation 62, nr 3, 350-371
- Jane Greenberg (2001a) - Automatic query expansion via lexical-semantic relationships - Journal of the American Society for Information Science and Technology 52, nr 5, 402-415
- Jane Greenberg (2001b) - Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology - Journal of the American Society for Information Science and Technology 52, nr 6, 487-498
- Jane Greenberg (2004) - User comprehension and searching with information retrieval thesauri - Cataloging & Classification Quarterly 37, nr 3/4, 103-120
- Martin Grumann (2000) - Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? - Bibliothek 24, nr 3, 297-318
- Marieke Guy, Emma Tonkin (2006) - Folksonomies: tidying up tags? - D-Lib Magazine 12, nr 1, <http://dlib.org/dlib/january06/guy/01guy.html>
- Tony Hammond, Timo Hannay, Ben Lund, Joanna Scott (2005) - Social Bookmarking Tools (I): A General Overview - D-Lib Magazine 11, nr 4, <http://dlib.org/dlib/april05/hammond/04hammond.html>
- T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, K. Järvelin (2004) - Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000–2002 - Information Retrieval 7, nr 1, 99-119
- Corey A. Harper (2006) - Encoding Library of Congress Subject Headings in SKOS: Authority Control for the Semantic Web - International Conference on Dublin Core and Metadata Applications, 3 - 6 October 2006 Conference Proceedings, <https://scholarsbank.uoregon.edu/dspace/bitstream/1794/3268/1/dc2006.pdf>
- Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, Ka-Ping Yee (2002) - Finding the Flow in Web Site Search - Communications of the ACM 45, nr 9, 42-49, <http://www.ischool.berkeley.edu/~hearst/papers/paper3.pdf>
- Rachel Heery, Harry Wagner (2002) - A Metadata Registry for the Semantic Web - D-Lib Magazine 8, nr 5, <http://www.dlib.org/dlib/may02/wagner/05wagner.html>
- Magda Heiner-Freiling (2003) - Die DDC in der Deutschen Nationalbibliografie - Dialog mit Bibliotheken 15, nr 3, 8-13, http://www.ddc-deutsch.de/publikationen/pdf/heiner-freiling_3_2003.pdf
- Djoerd Hiemstra (2001) - Using language models for information retrieval - Proefschrift Universiteit Twente, 19 januari 2001, <http://purl.org/utwente/1368>
- Linda Hill, Olha Buchel, Greg Janée, Marcia Lei Zeng (2002) - Integration of Knowledge Organization Systems into Digital Library Architectures - Proceedings of the 13th ASIST SIG/CR Workshop on "Reconceptualizing Classification Research", 62-68, <http://www.alexandria.ucsb.edu/~gjanee/archive/2002/kos-dl-paper.pdf>
- Vera Hollink, Jaap Kamps, Christof Monz, Maarten de Rijke (2004) - Monolingual Document Retrieval for European Languages - Information Retrieval 7, nr 1, 33–52

