Inhoudelijk toegankelijk maken van hybride bibliotheekcollecties

Een verkennend onderzoek naar huidige opvattingen, recente ontwikkelingen en toekomstverwachtingen

Eric Sieverts

Onderzoek uitgevoerd ten behoeve van de Koninklijke Bibliotheek, Den Haag

Opdrachtgever:	Koninklijke Bibliotheek	Stuurgroep TOO
Onderzoeker:	Dr. Eric G. Sieverts	
Versie 1.0	voorlopig deelrapport	07-07-2004
Versie 2.0	voorlopig eindrapport	22-10-2004
Versie 2.1	eindrapport	31-10-2004

Only librarians like to search, everyone else likes to find

Roy Tennant, California Digital Library

Wie heeft gevonden, heeft niet goed gezocht

Motto op de muur van Restaurant Zeldenrust, Den Haag

Inhoud

Inleiding

Bij de Koninklijke Bibliotheek, de Nederlandse nationale bibliotheek, is een project gestart onder de titel "Toekomst van de onderwerpsontsluiting". Als voorbereiding op dit project en op het uiteindelijke besluitvormingsproces over de onderwerps-ontsluiting, is een verkennend onderzoek uitgevoerd. Op grond daarvan wordt in dit rapport een overzicht gegeven van de "State of the art" op een aantal terreinen die nauw met de problematiek van de onderwerpsontsluiting samenhangen. Dit overzicht is tot stand gekomen op basis van desk-research (literatuuronderzoek en onderzoek op het web) en mondelinge of schriftelijke bevraging van een aantal toonaangevende experts op deze terreinen. Vooral op basis van de meningen en inzichten van deze experts wordt ook getracht enige verwachtingen voor toekomstige ontwikkelingen te formuleren.

De concrete onderzoeksvragen op basis waarvan dit onderzoek is uitgevoerd, luidden:

1. Zoeken vs. ontsluiten
- In hoeverre zijn technieken en mogelijkheden van information retrieval uit digitaal aanwezige tekst-informatie al zo geavanceerd dat geen behoefte meer bestaat aan gecontroleerde inhoudelijke ontsluiting van dit materiaal?

2. Gecontroleerde ontsluitingsmethoden
- Wat is de state-of-the-art op het terrein van thesauri, classificaties, taxonomieën, ontologieën?
- Wat zijn de huidige opvattingen over toepassing van deze ontsluitingsmethoden?
- Hoe kunnen deze methoden tegen elkaar worden afgewogen in situaties zoals bij de KB?

3. Automatische classificatie/verrijking van digitaal materiaal
- Wat is de state-of-the-art op het terrein van automatisch inhoudelijk karakteriseren van digitale documenten, zodat ze automatisch met thesaurustermen verrijkt of in klassen van een taxonomie kunnen worden ingedeeld?
- Wat zijn in het bijzonder randvoorwaarden hierbij, zoals aard van het te verrijken materiaal, maximum aantal klassen, e.d.?
- Hoeveel menselijke inspanning vereist het afregelen en trainen en het onderhoud van dergelijke systemen?

4. Kenmerkende verschillen in aanpak tussen digitaal en papieren materiaal
- Welke van de geautomatiseerde technieken voor digitale documenten kunnen ook worden toegepast op slechts beperkt digitaal aanwezige informatie?
- Welke methoden zijn er om van papieren materiaal op eenvoudige, geautomatiseerde wijze een voldoende uitgebreide digitale representatie te verkrijgen?

5. Mogelijkheid/wenselijkheid van een geïntegreerd uniform systeem voor gehele collecties
- Is het noodzakelijk het digitale deel van een collectie volledig op dezelfde wijze inhoudelijk toegankelijk te maken als het papieren gedeelte?
- Is het mogelijk en haalbaar om het digitale deel van een collectie volledig op dezelfde wijze inhoudelijk toegankelijk te maken als het papieren gedeelte?

Waar in de tekst personen met vermelding van een jaartal worden aangehaald, betreft het literatuurverwijzingen. Waar bij personen geen jaartal wordt vermeld, gaat het om meningen van de betrokkenen die uit vraaggesprekken of vraagformulieren verkregen zijn. Een selectie van de meest relevante geraadpleegde literatuur en websites en een lijst van de geraadpleegde experts is te vinden aan het eind van dit rapport.

Eric Sieverts

Samenvatting

Onder de term Knowledge Organisation Systems worden alle soorten systemen voor gecontroleerde inhoudelijke ontsluiting gevangen. Naast de klassieke soorten, als classificaties en thesauri, vallen hieronder ook nieuwe als ontologieën en Topic Maps.

Classificaties zijn vooral een belangrijk hulpmiddel om de gebruiker, al bladerend door de systematische indeling van categorieën bij de juiste categorie te laten uitkomen. Sinds kort staan ze weer sterk in de belangstelling. Dat is vooral een gevolg van het toenemende belang van het World Wide Web en de daar gebruikte browse- en linking-technieken. Klassieke bibliotheekclassificaties moeten wel sterk worden aangepast voor dergelijke toepassingen, om tot voldoende overzichtelijke scherm-presentaties en gemakkelijke bruikbaarheid te komen. Vooral van Dewey (DDC) en Library of Congress (LCC) classificaties bestaan intussen voorbeelden van dergelijke aanpassingen. Hoewel veel van dit soort projecten in aanzet vooral gericht zijn op ontsluiting van webpagina's, zijn ervaringen en conclusies voor met name de presentatiekant ook van belang voor het toegankelijk maken van - al dan niet digitale - bibliotheekcollecties. Belangstelling voor facetclassificaties neemt eveneens weer toe, maar aansprekende bibliotheektoepassingen zijn nog moeilijk te vinden. In het bedrijfsleven worden dergelijke "meerdimensionale" systemen, onder de noemer van taxonomieën, wel toegepast.

Thesauri zijn vooral een belangrijk hulpmiddel voor zoeksystemen waarin de gebruiker zelf zijn vraag kan formuleren. Bij de toepassing van thesauri liggen de belangrijkste ontwikkelingen op het terrein van de matching van zoektermen van gebruikers met de daarmee corresponderende thesaurustermen. Zo wordt getracht de discrepantie tussen gebruikerstaal en systeemtaal te overbruggen. Om ingetikte zoek-woorden van gebruikers met meer zekerheid met de juiste en meest in aanmerking komende thesaurustermen te associëren, kunnen dialoogsystemen gebruikt worden, zoals die nu voor natural language retrieval-systemen worden ontwikkeld. Bij voldoende strikt hiërarchisch gestructureerde thesauri vormt automatische expansie van zoekvragen met specifiekere termen uit de thesaurus een belangrijk middel om de recall van zoekacties te verbeteren.

In zowel ontologieën als topic maps kunnen op geformaliseerde wijze allerlei soorten relaties worden gelegd tussen begrippen. Daardoor zou informatie in principe beter toegankelijk gemaakt moeten kunnen worden, dan met de klassieke ontsluitings-methoden. In de praktijk bestaan nog geen toepassingen voor grote algemene informatiecollecties. Toepassing lijkt vooralsnog erg arbeidsintensief te zijn.

Bij toepassing van classificaties voeren DDC en LCC internationaal de boventoon. 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 onderwerps-domeinen mogelijk te maken. Voor precieze zoekacties daarin, zal toch van thesauri of van free-text/full-text zoeken gebruik gemaakt moeten worden. In hoeverre een systeem als van de Nederlandse GOO trefwoorden daarbij als thesaurus geschikt is, dient nog nader te worden onderzocht.

Computeranalyse van digitaal beschikbare teksten - meestal in een combinatie van statistische, linguïstische en regelgebaseerde technieken - kan resulteren in tamelijk betrouwbare automatische karakterisering van de teksten. Als een daarmee verkregen "vingerafdruk" van een document in een free-text zoeksysteem wordt gebruikt als substituut voor het oorspronkelijke document, kan dat al leiden tot betere precisie (en soms zelfs betere recall) van zoekacties. Deze vingerafdrukken kunnen ook gebruikt worden om, na training met voorbeeldmateriaal, nieuwe documenten met redelijke betrouwbaarheid te categoriseren of van trefwoorden te voorzien op basis van een bestaande classificatie of thesaurus. Door het instellen van betrouwbaarheidsdrempels kan handmatige ontsluiting dan beperkt blijven tot een klein percentage voor het computersysteem moeilijke gevallen. Dergelijke technieken kunnen ook worden toegepast als ondersteuning bij puur handmatige ontsluiting, opdat de menselijke indexeerder consistenter en vollediger termen of klassen toekent. Voor succesvolle automatische toepassing dient het aantal klassen of thesaurustermen liefst niet groter te zijn dan ca. 5000, en dienen deze onderling voldoende "orthogonaal" te zijn. Dit aantal stelt dus ook 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.

Voor materiaal waarvan onvoldoende tekst digitaal beschikbaar is, kunnen inhoudsopgaven en andere voor de inhoud representatieve onderdelen via scannen en OCR verwerkbaar gemaakt worden. Daarnaast zal men van steeds meer boeken relevante tekstonderdelen digitaal kunnen aanschaffen. Op basis hiervan kunnen de in eerdere hoofdstukken beschreven geautomatiseerde technieken worden toegepast.

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. In het algemeen wordt daarbij uitgegaan van collecties die allemaal al van gecontroleerde ontsluiting voorzien zijn. Collecties waarin alleen op basis van full-text retrieval wordt gezocht, worden daar niet ook bij betrokken. Andersom zal het bij toepassing van full-text retrieval als overkoepelend zoeksysteem 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.