- Michèle Hudon (2001) - Structuration du savoir et organisation des collections dans les répertoires du Web - Bulletin des bibliothèques de France 46, nr 1, 57-62
- Michèle Hudon (2004) - Conceptual and Lexical Compatibility in Thesauri Used to Describe and Access Moving Image Collections - Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science, 3-5 June 2004, Winnipeg, Manitoba, http://www.cais-acsi.ca/proceedings/2004/hudon_2004.pdf
- Antoine Isaac (2006a) - OWL: Web Ontology Language - Informatie Professional 10, nr 12, 30-33
- Antoine Isaac (2006b) - SKOS: Simple Knowledge Organisation System - Informatie Professional 10, nr 11, 40-43
- Péter Jacsó (2004) – Natural language searching – Online Information Review 28, nr 1, 75-79
- Diane Kelly, Xin Fu (2007) - Eliciting better information need descriptions from users of information search systems - Information Processing & Management 43, 30–46
- Kazuaki Kishida (2007) - Term disambiguation techniques based on target document collection for cross-language information retrieval: An empirical comparison of performance between techniques - Information Processing & Management 43, nr 1, 103-120
- Traugott Koch (2006) - Electronic thesis and dissertation services: semantic interoperability, subject access, multilinguality - E-Thesis Workshop, Amsterdam 2006-01-19/20, <http://www.ukoln.ac.uk/ukoln/staff/t.koch/publ/e-thesis-200601.html>
- Cornelis H.A. Koster, Marc Seutter (2003a) - Taming Wild Phrases - Proceedings 25th European Conference on IR Research, pp 161-176, <http://www.cs.ru.nl/peking/ecir03.pdf>
- Cornelis H.A. Koster, Marc Seutter, Jean G. Beney (2003b) - Multi-Classification of Patent Applications with Winnow - Proceedings PSI 2003, pp 545-554, <http://www.cs.ru.nl/peking/psi2003.pdf>
- Martin Kunz (2002) - Sachliche Suche in verteilten Ressourcen: ein kurzer Überblick über neuere Entwicklungen - 68th IFLA Council and General Conference, Glasgow, August 2002, <http://www.ifla.org/IV/ifla68/papers/007-122g.pdf>
- Patrice Landry (2004) - Multilingual subject access: the linking approach of MACS - Cataloging & Classification Quarterly 37, nr 3/4, 177-191
- Raija Lehtokangas, Eija Airio, Kalervo Järvelin (2004) - Transitive dictionary translation challenges direct dictionary translation in CLIR - Information Processing and Management 40, nr 6, 973-988
- Jessica Lindholm, Tomas Schönthal, Kjell Jansson (2003) - Experiences of Harvesting Web Resources in Engineering using Automatic Classification - Ariadne nr. 37, <http://www.ariadne.ac.uk/issue37/lindholm/intro.html>
- Ben Lund, Tony Hammond, Timo Hannay, Martin Flack (2005) - Social Bookmarking Tools (II): A Case Study: Connotea - D-Lib Magazine 11, nr 4, <http://dlib.org/dlib/april05/lund/04lund.html>
- George Macgregor, Emma McCulloch (2006) - Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool - Library Review 55, nr 5, 291-300
- Jens-Erik Mai (2003) - The future of general classification - Cataloging & Classification Quarterly 37, nr 1/2, 3-12
- O. Medelyan, I. Witten (2006) - Thesaurus Based Automatic Keyphrase Indexing - Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 296-297, <http://www.cs.waikato.ac.nz/~ihw/papers/06-OM-IHW-Autokeyphraseindex.pdf>

- A Miles, B Matthews, M D Wilson, D Brickley (2005) - SKOS Core: Simple Knowledge Organisation for the Web - Proc. International Conference on Dublin Core and Metadata Applications (DC-2005), Madrid, Spain, 12-15 Sep 2005, <http://epubs.cclrc.ac.uk/bitstream/675/dc2005skospaperssubmission1.pdf>
- Alistair Miles, Brian Matthews, Dave Beckett, Dan Brickley, Michael Wilson, Nikki Rogers (2005) - SKOS: A language to describe simple knowledge structures for the web - XTech 2005: XML, The Web and Beyond (XTECH 2005), Amsterdam, 2005, <http://epubs.cclrc.ac.uk/bitstream/685/SKOS-XTech2005.pdf>
- Christopher Miller (2003) - All new subject access to fiction: how a cultural zeitgeist with grayhair informed ALA's Guidelines – Cataloging & Classification Quarterly 36, nr 2, 89-98
- Jessica L. Milstead (1999) - Metadata: the content issue - Proceedings of the National Online Meeting, New York, May 1999, 313-319
- Aurélie Névéol, Alexandrina Rogozan, Stéfan Darmoni (2006) - Automatic indexing of online health resources for a French quality controlled gateway - Information Processing and Management 42, 695-709
- Dennis Nicholson, Susannah Neill (2001) - Interoperability in subject terminologies: The HILT project - The New review of Information Networking 7, 147-157
- Dennis Nicholson, Gordon Dunsire, Susannah Neill (2002) - Moving towards interoperability in subject terminologies - Journal of Internet Cataloging 5, nr 4, 97-111
- Louisa Nigg (2004) - Von automatischer Indexierung zur Klassifizierung - Seminar angewandtes Informaton Retrieval - Sommersemester 2004, http://www.unibas.ch/LIlab/studies/IR-SS2004/SeminarArbeit_Nigg.pdf
- Edward T. O'Neill, Lois Mai Chan (2003) - FAST (Faceted Application of Subject Terminology): A Simplified LCSH-based Vocabulary - 69th IFLA General Conference and Council, Berlin, August 2003, http://www.ifla.org/IV/ifla69/papers/010e-O'Neill_Mai-Chan.pdf
- Edward T. O'Neill, Eric Childress, Rebecca Dean, Kerre Kammerer, Diane Vizine-Goetz, Lois Mai Chan, Lynn El-Hoshy (2001) - FAST: Faceted Application of Subject Terminology - IFLA Satellite Meeting on "Subject Retrieval in a Networked Environment", Dublin, Ohio, August 2001, <http://www.oclc.org/research/projects/fast/dc-fast.doc>
- Manjula Patel, Traugott Koch, Martin Doerr, Chrisa Tsinaraki (2005) - Semantic Interoperability in Digital Library Systems - DELOS Network of Excellence, WP5 Deliverable D5.3.1, <http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/SI-in-DLs.pdf>
- A. Stephen Pollitt, Amanda J. Tinker, Patrick A.J. Braekevelt (1998) - Improving access to online information using dynamic faceted classification - Proceedings of the 22nd International Online Information Meeting, London, December 1998, 17-21
- Corine Quarles van Ufford (2004) - Monnikenwerk; milieu-informatie ontsluiten volgens het verdrag van Aarhus - Informatie Professional 8, nr 11, 30-33
- Jonathan Rothman (2002) - Bridging the Gap Between Materials-Focus and Audience-Focus: Providing Subject Categorization for Users of Electronic Resources - Journal of Internet Cataloging 5, nr 4, 67-80
- H. Saeed, A.S. Chaudry (2001) - Potential of bibliographic tools to organize knowledge on the internet: the use of Dewey Decimal Classification scheme for organizing web-based information resources - Knowledge Organization 28, nr 1, 17-26
- Salvador Sanchez-Alonso, Elena Garcia-Barriocanal (2006) - Making use of upper ontologies to foster interoperability between SKOS concept schemes - Online Information Review 30, nr 3, 263-277
- Jacques Savoy (2004) - Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval - Information Retrieval 7, nr 1, 121-148