Inhoudelijke ontsluiting van die delen van collecties die uit "fictie" bestaan, is in principe wel mogelijk. Toepassing van eenzelfde gedetailleerde ontsluiting als voor non-fictie biedt echter maar weinig toegevoegde waarde. De problematiek van de juiste afweging tussen zeer specifieke ontsluiting van hoog-specialistisch materiaal, zoals wetenschappelijke artikelen, en de veel minder specifieke concepten waarmee bijvoorbeeld leerboeken, algemene werken en verzamelwerken worden ontsloten, is wat minder eenvoudig op te lossen.

Aangezien de beste full-text retrieval nog altijd geen panacee biedt voor alle soorten zoekgebruik, lijkt gecontroleerde ontsluiting van alle materiaal op een vrij algemeen niveau, in combinatie met free-text zoeken op specialistische onderwerpen, op dit moment de beste oplossing te bieden om zowel de algemene zoeker als de specialistische wetenschappelijke zoeker voldoende aan zijn trekken te laten komen.

1. Methoden van inhoudelijke ontsluiting

1.1 Recente ontwikkelingen / state of the art

Onder de term Knowledge Organisation Systems worden alle soorten systemen voor gecontroleerde inhoudelijke ontsluiting gevangen. Van de klassieke soorten kennen classificaties recent een revival, vooral vanwege het browse-gemak dat daarmee gerealiseerd kan worden. Voor zoeken gaat nog steeds de belangstelling uit naar thesauri, hoewel gebruikers de zoekmogelijkheden daarvan vaak nog onvoldoende benutten. Met technische middelen valt daar echter wel wat aan te doen. Ontologieën en Topic Maps zijn nieuwe voorbeelden van Knowledge Organisation Systems.

De laatste jaren mag de standaardisatie van metadatasystemen zich in een snel groeiende belangstelling verheugen. Daarbij is het opmerkelijk dat vaak veel meer nadruk wordt gelegd op standaardisatie van de formats dan van de inhoud zelf 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.

Het is opmerkelijk dat zich, gedeeltelijk wat los van de groepen die zich specifiek met deze afzonderlijke klassieke ontsluitingsmethoden bezig houden, een onderzoekstak heeft ontwikkeld welke 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, die in dit rapport in een afzonderlijk hoofdstuk aan de orde komen, tot het werkterrein van de KOS behoren. In het kader van KOS-onderzoek wordt er ook aandacht aan besteed hoe gebruikers toegang moeten krijgen 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 wereldwijde opkomst van web-gebaseerde systemen. Voor het browsen door en het aanklikken van onderwerpsrubrieken zijn classificaties namelijk aanzienlijk geschikter dan woordsystemen zoals thesauri. Zij bieden in wezen een reactieve zoekwijze, waarbij de gebruiker niet zelf zijn onderwerp hoeft te omschrijven, maar waar hij uit gepresenteerde omschrijvingen van categorieën die keuze kan aanklikken die het meest in overeenstemming lijkt met zijn informatie-behoefte. In de uiteindelijk gekozen eindcategorie 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 kunnen bestaan uit een door de gebruiker geformuleerde combinatie van de concepten die in de gewenste documenten aan de orde moeten komen. Een hinderpaal bij toepassing van thesauri aan de gebruikerskant is meestal de discrepantie tussen de belevings-wereld van de gebruiker en het binnen een thesaurus vastgelegde voorkeurs-vocabulair. Dit is een van de oorzaken van het vaak gesignaleerde ondergebruik van dergelijke ontsluitingssystemen (zie bijvoorbeeld het overzichtsartikel over thesauri van Aitchison (2004) en een studie van Greenberg (2004)). Aan ontwikkelingen om iets aan deze discrepantie te doen en daarmee de mate van gebruik - soms zelfs voor de gebruiker ongemerkt - te stimuleren, wordt in §1.1.4 apart aandacht besteed.

Behalve dat gecontroleerde ontsluitingssystemen vaak als gebruiksonvriendelijkheid beschouwd worden, heeft de handmatige toekenning van categorieën of thesaurus-termen als grootste nadeel dat daar vaak onevenredig veel tijd in geïnvesteerd moet worden. Daarom is het interessant dat vooral voor classificaties, maar ook wel voor thesauri, geautomatiseerde methoden bestaan waarmee deze toekenning door de computer gedaan kan worden. Gezien het belang van deze methoden, wordt daarop in hoofdstuk 3 afzonderlijk ingegaan.

Naast klassieke ontsluitingssystemen als classificaties en thesauri, wordt ook geëxperimenteerd met nieuwe soorten systemen, zoals ontologieën en topic maps. Daaraan wordt in §1.1.4 aandacht besteed. Overigens is het woord ontologie tot een soort modewoord geworden, waaronder in de praktijk uiteenlopende soorten ontsluitingsmethoden worden verstaan. Volgens Van der Vet blijkt hieruit wel de convergentie tussen die verschillende methoden van ontsluiting, hetgeen door hem als een zeer veelbelovende ontwikkeling wordt gezien. Een ander voorbeeld van een dergelijke convergentie, dat onder andere door Brazier wordt genoemd, is de koppeling die steeds vaker wordt aangebracht tussen toonaangevende verschillend-soortige systemen zoals de Dewey-classificatie (DDC) en het Library of Congress trefwoordvocabulaire (LCSH).

1.1.1 Classificaties

Classificaties zijn vooral een belangrijk hulpmiddel om de gebruiker, al bladerend door de systematische indeling van categorieën bij de juiste categorie te laten uitkomen. Sinds kort is er een sterke herleving van de belangstelling voor classificaties. Dat is vooral een gevolg van het toenemende belang van het World Wide Web en de daar gebruikte browse- en linking-technieken. Klassieke bibliotheek-classificaties moeten wel sterk worden aangepast voor dergelijke toepassingen, om tot voldoende overzichtelijke schermpresentaties en gemakkelijke bruikbaarheid te komen. Vooral van Dewey (DDC) en Library of Congress (LCC) classificaties bestaat intussen een aantal voorbeelden van dergelijke aanpassingen. Hoewel veel van dit soort projecten in aanzet vooral gericht zijn op ontsluiting van webpagina's, zijn ervaringen en conclusies voor met name de presentatiekant van gelijkelijk belang voor het toegankelijk maken van - al dan niet digitale - bibliotheekcollecties.
Ook de belangstelling voor facetclassificaties neemt weer toe. Aansprekende bibliotheektoepassingen daarvan zijn echter nog moeilijk te vinden. Anderzijds worden in het bedrijfsleven, onder de noemer van taxonomieën, wel dergelijke "meerdimensionale" systemen toegepast.

Classificaties zijn enige tijd beschouwd als achterhaald en van beperkt nut bij het toegankelijk maken van grote hoeveelheden informatie in een digitale omgeving. Dit hing vooral samen met de opkomst van retrievalsystemen. Woordsystemen als thesauri maken het de zoekspecialist inderdaad eenvoudiger om op inhoudelijke elementen te zoeken, dan classificatiesystemen met hun vaak complex opgebouwde codes. In dit verband kunnen de hoogste niveaus van classificaties wel nog van nut zijn 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 de 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. Met de introductie van het begrip taxonomie voor informatie-verzamelingen 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 UDC moeilijk te vertalen zijn in overzichtelijke in woorden omschreven keuzemenu's en aanklikbare hiërarchieën. Riesthuis gaf aan dat internationaal de belangstelling op dit moment vooral uitgaat naar de Dewey Decimale Classificatie, 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 Congres Classificatie bestaat in dit opzicht minder hernieuwde belangstelling. De Nederlandse Basisclassificatie acht hij voor dit soort inhoudelijke ontsluiting - om deels andere redenen - eveneens ongeschikt.

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 volgordes, 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 ook goede 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 LCSH. Ook hij geeft aan dat DDC in principe een beter uitgangspunt had geboden, maar dat moest hij in zijn situatie als politiek onhaalbaar verwerpen. Ook voor de Duitse Nationale Bibliografie wordt op dit moment 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 een apart hoofdstuk aan de orde.

Evenals Davis (2002) benadrukt ook Van Gent dat degenen die boomstructuren gebruiken, niet te vaak willen klikken om bij hun informatie te komen. Dat stelt 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. Bij een maximum van drie niveaus en niet meer dan circa vijftien vervolgkeuzes per categorie, zou dat het aantal mogelijke categorieën beperken tot maximaal ruim 3000. Anderzijds zien we bij de grote onderwerpstaxonomieën op het web (Yahoo, OpenDirectory) in de praktijk juist 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 soms wel acht niveaus diep. Dat is ook weer niet zo verwonderlijk want in dergelijke grote systemen met enkele miljoenen items, zou anders het aantal items per categorie onhanteerbaar groot worden.

Van Gent meent dat bij dergelijke grote aantallen, die veelal voortkomen uit precoördinatieve combinaties van meer onderwerpsfacetten, beter van trefwoorden gebruik gemaakt zou kunnen worden. Bij de webtaxonomieën wordt de zoekfunctie dan ook vaak aangeraden als methode om snel de juiste rubrieken voor een bepaald onderwerp te vinden. 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.

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 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. In de praktijk blijkt de basis van de speciaal voor het web ontwikkelde classificaties vaak een heel andere te zijn dan die van de klassieke bibliotheekclassificaties. Dat blijkt uit een

vergelijking door Hudon (2001) van Yahoo en een Canadese portal 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 praktisch 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. Beagle (2003) heeft dat ook voor een aanpassing van LCC gedaan, bij Belmont Abbey College (Canada). Ook hij combineert dit met een zoekmogelijkheid als alternatieve methode om de gebruiker bij de juiste rubriek te laten uitkomen.

In het kader van webtoepassingen, noemt Riesthuis ook het belang van concordanties - in de praktijk met name tussen DDC en LCC - waarmee het mogelijk wordt 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 door-zocht 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 hier uitegbreider op teruggekomen.

Los van eventuele webtoepassingen, ziet Riesthuis bij classificaties op dit moment een tendens in de richting van facetsystemen. Dat is juist in UDC weer veel makkelijker te realiseren, dan in de veel sterker precoördinatieve LCC. Voor UDC bestaan ook 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. In dit opzicht is UDC dus juist weer veel geschikter dan DDC om als basis voor een zoeksysteem te dienen. Toch beschrijft Pollitt (1998) hoe juist DDC aangepast kan worden tot een facetclassificatie voor toegang tot online informatie. De volledig facet-gebaseerde BLISS classificatie wordt ook specifiek voor web-bronnen opnieuw onder de aandacht gebracht (Broughton 2000).

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. Voor verschillende invalshoeken waar-onder informatie toegankelijk gemaakt kan worden, worden dan afzonderlijke, meestal vrij eenvoudige classificaties of taxonomieën ontwikkeld, waarin onafhankelijk gebrowsed kan worden. Als resultaat wordt telkens de doorsnede van de geselecteerde, als verschillende dimensies te beschouwen categorieën getoond. Een bekende software-leverancier als Verity noemt dit zijn "parametric search".

Riesthuis ziet verder geen fundamenteel nieuwe ontwikkelingen op het terrein van classificaties. Wel meent hij dat er aanwijzingen zijn dat ontsluiting met behulp van een classificatiesysteem minder gevoelig is voor menselijke fouten en inconsistenties dan ontsluiting met behulp van een thesaurus.

1.1.2 Thesauri

Thesauri zijn vooral een belangrijk hulpmiddel voor zoeksystemen waarin gebruikers zelf hun vraag kunnen formuleren. Bij toepassing van thesauri liggen de belangrijkste ontwikkelingen op het terrein van de matching van zoektermen van gebruikers met corresponderende thesaurustermen. Zo kan de discrepantie tussen gebruikerstaal en systeemtaal overbrugd worden. Bij voldoende strikt hiërarchisch gestructureerde thesauri vormt automatische expansie van zoekvragen met specifiekere termen uit de thesaurus een belangrijk middel om de recall van zoekacties te verbeteren.

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 tendens naar postcoördinatie ziet Riesthuis de laatste tijd nog sterker worden, zij het ook weer niet zo extreem als bij het Uniterm-systeem. Deze tendens zien we bijvoorbeeld ook bij een gefaceteerde toepassing van LCSH in FAST bij OCLC (O'Neill, 2001, 2003). Deze facettering blijkt zich in de praktijk echter 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 nadeel dat ze niet erg gebruiksvriendelijk zijn. Hoe weet de zoeker immers welke termen de "juiste" zijn waarop gezocht moet worden? Omdat thesauri vrijwel nooit - zoals classificaties - vanuit een beperkt aantal hoofdcategorieën zijn opgebouwd, lenen ze zich 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 eerste methode is die waarbij elke thesaurusterm geassocieerd is met een voldoende groot aantal synonieme of quasi-synonieme termen, zodat bij het gebruik van enig van die termen automatisch de juiste thesaurusterm in de zoekactie gesubstitueerd kan worden. Dit associëren kan onder meer gebeuren:
- doordat een voldoende groot aantal synoniemen en nauw verwante woorden en begrippen als verwijstermen in de thesaurus is opgenomen (zo worden in PubMed, MeSH-termen die met ingetikte zoekwoorden corresponderen, automatisch aan zoekvragen toegevoegd); 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, bijvoorbeeld via logfiles, de door gebruikers gebezigde zoektermen te analyseren en die zo goed mogelijk af te beelden op de termen uit de thesaurus, zodat de mapping tussen gebruikerstaal en systeemtaal geleidelijk steeds volledíger wordt; in feite leidt dit ook tot zo'n user-thesaurus;

- door de 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.

In het FACET-project worden zulke 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 descriptoren die met verschillende betekenissen van de oorspronkelijke zoekterm corresponderen, dan kan dat al snel leiden tot een onaanvaardbaar slechte precisie van de zoekresultaten. Aan methoden van gebruikersinteractie om daar weer iets aan te doen, zal in §1.1.4 nog aandacht worden besteed. In een - overigens niet zeer representatief - onderzoek door Greenberg (2004), bleek dat gebruikers van systemen die zoektermen met thesaurustermen proberen te associëren in overgrote meerderheid eigen keuzevrijheid, middels selectie van termen uit keuzelijstjes, prefereerden.

Wanneer ook titels, samenvattingen en eventueel nog meer niet-geformaliseerde tekst van documenten digitaal beschikbaar is en dus doorzocht kan worden, is nog een andere methode mogelijk. Zoekresultaten die zijn verkregen op basis van de door de zoeker gebruikte zoekwoorden, kunnen statistisch geanalyseerd worden op de daarin aanwezige thesaurustermen, waarna de zoekactie - al dan niet automatisch - op basis van die termen wordt 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 is het zeker gewenst dat de gebruiker zelf kan kiezen welke thesaurustermen uiteindelijk wel en niet 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 onvoldoende bewust van het feit dat het zoeken op specifiekere begrippen in collecties met veel gespecialiseerd materiaal, zoals wetenschappelijke artikelen en hoofdstukken uit monografieën, vrijwel altijd veel meer resultaten oplevert dan zoeken op meer algemene begrippen. Greenberg (2001-1,2) 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 expansion met ruimere en verwante begrippen (BT en RT)) ging dit 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 innovatieve toepassing van een thesaurus wordt beschreven door Silveira (2004). Daarbij 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.

1.1.3 Ontologieën en topic maps

In zowel ontologieën als topic maps kunnen op geformaliseerde wijze allerlei soorten relaties worden gelegd tussen begrippen. Daardoor zou informatie in principe beter toegankelijk gemaakt moeten kunnen worden, dan met de klassieke ontsluitings-methoden. In de praktijk bestaan nog geen toepassingen voor grote algemene informatiecollecties. Toepassing lijkt vooralsnog erg arbeidsintensief te zijn.

Vooral in het kader van het semantisch web, wordt nadruk gelegd op kennis-representaties in de vorm van ontologieën. Daarin wordt getracht op geformaliseerde, computerinterpreteerbare wijze een gedetailleerde beschrijving te geven van (een stukje van) de werkelijkheid. Daarbij kunnen aan begrippen allerlei rollen worden toegekend en kunnen allerlei soorten relaties tussen woorden of begrippen worden gedefinieerd. Door de computer bij het zoeken naar informatie ook van die rollen en relaties gebruik te laten maken, zou toepassing van dergelijke kennisrepresentaties vooral tot een hogere precisie moeten leiden, al wordt ook wel van hogere recall gesproken. Echt hard bewijs voor verbeterde zoekresultaten bestaat echter nog niet.

Een voorbeeld van het omwerken van een bestaande thesaurus van de FAO naar een ontologie wordt gegeven door Soergel (2004). Daarin bespreekt hij ook tekort-komingen van klassieke thesauri. Over ontologieën die hij "semantically rich knowledge organization systems" noemt, heeft Soergel hooggestemde verwachtingen. Zij moeten uiteindelijk leiden tot:
- betere gebruikersinteractie en daardoor betere queries,
- automatische intelligente query-expansie,
- intelligente ondersteuning van menselijke indexeerders en betere automatische indexering,
- ondersteuning van semantisch web toepassingen.

Voorlopig lijkt het bouwen en toepassen van dergelijke ontologieën echter alleen realiseerbaar voor zeer duidelijk afgeperkte onderwerpsdomeinen. Voor bredere terreinen loont het zeker nog niet.

Vanuit de XML-wereld zijn ook "topic maps" geïntroduceerd als middel om informatie op een intelligente manier toegankelijk te maken. Ook daarin kunnen allerlei rollen en relaties worden vastgelegd. Veel sterker dan bij ontologieën gaat het hier echter om formele concepten waartussen de relaties liggen. Die concepten op hun beurt kunnen weer door allerlei woorden en begrippen - ook in verschillende talen - gerepresenteerd worden. Het al voor ontologieën genoemde nadeel dat toepassing zich voorlopig beperkt tot specialistische onderwerpsdomeinen, lijkt ook hiervoor geldig te zijn. Hoewel topic maps voor verschillende domeinen in principe koppelbaar zouden moeten kunnen zijn, is niet te verwachten dat dit al op korte termijn tot toepassingen voor zeer brede collecties zal leiden.

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

1.1.4 Interactie tussen gebruiker en ontsluitingssysteem

Om ingetikte zoekwoorden van gebruikers met meer zekerheid met de juiste en meest in aanmerking komende thesaurustermen te associëren, kunnen dialoogsystemen gebruikt worden, zoals die nu voor natural language retrieval-systemen worden ontwikkeld.