- Jacques Savoy (2005) - Bibliographic database access using free-text and controlled vocabulary: an evaluation - *Information Processing & Management* 41, nr 4, 873-890

- Maarten van Schie (2006) - RDF: Resource Description Framework - *Informatie Professional* 10, nr 10, 40-43

- Fabrizio Sebastiani, Evgeniy Gabrilovich (2006) - Bibliography on Automated Text Categorization - <http://iinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html>

- Keith E. Shafer (2001) - Automatic subject assignment via the Scorpion system - *Journal of library administration* 34, nr 1/2, 187-189

- A. Shiri, C. Revie (2006) - Query expansion behavior within a thesaurus-enhanced search environment: a user-centered evaluation - *Journal of the American Society for Information Science* 57, nr 4, 462-478

- Maria L. Silveira, Berthier Ribeiro-Neto (2004) - Concept-based ranking: a case study in the juridical domain - *Information Processing and Management* 40, nr 5, 791-805

- Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer, Stephen Katz (2004) - Reengineering Thesauri for New Applications: the AGROVOC Example - *Journal of Digital Information* 4, nr 4, <http://jodi.tamu.edu/Articles/v04/i04/Soergel/>

- Bruce Sterling (2005) - Order Out of Chaos; What's the best way to tag, bag, and sort data? Give it to the unorganized masses - *Wired* 13, nr 4, <http://www.wired.com/wired/archive/13.04/view.html?pg=4>

- Amanda J. Tinker, A. Steven Pollitt, Ann O'Brien, Patrick A. Braekevelt (1999) - The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation - *Knowledge Organisation* 26, nr 2, 80-96

- E. Toth (2002) - Innovative solutions in automatic classification: a brief summary - *Libri* 52, 48-53

- Andrew Trotman (2005) - Choosing document structure weights - *Information Processing & Management* 41, 243-264

- Douglas Tudhope, Ceri Binding, Dorothee Blocks, Daniel Cunliffe (2002) - Compound Descriptors in Context: A Matching Function for Classifications and Thesauri - *Proceeding of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*

- Douglas Tudhope, Ceri Binding, Dorothee Blocks, Daniel Cunliffe (2006a) - Query expansion via conceptual distance in thesaurus indexed collections - *Journal of Documentation* 62, nr 4, 509-533

- Douglas Tudhope, Traugott Koch, Rachel Heery (2006b) - Terminology Services and Technology; JISC state of the art review - UKOLN, September 2006, <http://www.ukoln.ac.uk/terminology/TSreview-jisc-final-Sept.doc>, http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