Zoals eerder aangegeven, is een belangrijke hinderpaal bij het nuttig gebruiken van gecontroleerde ontsluitingssystemen tijdens het zoekproces, dat er een discrepantie bestaat tussen het door de gebruiker in zijn zoekvraag gehanteerde vocabulaire en dat van het systeem. 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 reactieve wijze van gebruik van de meeste van dergelijke systemen - de gebruiker kiest uit de gepresenteerde rubrieksomschrijvingen die welke waarschijnlijk het best correspondeert 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 eerder 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 bij de juiste thesaurustermen te laten uitkomen. Het is interessant dat Van Gent in dit verband opmerkt dat juist ook vanuit de wereld van de free-text retrieval de aandacht sinds kort uitgaat naar dialoogsystemen, waarmee de computer probeert er achter te komen wat de gebruiker precies bedoelt. Dergelijke dialogen kunnen namelijk 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 uitsluitsel 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. Onder meer in het IMIX-programma van NWO worden projecten uitgevoerd waar dit aspect - weliswaar vooral gericht op spraaktoepassingen - een belangrijke rol speelt.

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 gebeurd (zie hoofdstuk 2). Omdat een zoekvraag over het algemeen veel minder tekst bevat dan een document, zal de zoeker ook hierbij meestal via een dialoogsysteem om aanvullende informatie gevraagd moeten worden.

1.2 Praktische toepassing en voorkeuren

Bij toepassing van classificaties voeren DDC en LCC internationaal de boventoon. 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 documentcollecties en voor specialistische publicaties eigenlijk alleen geschikt om in zoeksystemen voorselectie op bepaalde onderwerpsdomeinen te bieden. Voor precieze zoekacties daarin, zal toch van thesauri of van free-text zoeken gebruik gemaakt moeten worden. In hoeverre de Nederlandse GOO trefwoorden hierbij als thesaurus geschikt zijn, dient nog nader te worden onderzocht.

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. Voorname 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. 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 5 miljoen documenten, zal in een classificatie met 5000 klassen, elke klasse gemiddeld nog 1000 documenten bevatten, hetgeen uiteraard veel te veel is 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 daar gezorgd wordt 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 dat materiaal. Probleem hierbij is wel dat de best uitgewerkte thesauri vrijwel allemaal voor beperkte onderwerpsdomeinen zijn ontwikkeld. In het kader van inter-operabiliteit (zie hoofdstuk 5) worden wel verschillende vocabulaires op elkaar gemapt, maar dat betreft meestal algemene vocabulaires, waarvan de onderwerps-gebieden elkaar vrijwel geheel overlappen.

In het Nederlandse taalgebied vormen de bij GOO behorende trefwoorden wel een vocabulaire dat alle vakgebieden omvat en waarin hiërarchische thesaurusrelaties zijn aangebracht. Zo evenwichtig en consequent opgebouwd als de betere specialistische vakthesauri is het GOO trefwoordsysteem echter nog niet. In vergelijking tot de

vereenvoudigde vormen van classificaties zoals die in een webomgeving worden toegepast, bieden de GOO trefwoorden wel tamelijk specifieke ontsluitingsmogelijkheden. Anderzijds worden ze maar in beperkte mate gebruikt voor de ontsluiting van zeer specialistisch materiaal. Zo specifiek als de voor de ontsluiting van tijdschriftpublicaties ontwikkelde vakthesauri zijn de GOO-trefwoorden dus maar zelden.

1.3 De toekomst van inhoudelijke ontsluiting

Voor de klassieke methoden van inhoudelijke ontsluiting lijken geen baanbrekende nieuwe ontwikkelingen op komst te zijn. Wat dat betreft moeten we overigens wel duidelijk onderscheid maken tussen de technieken voor het opzetten van dergelijke kennisrepresentaties, in de vorm van rubrieksindelingen, taxonomieën of thesauri, en het ontsluitingsproces zelf. Juist bij dat ontsluitingsproces, waarbij materiaal wordt ingedeeld of ontsloten met behulp van dergelijke systemen, staat veel te gebeuren. Daarbij moet gedacht worden aan 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.

Daarnaast zijn technieken in ontwikkeling die tijdens het zoekproces, voor "gewone" zoekvragen van gebruikers, op de achtergrond toch gecontroleerde ontsluitingssystemen toepassen. Dat maakt vrijwel automatisch een einde aan het alom gesignaleerde ondergebruik van de huidige ontsluitingssystemen. In dat kader verwacht Van der Vet op de langere termijn interessante resultaten van op kunstmatige intelligentie (AI) gebaseerde technieken voor een betere interpretatie van informatiebehoeften en bedoelingen van de gebruiker.

Over de acceptatie en toepassing van betrekkelijk nieuwe methoden als ontologieën en topic maps is nog niet veel 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.

2. Geautomatiseerde methoden van inhoudelijke karakterisering, classificatie en verrijking

Computeranalyse van digitaal beschikbare teksten - meestal in een combinatie van statistische, linguïstische en regelgebaseerde technieken - kan resulteren in tamelijk betrouwbare automatische karakterisering van die teksten. Wanneer een daarmee verkregen "vingerafdruk" van een document in een free-text zoeksysteem wordt gebruikt als substituut voor het oorspronkelijke document, kan dat al leiden tot betere precisie (en soms zelfs betere recall) van zoekacties. Deze vingerafdrukken kunnen ook gebruikt worden om, na training met voorbeeldmateriaal, nieuwe documenten met redelijke betrouwbaarheid te categoriseren of van trefwoorden te voorzien op basis van een bestaande classificatie of thesaurus. Door het instellen van betrouwbaarheidsdrempels kan handmatige ontsluiting dan beperkt blijven tot een klein percentage van de documenten, de voor het computersysteem moeilijke gevallen. Dergelijke technieken kunnen ook worden toegepast als ondersteuning bij puur handmatige ontsluiting, opdat menselijke indexeerders consistenter en vollediger termen of klassen zullen toekennen.
Voor succesvolle automatische toepassing dient het aantal klassen of thesaurustermen liefst niet groter te zijn dan ca. 5000, en dienen deze onderling voldoende "orthogonaal" te zijn. Dit aantal stelt dus ook 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.

2.1 Recente ontwikkelingen / state of the art

Er is intussen al een aantal jaren ervaring met systemen die kunnen zorgen 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 er eigenlijk relatief weinig bekend is van de menselijke analyseprocessen die ten grondslag liggen aan handmatige ontsluiting. Al is er een ISO-standaard hoe een indexeerder de onderwerpen van een document moet bepalen, toch spelen daarbij in de praktijk veel meer subjectieve elementen mee, dan bij geautomatiseerd classificeren. Dat geautomatiseerde ontsluiting vaak even goed beoordeeld wordt als handmatige, hoeft dus misschien niet eens zo 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 vingerafdrukken staat intussen een heel arsenaal aan taaltechnologische methoden ter beschikking. Een overzicht van veel toegepaste technieken wordt onder meer gegeven door Anderson (2001, part II). In feite zijn dat voor een belangrijk deel dezelfde soort methoden die ook worden toegepast voor verbetering van free- of 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 kan worden bepaald wat de inhoudelijk belangrijkste en meest karakteristieke woorden uit een document zijn. Met kennisregels kan het belang van bepaalde woorden worden afgeleid, op basis van logische regels, op basis van hun positie, 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 aanwezige woorden tot hun morfologische woordstam, door zogenaamde "decompounding" van samengestelde woorden in hun losse componenten, door syntactische analyse en door beschikbaarheid van semantische kennis over mogelijke betekenisrelaties tussen woorden. In de praktijk blijken deze linguïstische technieken niet altijd allemaal te hoeven worden toegepast om al redelijk betrouwbare resultaten te verkrijgen. Wel wordt vrijwel altijd een combinatie van minstens statistisch + regelgebaseerd of statistisch + linguïstisch toegepast.

Deze 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 gebruikt worden 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 voor alle klassen uit de taxonomie of voor alle termen uit de thesaurus ook kunnen beschikken over dergelijke vingerafdrukken. Elke klasse of thesaurusterm is dan als het ware verrijkt met een verzameling termen die, ieder met een eigen gewicht, gezamenlijk karakteristiek zijn voor de betreffende klasse of thesaurusterm. Men spreekt daarbij ook wel van equivalentieklassen.

Die aan de klassen of termen gehechte vingerafdrukken kunnen geheel door menselijke tussenkomst zijn opgesteld, in de vorm van een soort kennisregels. Veel systemen bieden echter de mogelijkheid deze vingerafdrukken door training van het systeem automatisch tot stand te laten komen. In het bijzonder wanneer een collectie digitale documenten beschikbaar is, die al handmatig is ontsloten op basis van de betreffende taxonomie of thesaurus, kan dit aantrekkelijke mogelijkheden bieden. In een dergelijk geval kan het hele trainingsproces voor het grootste gedeelte geautomatiseerd worden. Daarbij zullen meestal minimaal tien zeer karakteristieke documenten per klasse nodig zijn. Als dergelijk al ontsloten materiaal niet 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 wat grotere taxonomie met bijvoorbeeld 1000 klassen vereist al dat 10.000 van dergelijke karakteristieke documenten worden geselecteerd en handmatig ingedeeld of ontsloten.

Daarnaast is het bovendien 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 van Van Gent 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 leken 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 de titels van wetenschappelijke artikelen wellicht voldoende concreet beschrijvend zijn, dat alleen op grond daarvan al indeling mogelijk is.

Over het algemeen zullen voor het classificatieproces drempelwaarden voor betrouwbaarheid ingesteld kunnen worden. Bij een hoge betrouwbaarheidsdrempel zal een hoog percentage van de geclassificeerde documenten - vaak ruim meer dan 90% - in de correcte klasse(n) terechtkomen, 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, completer en consistenter kan ontsluiten, zeker ook in situaties waar het om niet-professionele ontsluiters gaat.

Een andere interessante mogelijkheid die Van Gent noemt, is om het automatisch classificeren op zoekvragen toe te passen. Daarmee kunnen ook 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. Ook deze techniek beoogt dus een oplossing te bieden voor de 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.

Een uitgebreide bibliografie op het terrein van geautomatiseerde ontsluiting is samengesteld door Fabrizio Sebastiani (Bibliography on Automated Text Categorization; http://faure.iei.pi.cnr.it/~fabrizio/). In oktober 2004 bevatte deze bibliografie verwijzingen naar 509 artikelen, congresbijdragen en andere soorten publicaties.

2.2 Vereisten voor het toepassen van geautomatiseerde inhoudelijke ontsluiting

In dit verband beperk ik me 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 dus een noodzakelijke voorwaarde dat respectievelijk een classificatie of taxonomie, dan wel een thesaurus of andere collectie gecontroleerde termen, al beschikbaar zijn. Voorts dient het systeem, zoals in de vorige paragraaf beschreven, op alle daarin voorkomende klassen of termen getraind te zijn.

Van Gent geeft aan dat ontsluiting met dergelijke systemen nog goed kan werken voor schema's met maximaal circa 5000 klassen. Voor schema's met grotere aantallen termen raadt hij aan om voor de meest specifieke begrippen daaruit geen gecontroleerde ontsluiting toe te passen, maar bij het zoeken op full-text retrieval te vertrouwen. Nog even afgezien van de al genoemde problematiek van het trainen van zeer grote aantallen klassen, kan het daarbij namelijk ook een probleem worden dat de computer steeds moeilijker onderscheid kan maken tussen verschillende in de praktijk nauw verwante begrippen. Als illustratie hierbij kan enige eigen ervaring op dit gebied dienen. Bij een voor de UB Utrecht uitgevoerd experiment bleek bijvoorbeeld dat het gebruikte systeem nauwelijks onderscheid kon maken tussen artikelen op het terrein van psycholinguïstiek, die ontsloten waren met de trefwoorden "leren lezen", "dyslexie" en "spelling". Vingerafdrukken voor deze termen bleken na training - misschien niet echt verwonderlijk - onvoldoende onderscheidend. Dit illustreert dat "orthogonaliteit" van de klassen uit de taxonomie of de thesaurus - d.w.z. niet te veel inhoudelijke overlap - ook een eis voor succesvol automatisch classificeren is.

Anderzijds bleek uit intussen al 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.

Van de te ontsluiten informatie dient uiteraard een voldoende hoeveelheid digitaal beschikbaar te zijn. Wat in dit verband "voldoende" is, kan per domein of document-type verschillen. Voor teksten met zeer concreet en specifiek taalgebruik kunnen korte samenvattingen van tien regels al voldoende zijn. In andere gevallen - of waar geen tekst met het karakter van een samenvatting beschikbaar is - zal meer tekst digitaal beschikbaar moeten zijn.

2.3 Praktische toepassingen van systemen voor geautomatiseerde inhoudelijke ontsluiting

Op het terrein van bedrijfsinformatie en kennissystemen worden vormen van geautomatiseerde inhoudelijke ontsluiting al enige tijd toegepast. Toepassingen in een meer klassieke bibliotheekomgeving zijn er ook, vrij veel daarvan overigens in de vorm van projecten, waarvan niet altijd duidelijk werd of ze uiteindelijk tot in de praktijk werkzame systemen hebben geleid. Een overzicht van projecten waarbij voor de ontsluiting vooral wordt uitgegaan van bestaande bibliotheekclassificaties wordt gegeven door Toth (2002). De belangrijkste die in dat artikel genoemd worden, komen ook in het overzicht hieronder aan de orde. (URL's van de de vermelde projecten zijn achterin dit rapport 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 zorgt de module AUTINDEX voor automatische indexering en classificatie, tijdens het productieproces van documenten, in zowel Engels als Duits. Hierin wordt gebruik gemaakt van bestaande thesauri op de terreinen van techniek en fysica. Er lijken geen recentere gegevens beschikbaar te zijn dan uit 2001.

CARMEN
Content Analysis, Retrieval and MetaData: Effective Networking. 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.

CORA
CORA was een portal en zoekmachine voor artikelen op het terrein van "computer science". Op basis van "machine learning" technieken werden artikelen automatisch in de categorieën van de portal ingedeeld. De hierbij gebruikte technieken worden beschreven door McCallum (1999, 2000). Op dit moment is het CORA-systeem niet meer online beschikbaar.

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 gebruikt van de Engineering Index classificatie. COMBINE was de harvester om op te nemen informatie te oogsten (OAI bestond 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.

ET-MAP
Prototype voor een categorisatiesysteem voor webpagina's, ontwikkeld door het Artificial Intelligence Lab van de University of Arizona. Het project liep van 1994 tot 1998.

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).

INTERSPACE
Project waarin prototypes werden ontwikkeld voor "concept extraction" van tekst- en image-bronnen, voor het automatisch genereren van thesauri (concept space generation), voor categorievisualisatie (category map generation) en voor het automatisch toekennen van onderwerpsontsluiting (concept assignment). Het project werd afgesloten in 1998.

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 woord-betekenissen is daar een belangrijk aspect van. In één van de deelprojecten worden documenten uit een corpus van Reuter-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 precision 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. Van Gent 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 digitaal beschikbaar zijn, kunnen ze 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 OCLC-product Connexion. Voor ondersteuning van dit catalogiseerproces zijn projecten voor automatische indexering uitgevoerd.

Scorpion beoogde hulpmiddelen - ook software - te ontwikkelen voor automatische onderwerpsherkenning op basis van bekende classificaties als DDC (Thompson 1997, Godby & Reighart 1998-1, Shafer 1997, 2001). Daarbij werd een combinatie van statistische en linguïstische technieken toegepast. Nieuwe documenten vormden zoekacties in vector-representatie tegen een database met 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 & Reighart 1998-2, 2001). Ook Wordsmith werd in 2000 afgesloten. Resultaten van deze 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 LCSH-vocabulaire. De FAST-database op zijn beurt wordt dan weer gebruikt als kennisbank voor ondersteuning van automatisch classificeren. (Godby & Reighart 2001, Godby & Stuler 2001, Chan 2001)

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 (2003-1,2). Het project is afgesloten in 2003.

PHAROS
In het PHAROS-project zijn tools ontwikkeld voor een schaalbare methode van automatische filtering en selectie van informatiebronnen op het web. Daarbij werden die bronnen toegekend aan LCC-categorieën. PHAROS maakte deel uit van het Alexandria Digital Library Project van de University of California. De Alexandria Digital Library richt zich op dit moment vooral op informatieontsluiting via zijn geografische component - een "Gazetteer"-benadering (Hill 2002). Het PHAROS-project is in 1998 afgesloten. Onderdelen van Scorpion (OCLC) zijn op dit project geïnspireerd. (Dolin 1998, 1999)

WWLib
Wolverhampton Web Library: webpagina's werden automatisch ingedeeld op basis van DDC. Er lijkt niet meer aan dit systeem te worden gewerkt. (Jenkins 1998).

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 met vooral dat eerste punt moeten we wel het gevaar onder ogen zien, zoals gesignaleerd door Brazier, dat zodanig op uitsluitend nog geautomatiseerde technieken wordt ingezet, dat uiteindelijk alleen nog maar digitaal beschikbaar materiaal wordt ontsloten en niet-digitaal materiaal inhoudelijk onvindbaar wordt.

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, en of ook tevoren kan worden ingeschat welk materiaal zodanig belangrijk is dat handmatige controle en nabehandeling wel vereist is, zullen voor elke organisatie en elke collectie anders kunnen uitvallen. Het zijn echter wel zaken die moeten worden afgewogen voordat aan geautomatiseerde ontsluiting wordt begonnen.

3. Moderne methoden voor free-text retrieval

Dankzij combinaties van technieken is de kwaliteit van free-text retrieval aanzienlijk verbeterd. Welke combinaties van technieken in welke omstandigheden de grootste verbetering geven, blijkt evenwel nog een punt van discussie. Systemen met "de beste combinatie van technieken" zijn dan ook nog niet standaard te koop. Ook de vraag of de meest optimaal afgeregelde systemen in de meest optimale situatie al even goed presteren als eenvoudige zoeksystemen die zijn gebaseerd op gecontroleerde ontsluiting, kan nog niet eensluidend beantwoord worden, vooral door gebrek aan vergelijkingsexperimenten aan voldoende grote corpora. De al meer dan een decennium geleden gedane belofte dat de retrieval-technologie nog maar een kleine stap verwijderd is van de oplossing van problemen met niet-eenduidigheid van taal, is echter nog steeds niet ingelost.

3.1 Recente ontwikkelingen / state of the art

Mede onder invloed van de TREC-competitie (en CLEF 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. Zowel de precieze wijze van toepassing van veel van die technieken, als hun mate van belang zijn echter sterk taalafhankelijk en soms ook wel contextafhankelijk (Savoy 2004). Dat is waarschijnlijk een van de redenen dat in de meeste commerciële zoeksoftware maar een beperkt aantal van deze technieken wordt toegepast - meestal alleen statistiek en word-stemming en geen verdere linguïstische technieken. Van Gent ontmoet bij commerciële zoekmachine-fabrikanten vaak ook maar weinig kennis op dit terrein.