- T. van Veen, B. Oldroyd (2004) - Search and Retrieval in The European Library, a new approach - *D-Lib Magazine* 10, nr 2, <http://www.dlib.org/dlib/february04/vanveen/02vanveen.html>

- Diane Vizine-Goetz (2002) - Classification Schemes for Internet Resources Revisited - *Journal of Internet Cataloging* 5, nr 4, 5-18

- D. Vizine-Goetz, C. Hickey, A. Houghton, R. Thompson (2004) - Vocabulary Mapping for Terminology Services - *Journal of Digital Information* 4, nr 4, <http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/>

- Diane Vizine-Goetz, Andrew Houghton, Eric Childress (2006) - *ASIS&T Bulletin*, june/july 2006, http://www.asist.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html

- Piek Vossen, Elton Glaser, Hetty van Zutphen, Rachel Steenwijk (2004) - Validation of MEANING - WP8.1 Deliverable 8.1, MEANING, IST-2001-34460, Irion Technologies BV, Delft, The Netherlands, http://www.vossen.info/docs/2004/EVALUATION_REUTERS5.pdf

- H.-J. Wätjen, B. Diekman, G. Möller, K.-U. Carstensen (1998) - Bericht zur DFG-Projekt GERHARD: German Harvest Automated Retrieval and Directory (16-6-1998)

- Michael Williams (1998) - An Evaluation of Passage-Level Indexing Strategies for a Technical Report Archive - LIBRES: Library and Information Science Research 8, nr 1, <http://libres.curtin.edu.au/libre8n1/williams.htm>

- Ka-Ping Yee, Kirsten Swearingen, Kevin Li, Marti Hearst (2003) - Faceted Metadata for Image Search and Browsing - Proceedings of the SIGCHI conference on Human factors in computing systems, Ft. Lauderdale 2003, 401-408, <http://flamenco.berkeley.edu/papers/flamenco-chi03.pdf>

- M.L. Zeng, L.M. Chan (2004) - Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems - Journal of the American Society for Information Science and Technology 55, nr 5, 377-395

- M.L. Zeng, L.M. Chan (2006) - Metadata interoperability and standardization: a study of methodology; Part 2. Achieving interoperability at the record and repository levels - D-Lib Magazine 12, nr 6, <http://dlib.org/dlib/june06/zeng/06zeng.html>

*Dit artikel is voor een deel gebaseerd op een voor de Koninklijke Bibliotheek geschreven onderzoeksrapport en het daarvoor verrichte literatuuronderzoek:
Inhoudelijk toegankelijk maken van hybride bibliotheekcollecties; een verkennend onderzoek naar huidige opvattingen, recente ontwikkelingen en toekomstverwachtingen / Eric G. Sieverts, Amsterdam, 31 oktober 2004 (50 blz.), <http://www.kb.nl/bst/too/rapport-KB.pdf>*

Gegevens van relevante projecten, producten en websites

(URL's gecontroleerd op 7-1-2007)

ADIURI

<http://www.adiuri.com/facet.htm>

bedrijf dat facetclassificatie in zijn zoeksystemen verwerkt

ADL (Alexandria Digital Library Project)

<http://www.alexandria.ucsb.edu/>

project met terminology service voor geografische thesaurus

ARION

http://www.dl-forum.de/englisch/projekte/projekte_eng_526_ENG_HTML.htm

<http://www.cultivate-int.org/issue4/arion/>

Architecture for Accessing Scientific Collections

Bibliography on Automatic Text Categorization

<http://iinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html>

zoek- en browse-bare collectie van 577 artikelen

BINDEX

<http://www.iai.uni-sb.de/iaien/en/bindex.htm>

Bilingual Automatic Parallel Indexing and Classification

CARMEN (Content Analysis, Retrieval and Metadata: Effective Networking)

<http://www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml>

project afgesloten 2002

CATCH

<http://www.nwo.nl/catch>

NWO-programma Continuous Access To Cultural Heritage

CERES (California Environmental Resources Evaluation System)