De belangrijkste van de technieken waarop hierboven gedoeld werd (zie bijvoorbeeld Ruge, 1998 en Hiemstra, 2001), zijn:

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 - bijvoorbeeld de Porter-algoritme voor Engels - en niet voor alle talen even goed mogelijk. Probleem met deze methode is dat er natuurlijk altijd uitzonderingen kunnen voorkomen, waarbij de algemene stemmingregels niet tot correcte resultaten leiden (bijvoorbeeld: communism - community - communication). Stemming kan ook berusten op gebruik van woordenlijsten en zelfs bestaan er trainbare, deels statistisch afgeleide morfologieën. Om word-stemming effectief te laten zijn, moet gezorgd worden dat noch "under-stemming", noch "over-stemming" optreedt. Moet bijvoorbeeld 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 ligt dat bij het woord "chemical" net zo? Stemming

heeft meer effect op verbetering van de recall, naarmate de hoeveelheid digitaal doorzoekbare tekst van documenten kleiner is. In lange documenten zullen verschillende woordvormen toch vaak al naast elkaar aanwezig zijn.

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 2004-2). 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 ook zogenaamde head-modifier constructies (zeil-makerij) juist wel gesplitst worden.

Desambiguering
Op basis van woordrelaties in semantische netwerken kunnen verschillende betekenissen van woorden in teksten worden onderscheiden. Overigens meent Riesthuis dat inzet van een semantisch netwerk hiervoor niet zo zinvol is, omdat polysemie een groter probleem vormt dan homonymie.

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 min of meer synoniem met die zoekwoorden.

Fuzzy zoeken
Er zijn verschillende technieken om te kunnen zoeken op woorden die sterk lijken op de woorden in de zoekvraag. Hiermee kan gecompenseerd worden voor spelfouten in zowel de documenten als de zoekvragen en voor spellingsvariaties en morfologische varianten. De meest gebruikte technieken zijn gebaseerd op het opdelen van woorden in tri- en 4-grammen, maar ook andere technieken zijn in gebruik.

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 verhogen. Door de vele uiteenlopende vormen die dergelijke phrases kunnen hebben, is het anderzijds lastig om phrases met dezelfde betekenis als zodanig te herkennen en gebruiken.

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 (2004-2) wordt ook de effectiviteit van diverse bestaande technieken vergeleken, zowel voor decompounding als voor word-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 op veel terreinen meer winst dan verlies opleveren, reageren gebruikers toch veelal zeer afwijzend als daardoor - al is het maar af en toe - onzinnige woordvarianten (stemmings-uitzonderingen, 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. Van Gent gaf dan ook aan dat retrieval op basis van goede samenvattingen vaak betere resultaten biedt dan gebruik van volledige teksten. Als toch volledige teksten van lange documenten genomen worden, dan kunnen die beter apart per pagina verwerkt worden in plaats van als geheel. Dat is ook beter dan alleen de eerste 1000 woorden te nemen. Soortgelijke resultaten rapporteert Williams (1998) op basis van experimenten met het opdelen van documenten met overlappende "windows" van 250 - 1000 woorden.

In dit kader kan ook gekeken worden welke onderdelen van documenten het belangrijkst zijn voor hun terugvindbaarheid, zowel in termen van precisie als van recall, zodat zoekacties tot die onderdelen beperkt kunnen 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. Een meer formele analyse hiervan voor goed gestructureerde documenten is al uitgevoerd door Paradis (1996).

Veel van het onderzoek aan de kwaliteit van retrieval-technieken is gebaseerd op Engelstalige documenten en corpora. Hollink (2004) heeft recent onderzocht welke van de eerder genoemde technieken voor andere Europese talen tot belangrijke verbetering van retrieval-resultaten leiden. Zij onderscheidt daarbij taalafhankelijke en taalonafhankelijke technieken. Uiteindelijk blijkt het sterk taalafhankelijk en zelfs zeer variabel 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 de verschillende talen hoeft te formuleren, wordt in het kader van CLEF en CLIR onderzoek gedaan aan twee- of meertalige zoek-technieken. Ook daarbij worden combinaties van allerlei technieken toegepast, deels dezelfde soorten als voor enkeltalige retrieval, maar 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 2004-1, Chen 2004, Hedlund 2004, Lehtokangas 2004).

Een techniek die de laatste tijd zowel op het web als in commerciële retrieval-software veel wordt toegepast is het clusteren van verkregen zoekresultaten in groepen documenten die onderlinge gelijkenis vertonen. Dat kan zijn op basis van statistiek of op basis van al toegekende trefwoorden of categorieën. Daarbij worden voor een deel vaak dezelfde technieken toegepast als die welke in hoofdstuk 2 werden beschreven bij het automatisch classificeren van documenten. Zulke clustering disambigueert vaak ook automatisch verschillende betekenissen van door zoekers gebruikte zoekwoorden. In het kader van het Scorpion-project van OCLC werd al een dergelijke techniek beschreven door Subramanian (1997). Op dit moment hoeven we hiervoor maar te kijken naar (meta-)zoekmachines als Vivisimo, Teoma of Wisenut of naar software van leveranciers als Verity.

Het merendeel van het praktisch gerichte retrieval-onderzoek - zeker ook dat in TREC-verband - vindt plaats op betrekkelijk grote tekst-corpora en veelal tekstrijke documenten. Toch worden methoden zoals hier beschreven een enkele keer ook wel toegepast op tekstarmere corpora uit bibliotheekcatalogi. Een voorbeeld daarvan is onderzoek van Grummann (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 . Verbetering van recall en precisie bleek bij die OB-collecties veel minder groot dan de eerder uit het MILOS-project gerapporteerde verbeteringen.

3.2 Afweging tussen free-text retrieval en inhoudelijke ontsluiting

Voor free-text retrieval zijn er, vooral via TREC, op basis van omvangrijke gecontroleerde corpora, zeer veel vergelijkende gegevens over de kwaliteit van allerlei retrieval-systemen. Soortgelijke vergelijkingen tussen zoekresultaten op basis van free-text retrieval en op basis van gecontroleerde handmatige ontsluiting bestaan vrijwel niet - en zeker niet grootschalig. Riesthuis krijgt uit de paar kleinschalige onderzoeken die hij kent, wel de indruk dat daarin beide methoden ongeveer even goed uit de bus kwamen. Alleen moesten voor de free-text retrieval dan wel alle voor dat specifieke materiaal in die specifieke taal meest geëigende technieken (zoals in de voorgaande paragraaf opgesomd) worden ingezet. Van der Vet verwacht dat onder die omstandigheden de schaal zelfs wel eens in het voordeel van free-text retrieval zou kunnen doorslaan. Anderzijds blijken de uitkomsten van wat in allerlei situaties de meest geëigende technieken zijn, nog allerminst eenduidig. Als we de opvatting van Van der Vet bovendien naast de eerdere observatie van Van Gent leggen, dat veel van de hier genoemde technieken nog nauwelijks in commerciële producten terug te vinden zijn, moet daar buiten universitaire laboratoriumcondities toch nog wel een groot vraagteken bij geplaatst worden.

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 ook niet zo verwonderlijk, omdat in onze huidige catalogi erg weinig digitale tekst beschik-baar is waarop die taaltechnieken zouden kunnen worden losgelaten. (Op die problematiek van de tekstarmoede van catalogusrecords wordt in het volgende hoofdstuk nog nader ingegaan). Ook het "schot in het duister" met een combinatie van wat willekeurige zoektermen, dat op het web met zijn 4 miljard webpagina's altijd nog wel enig zinnig resultaat oplevert, zal voor de veel kleinere collecties van bibliotheken inderdaad meestal bedroevend slechte resultaten opleveren, ook als de tekstarmoede van catalogusrecords genezen zou zijn. Uiteindelijk zal waarschijnlijk een combinatie van moderne retrieval-technieken, automatische karakterisering van documenten en met moderne middelen opgewaardeerde klassieke ontsluitings-systemen nodig zijn om Davis uiteindelijk tevreden te kunnen stellen.

3.3 De toekomst van free-text retrieval methoden

De steeds verdere verbetering van taaltechnologische technieken die zich manifesteert in geleidelijk steeds iets betere resultaten in TREC zal nog wel enige tijd doorgaan. Of de best haalbare recall- en precisie-cijfers uiteindelijk asymptotisch naar een eigenlijk nog niet acceptabele waarde naderen of werkelijk aanmerkelijk beter blijven worden, valt echter moeilijk te voorspellen. In dat verband moeten we ons inderdaad realiseren dat niet iedereen even tevreden is als de mensen uit de information-retrieval-gemeenschap zelf. Kunz bijvoorbeeld heeft weinig vertrouwen in de uiteindelijk te bereiken kwaliteit van full-text retrieval, refererend aan onze huidige ervaringen met web-zoekmachines. Ook memoreert hij dat ons al meer dan een decennium wordt voorgehouden dat we nog slechts een kleine stap verwijderd zijn van de uiteindelijke oplossing van de problemen met de niet-eenduidigheid van taal, die ons daarbij opbreekt. Ook in de komende tien jaar zou het best nog steeds diezelfde kleine stap kunnen blijven die ons van "de" oplossing scheidt.