<http://ceres.ca.gov/thesaurus/>

CERES/NBII Thesaurus Partnership Project

CIDOC Conceptual Reference Model

<http://cidoc.ics.forth.gr/index.html>

structuur van concepten en relaties in cultureel erfgoed documentatie

CRIS

<http://www.eurocris.org:8080/Lenya/euroCRIS/live/index.htm>

Europees Research Information System

CROSSWALK STW-SWD

<http://www.uni-kiel.de/IfW/zbw/projekte/konkordanz-e.html>

concordantie economiethesaurus, met algemene standaardthesaurus

CYC

<http://www.cyc.com/cyc/technology/whatisyc>

De Cyc upper ontology

DARPA Unfamiliar Metadata Project

<http://metadata.sims.berkeley.edu/GrantSupported/unfamiliar.html>

mapping entry vocabulary to metadata vocabularies

DDC-DEUTSCH

<http://www.ddc-deutsch.de>

introdactie van DDC in de Duitse Nationale bibliografie

DELOS

<http://www.delos.info/>

<http://www.lub.lu.se/tk/publ/SI-in-DLs.htm>

Network of Excellence on Digital Libraries

DESIRE

<http://www.lub.lu.se/desire/demonstration.html>

<http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003489>

Automatic Classification and Content Navigation Support

EEL (Engineering Electronic Library)

<http://www.lub.lu.se/eel/home.html>

<http://www.lub.lu.se/eel/aboutsubj.html>

How EELS classifies subjects; EELS classification system

ETB European Treasury Browser

<http://etb.eun.org/eun.org2/eun/en/etb/content.cfm?lang=en&ov=7208>

European Schoolnet multilingual thesaurus for educational material

FACET (University of Glamorgan)

<http://rapid.isd.glam.ac.uk/FACET/live/overview.asp>

faceted thesauri for retrieval from multimedia collections

FLAMENCO

<http://flamenco.berkeley.edu/>

demonstratie van zoekinterfaces die gebruik maken van

facetclassificatie

HEREIN

<http://www.european-heritage.net/sdx/herein/thesaurus/introduction.xsp>

European Heritage Information Network

HILT (JISC)

<http://hilt.cdrl.strath.ac.uk/>

high-level thesaurus for cross-searching and browsing

<http://hilt.cdrl.strath.ac.uk/hilt2web/Sources/thesauri.html>

A-Z lijst van online beschikbare thesauri

INTERSPACE

<http://www.canis.uiuc.edu/interspace/>

prototype environment for semantic indexing

iVia Research and Development Projects

<http://ivia.ucr.edu/projects/>

diverse projecten voor automatische verrijking en conversie

KBS/Ontology Projects

<http://www.cs.utexas.edu/users/mfkb/related.html>

overzicht van ontologie-projecten en -groepen wereldwijd

MACS (Multilingual access to subjects)

<https://macs.vub.ac.be/pub/>

project afgesloten 2001

MEANING

http://cordis.europa.eu/data/PROJ_FP5/ACTIONeqDndSESSIONeq112422005919ndDOCe640ndTBLeqEN_PROJ.htm

Multilingual Web-scale Language Technologies

MILOS (Universitäts- und Landesbibliothek Düsseldorf)

http://www.ub.uni-duesseldorf.de/home/ueber_uns/projekte/abgeschlossene_projekte/milos
project afgesloten 2001

MXD

http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_v1.1.0.pdf

uitwisselingsformat Deense Nationale Research Database

NACTEM (National Centre for Text Mining)

<http://www.nactem.ac.uk/resources.php?view=5>

overzicht van biomedische ontologieën

NKOS

<http://nkos.slis.kent.edu/>

Networked Knowledge Organization Systems/Services

NSDL (National Science Digital Library)

<http://nsdl.org/about/index.php?pager=projects>

Overzicht van projecten

OCLC projecten

<http://www.oclc.org/research/projects/default.htm>

Algemeen overzicht van projecten

<http://www.oclc.org/research/software/scorpion/>

Scorpion project van OCLC

http://www.oclc.org/research/projects/auto_class/

Automatisch classificeren binnen Scorpion

<http://www.oclc.org/research/projects/fastac/>

FAST as a knowledge base for automatic classification

<http://www.oclc.org/research/projects/mswitch/default.htm>

Metadata Switch: diverse metadata-projecten

Open Directory Project

<http://dmoz.org>

web directory met 590.000 onderwerpscategorieën

PEKING

<http://www.cs.ru.nl/peking/>

project afgesloten 2003

Portal Thesaurus Project

<http://www.e.govt.nz/archive/standards/nzxls/interim-thesaurus>

<http://www.e.govt.nz/standards/nzxls/thesauri>

Overheidsthesaurus voor E-Government in NieuwZeeland

Projects addressing or relating to interoperability issues

<http://www.und.nodak.edu/dept/library/Departments/abc/SACSEM-InteroperabilityProjects-Lois.htm>

overzichtspagina met links naar 18 projecten + bibliografie

QUT (Queensland University of Technology)

http://sky.fit.qut.edu.au/~middletm/cont_voc.html

links naar vrij toegankelijke gecontroleerde vocabulaires

RENARDUS

<http://www.renardus.org/>

integratie van Europese subject gateways (afgesloten in 2002)

SCOT (Schools Online Thesaurus project)

<http://www.curriculum.edu.au/scis/partnerships/scot.htm>

thesaurus voor beschrijven van materiaal voor primair onderwijs.

SLAIS (School of Library, Archival & Information Studies)

<http://www.slais.ubc.ca/resources/indexing/database1.htm>

bibliografie en links over gecontroleerd vocabulaire

STITCH

<http://www.cs.vu.nl/STITCH/>

Semantic Interoperability To access Cultural Heritage (bij CATCH)

SWAD-Europe Thesaurus Activity

<http://www.w3.org/2001/sw/Europe/reports/thes/>

Europees project voor thesaurus-activiteiten met gebruik van SKOS

TASI (Technical Advisory Service for Images)

<http://www.tasi.ac.uk/resources/vocabs.html>

overzicht van online beschikbare gecontroleerde vocabulaires

Taxonomy Warehouse

<http://www.taxonomywarehouse.com/>

overzicht van beschikbare gecontroleerde vocabulaires

Thunderstone

<http://search.thunderstone.com/texis/websearch>

door automatisch classificeren gevulde webgids

Webbrain

http://www.webbrain.com/html/default_win.html

visualisatie-tool voor de Open Directory

Willpower

<http://www.willpowerinfo.co.uk/thesbibl.htm>

bibliografie op het terrein van thesauri, links naar lijsten van thesauri

Wordnet

<http://www.cogsci.princeton.edu/>

Wordnet semantisch netwerk / upper ontology

XMDR (eXtended MetaData Registry (XMDR) Project)

<http://www.xmdr.org/>

project voor standaardisatie van Metadata Registries

Zthes

<http://zthes.z3950.org/>

protocol voor thesaurus-representatie, -toegang en -navigatie

Gebruikte afkortingen

AAT	Art & Architecture Thesaurus
AI	Artificial intelligence
ALA	American Library Association
BLISS	(geen afkorting maar de eigen naam van de ontwerper van de gelijknamige facet-classificatie)
BT	Broader Term (in thesaurus)
CDWA	Categories for the Description of Works of Art
CLEF	Cross-Language Evaluation Forum
CLIR	Cross-Language Information Retrieval
CORC	Cooperative Online Resource Catalog
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
DDC	Dewey Decimale Classificatie
ERIC	Educational Resources Information Center
FAO	Food & Agricultural Organization (van de Verenigde Naties)
FAST	Faceted Application of Subject Terminology
GERHARD	German Harvest Automated Retrieval and Directory
ISO	International Organization for Standardization
JISC	Joint Information Systems Committee
LCC	Library of Congress Classification
LCSH	Library of Congress Subject Headings
LOM	Learning Object Metadata
MARC	MAchine-Readable Cataloging
MeSH	Medical Subject Headings
MILOS	Maschinelle Indexierung zur erweiterten Literaturschließung in Online-Systemen
MODS	Metadata Object Description Schema
NT	Narrower Term (in thesaurus)
NWO	Nederlandse organisatie voor Wetenschappelijk Onderzoek
OAI-PMH	Open Archive Initiative - Protocol for Metadata Harvesting
OCLC	Online Computer Library Center
OCR	Optical Character Recognition
OPAC	Online Public Access Catalog
OWL	Web Ontology Language
RAMEAU	Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié
RDF	Resource Description Framework
RT	Related Term (in thesaurus)
SKOS	Simple Knowledge Organisation System
SRU/SRW	Search-Retrieve by URL / Search-Retrieve Webservice
SWD/RSWK	Schlagwortnormdatei/Regeln für den Schlagwortkatalog
STW	Standard-Thesaurus Wirtschaft
TREC	Text REtrieval Conference
UDC	Universele Decimale Classificatie
UMLS	Unified Medical Language System
VRA	Visual Resources Association
XML	Extensible Mark-up Language