4. Ontsluiting van niet-digitaal materiaal

Voor materiaal waarvan onvoldoende tekst digitaal beschikbaar is, kunnen inhoudsopgaven en andere voor de inhoud representatieve onderdelen via scannen en OCR verwerkbaar gemaakt worden. Daarnaast zal men van steeds meer boeken relevante tekstonderdelen digitaal kunnen aanschaffen. Op basis hiervan kunnen de in eerdere hoofdstukken beschreven geautomatiseerde technieken worden toegepast.

Om geautomatiseerde technieken te kunnen toepassen, of dat nu state-of-the-art text-retrieval-technieken betreft, of methoden voor de automatische ontsluiting van bibliotheekmateriaal, zal altijd een voldoende hoeveelheid informatie over of uit de documenten/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 belangrijk deel is het dat echter nog niet en zal het dat in de nabije toekomst waarschijnlijk ook niet worden. Een mogelijke oplossing voor dat materiaal is uiteraard scannen en OCR-en. Hoewel klassieke catalogiseerprocessen zelf altijd al tamelijk arbeidsintensief zijn, blijkt men in de praktijk vooral deze technische stappen in het verwerkingsproces van het papieren document als te arbeidsintensief te ervaren. Als verbeterde logistiek en workflow - en misschien ook nog wel verbeteringen in de techniek zelf - tot een versnelling van de uitvoering kan leiden en tevens het besef van het belang ervan toeneemt, moet deze extra stap in het catalogiseerproces zeker worden overwogen.

Daarbij moet in principe wel een keuze worden gemaakt welke onderdelen van de documenten op deze wijze gedigitaliseerd moeten worden. In §3.1 werd al aangegeven wat de meest representatieve en nuttige onderdelen van documenten zijn om in te zoeken. Bij veel boekmateriaal zal men zich bijvoorbeeld kunnen beperken tot inhoudsopgaven, omdat daarin meestal alle in het boek behandelde onderwerpen (weliswaar kort) in de vorm van hoofdstuk- en paragraaf-titels worden beschreven. Als een samenvatting of zelfs 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 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 uitgevers zelf, hetzij van hierin gespecialiseerde bedrijven, zoals Syndetics er al één is. 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 indexering zoals besproken in hoofdstuk 2.

5. Noodzaak van uniforme aanpak voor gelijktijdig te doorzoeken (deel-) collecties

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. In het algemeen wordt daarbij uitgegaan van collecties die allemaal al van gecontroleerde ontsluiting voorzien zijn. Collecties waarin alleen op basis van full-text retrieval wordt gezocht, worden daar niet ook bij betrokken. Andersom zal het bij toepassing van full-text retrieval als overkoepelend zoeksysteem 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.
Inhoudelijke ontsluiting van die delen van collecties die uit "fictie" bestaan, is in principe wel mogelijk. Toepassing van eenzelfde gedetailleerde ontsluiting als voor non-fictie biedt echter maar weinig toegevoegde waarde. De problematiek van de juiste afweging tussen zeer specifieke ontsluiting van hoog-specialistisch materiaal, zoals wetenschappelijke artikelen, en de veel minder specifieke concepten waarmee bijvoorbeeld leerboeken, algemene werken en verzamelwerken worden ontsloten, is wat minder eenvoudig op te lossen. Op dit moment lijkt gecontroleerde ontsluiting van alle materiaal op dat algemene niveau, in combinatie met free-text zoeken op specialistische onderwerpen de beste oplossing te bieden om zowel de algemene zoeker als de specialistische wetenschappelijke zoeker aan zijn trekken te laten komen.

5.1 Interoperabiliteit

Door de opkomst van het web is een netwerkomgeving ontstaan waarin sterke nadruk ligt op interoperabiliteit van informatiecollecties, de mogelijkheid om geografisch gescheiden collecties gelijktijdig te kunnen doorzoeken. Probleem daarbij is dat dergelijke collecties vaak op heel verschillende wijze zijn ontsloten. In het kader van interoperabiliteit is daarom nogal wat onderzoek gedaan naar de mogelijkheden van mapping of concordanties (ook wel "crosswalks" genoemd) tussen de vocabulaires van verschillende ontsluitingssystemen. Voor de hier gestelde vraag naar de noodzaak van uniforme aanpak van ontsluiting is dit een relevant onderwerp. Immers als interoperabiliteit makkelijk te bewerkstelligen is, is er ook voor de verschillende deelcollecties binnen eenzelfde organisatie geen noodzaak om in elke situatie een uniforme aanpak van ontsluiting na te streven. De problematiek van meertaligheid, die bij interoperabiliteit tussen verschillende organisaties in verschillende landen meestal een belangrijke extra complicatie vormt, speelt voor concordanties binnen een enkele organisatie meestal vrijwel niet.

Doerr (2001) geeft een uitgebreid overzicht van toe te passen technieken en van voorwaarden en beperkingen onder verschillende condities. 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, zoals 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, kunnen het erg lastig maken tot een voldoende exacte mapping tussen verschillende systemen te komen. Bij veel projecten speelt bovendien niet alleen deze problematiek een rol, maar ook die van taalverschillen en meertaligheid. Koppeling van vocabulaires kan onder meer plaats vinden met klassieke concordanties, via een centrale "tussentaal" of door linking tussen zogenaamde "subject authority files", maar ook door meer statistische methoden. Zeng verwacht dat voorlopig nog veel menselijke intellectuele inzet vereist blijft, maar dat geautomatiseerde methodes wel steeds belangrijker worden, bijvoorbeeld op basis van materiaal dat in het verleden al met verschillende systemen is ontsloten. Voorlopig zullen die methodes echter nog wel naast elkaar blijven bestaan.

Binding (2004) ziet als voornaamste hinderpaal voor de interoperabiliteit tussen ontsluitingssystemen dat er nog geen geaccepteerde standaarden bestaan voor de toegang tot "knowledge organization systems" zoals classificaties en thesauri en evenmin voor de uitwisseling van gegevens daartussen. Een web-demonstrator voor een thesaurusprotocol moet illustreren hoe dit in de praktijk zou kunnen werken. Ook Vizine-Goetz (2004) herkent een soortgelijk probleem en introduceert het concept van een via het mechanisme van web-services werkende "terminology service". Via het OAI-PMH protocol zou dat toegang moeten verschaffen tot vocabulairesystemen en concordanties.

Op dit moment zijn het vooral ontsluitingssystemen voor beperkte onderwerps-domeinen waarvoor met succes interoperabiliteit bereikt wordt. In het medische domein dient daarbij bijvoorbeeld de UMLS, het Unified Medical Language System, als de centrale spil waaromheen ontsluitingsvocabulaires met elkaar in relatie gebracht worden. Bij enkele projecten in Duitsland (CARMEN, Crosswalk STW-SWD gaat het om concordanties 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 (en waarin ook het eerder genoemde concept van een "terminology service" wordt voorgesteld). Zij besteedt ook 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. Ook binnen Renardus werd van die concordantie gebruik gemaakt. Hoewel het project al in 2001 is afgerond, heeft Riesthuis de indruk dat daar verder toch niet zo heel veel meer mee gebeurt. Bij de KB zelf zal wellicht informatie beschikbaar zijn of in het kader van TEL hiermee weer wordt verdergegaan.

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 komt vooral door

onvoldoende kwaliteit van de handmatig toegekende ontsluiting, en dat in het bijzonder voor technische en bètawetenschappen. Overigens zal men altijd moeten accepteren dat er enig verlies aan kwaliteit optreedt, doordat verschillende systemen nooit exact 1-op-1 op elkaar af te beelden zijn. De mate van dat verlies kan overigens sterk domeinafhankelijk 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 voor niet met het lokale systeem bekende gebruikers een globale toegang tot een collectie kan garanderen. Aangevuld met meer gespecialiseerde ontsluitingssystemen kan dan een veel gedetailleerder en precieze toegang geboden worden, vooral voor de gebruikers die wel met het systeem vertrouwd zijn. Juist voor de wetenschappelijk onderzoeker is die heel precieze en gecontroleerde toegang, ook via het web, echter volstrekt 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.

Gegevens over een aantal projecten op dit terrein:

CARMEN
In dit ook 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).

CERES
Project waarin een milieuthesaurus wordt opgezet via de integratie van verschillende al bestaande thesauri op de terreinen van biologie en milieu . Dit gebeurt samen met de ontwikkeling van tools voor navigatie, toekennen van metadata en zoeken,

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

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. Het 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. Moet ontsluitingssystemen van archieven, onderwijsinstellingen, bibliotheken, musea en het Resource Discovery Network vanuit diverse onderwerpsdomeinen in het Verenigd Koninkrijk aan elkaar koppelen. In elk geval moeten hiertoe 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 de gebruiker om betekenissen van begrippen te definiëren en zonodig te onderscheiden, maar ook met de systemen door die begrippen te vertalen in de termen of zonodig combinaties van termen die daarvoor in de te doorzoeken systemen worden gebruikt.

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

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

5.2 Digitaal versus niet-digitaal

In de hybride bibliotheek komen de digitale en de fysieke collectie als twee nogal ongelijksoortige deelcollecties samen. Bij de vraag of digitaal en niet-digitaal beschikbaar materiaal op dezelfde wijze ontsloten moet en kan worden, staat natuurlijk de vraag centraal of en hoe op het digitale materiaal al geautomatiseerde systemen worden toegepast. Daarbij kunnen zich een aantal verschillende situaties voordoen. Op de twee meest waarschijnlijke ga ik hieronder in meer detail in.

1. 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. Alleen als op de in hoofdstuk 4 beschreven wijze ook nog voldoende digitale informatie over dit materiaal verkregen kan worden, vervalt dit onderscheid. Anders zal op dat materiaal een liefst gecontroleerde vorm van handmatige inhoudelijke ontsluiting moeten worden toegepast. Om in dat geval 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 zoeksysteem echter methoden van relevantieordening en relevantieterugkoppeling 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. 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 §1.1.4 besproken methoden om de gebruiker - misschien zelfs ongemerkt - bij de juiste gecontroleerde zoektermen te laten uitkomen, kunnen hier dus op de gehele collectie worden toegepast. Toch dreigt hier ook nog wel 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 echter toch weer van die zoekactie uitgesloten.

In beide situaties kan voor de gebruikers van de collectie(s) in elk geval dus nog wel enige mate van uniformiteit gegarandeerd worden. Aan de ontsluitingskant vraagt dat - indien men zich voor het digitale materiaal niet geheel 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.

5.3 Wetenschappelijke publicaties versus algemeen depot-materiaal; fictie versus non-fictie.

Er zijn meer criteria op basis waarvan de collecties van bibliotheken in ongelijk-soortige 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 of tussen gecollectioneerd en voor het depot ontvangen materiaal. Van alle non-fictie, wetenschappelijk of niet, gecollectioneerd of gratis ontvangen, kan immers in principe worden bepaald, waar het inhoudelijk over gaat. Hoogstens kan er verschil zijn in de mate van specialisatie of de 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 steeds 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 American Library Association, waarin sterk de nadruk wordt gelegd op het belang van *aboutness* en *whatness*, ook voor fictie. Vooral voor historisch onderzoek acht hij de 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 meent ook Riesthuis dat het daarom weinig zinvol en zelfs wat kunstmatig is om te proberen dit materiaal helemaal op dezelfde wijze, met hetzelfde vocabulaire als non-fictie, inhoudelijk toegankelijk te maken. Als bepaalde inhoudelijke elementen van fictie toch worden ontsloten, dient het echter wel mogelijk te zijn desgewenst in alles tegelijk te zoeken. In dit verband wees hij er op dat in de catalogus van de nationale bibliotheek van Slovenië, gebruik van onderwerpsingangen inderdaad zowel non-fictie als fictie oplevert. Anderzijds moet fictie via materiaaltype of vormrubriek natuurlijk ook in zoekacties kunnen worden uitgesloten. Heiner-Freiling meldde dat ook in de Duitse nationale bibliotheek de SWD/RSWK tevens voor fictie wordt toegepast. Brazier meldde ervaringen op basis van ALA-richtlijnen bij de British Library. Onderzoek hoe nuttig het zoeken van fictie op inhoudelijke elementen door onderzoekers wordt gevonden, ontbreekt echter nog.

5.4 Gespecialiseerde versus algemene publicaties

Een derde scheidslijn waarlangs de collectie van een hybride bibliotheek in ongelijksoortige delen uiteen kan vallen is die tussen algemene en zeer specialistische publicaties. Hoewel collecties van klassieke papieren bibliotheken naast algemene werken ook tamelijk specialistische monografieën kunnen bevatten, is dat bij een hybride bibliotheek veel sterker het geval, als daarbij ook individuele tijdschriftartikelen tot de collectie 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 nog specialistischer deelonderwerpen dan met monografieën al het geval was.

Franklin (2003) geeft aan dat het voor de wetenschappelijke onderzoeker volstrekt noodzakelijk is op een dergelijk zeer specifiek niveau te kunnen zoeken. Ook Heiner-Freiling benadrukte dat voor dit doel additionele specifieke onderwerpsingangen nodig zijn. Voor bibliotheken met een collectie op vrijwel alle onderwerpsgebieden levert dit echter het probleem dat voldoende specifieke ontsluitingssystemen die ook nog voor gebruikers acceptabel zijn, niet voor alle disciplines bestaan en zeker nog niet geïntegreerd beschikbaar zijn. Ook als in het kader van interoperabiliteit oplossingen voor enerzijds integratie en anderzijds gebruikersacceptatie worden aangedragen, hebben die op dit moment nog als beperking dat ze in de opbouwfase te bewerkelijk zijn om die op voldoende specifiek niveau voor alle onderwerpsdomeinen te kunnen uitwerken.

Door Brazier wordt in dit verband nog een ander probleem genoemd: dat van 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 toch een praktische grens liggen aan het aantal zeer specifieke termen dat aan een dergelijk object kan worden toegekend. Toch is dat mijns inziens geen volstrekt nieuw probleem voor uitsluitend digitale objecten. Ook in fysieke collecties van bibliotheken kwam dat, onder meer bij verzamelbundels, al regelmatig voor. Hoogstens kan daar sprake zijn van een schaalverschil ten opzichte van sommige digitale objecten.

Anderzijds moet misschien 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 immers vaak al gespecialiseerde niet collectiegebonden bibliografische hulpmiddelen ter beschikking. De onmiddellijke online beschikbaarheid van de volledige tekst van digitaal beschikbare artikelen kan echter een sterk argument zijn om deze wel als integraal onderdeel van de lokale collectie te presenteren en te laten 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 gebruikersvocabulaire. Anderzijds zagen we ook 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.

Enige algemene conclusies

Bij elke paragraaf uit dit rapport werd al een samenvatting van de belangrijkste bevindingen gegeven. Om die te integreren volgen hier nog wat algemene conclusies.

Op het terrein van free-text/full-text retrieval is dankzij combinaties van allerlei taalkundige en statistische technieken de kwaliteit van zoekresultaten sterk verbeterd. Die verbetering is echter veel sterker op het terrein van de precisie van de resultaten, dan wat betreft de recall. Zeker als tamelijk generieke concepten onderdeel uitmaken van zoekvragen zullen zoekresultaten meestal erg onvolledig blijven. Van documenten die niet van oorsprong al digitaal zijn, zal bovendien altijd nog een voldoende hoeveelheid karakteristieke tekst gedigitaliseerd moeten worden.

Classificaties zijn geschikt voor gebruikersvriendelijke toegang van collecties via browsen. In principe kan daarmee alle materiaal, zelfs over niet-specialistische onderwerpen tamelijk volledig bij elkaar gevonden worden. Door koppeling met gebruikersvocabulaire en ontwikkelingen op het terrein van interoperabiliteit kan de toegankelijkheid nog verder worden verbeterd. Het is echter moeilijk, zo niet onmogelijk om op deze wijze specialistisch materiaal over heel specifieke onderwerpen met voldoende precisie toegankelijk te maken. Bij grote collecties is er bovendien een gerelateerd probleem van schaalgrootte. Voor toegang via browsen dienen de aantallen documenten per categorie niet te groot te zijn. Voor een collectie van bijvoorbeeld 5 miljoen documenten zullen dan tenminste 100.000 à 200.000 categorieën nodig zijn, maar dat heeft als bezwaren:
1. dat dat er erg veel zijn voor verantwoord beheer van het ontsluitingssysteem,
2. dat methoden voor automatische categorisatie van documenten voorlopig nog niet zijn toegesneden op zulke grote aantallen,
3. dat de hiervoor noodzakelijke diepte van tenminste vijf niveaus veelal te groot geacht wordt voor gebruiksvriendelijk browsen.

Thesauri en soortgelijke woordsystemen kunnen in principe wel goed zijn voor de precisie van zoekacties, mits het vocabulaire voldoende specifieke termen kent. Bij aanwezigheid van goede hiërarchische relaties kan tevens tamelijk volledig (met goede recall) op generieke concepten gezocht worden. Door koppeling met gebruikersvocabulaire en ontwikkelingen op het terrein van interoperabiliteit kan de toegankelijkheid nog verder worden verbeterd. Hier speelt echter de problematiek dat waarschijnlijk nog geen voldoende gespecialiseerd vocabulaire beschikbaar is op alle vakgebieden die in een algemene collectie aanwezig zijn. Als dergelijk vocabulaire wel beschikbaar zou zijn, speelt daarnaast de problematiek van de beheersbaarheid van een dergelijk zeer uitgebreid vocabulaire. Bovendien is automatisch toekennen van termen uit zo'n omvangrijk vocabulaire niet goed mogelijk.

De beste oplossing voor zowel specialistische als algemene zoekers lijkt daarom
- enerzijds een classificatie of een betrekkelijk globale thesaurus voor zoeken of beperken op tamelijk globale onderwerpselementen, waarbij geautomatiseerde categorisatie of trefwoordtoekenning wordt toegepast, in combinatie met
- 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 onderwerpselementen.

Een selectie van gebruikte en aangehaalde literatuur

- Jean Aitchison, Stella Dextre Clarke - The thesaurus: a historical viewpoint, with a look to the future - Cataloging & Classification Quarterly 37 (2004) nr 3/4, 5-21

- James D. Anderson, José Pérez-Carballo - 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 (2001) 231-254
Part II: Machine indexing, and the allocation of human versus machine effort - Information Processing and Management 37 (2001) 255-277

- Anders Ardö, Traugott Koch - 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/

- Donald Beagle - Visualizing Keyword Distribution Across Multidisciplinary C-Space - DLib Magazine 9 (june 2003) nr 6, http://www.dlib.org/dlib/june03/beagle/06beagle.html

- Nuria Bel, Cornelis H.A. Koster, Marta Villegas - Cross-Lingual Text Categorization - Proceedings ECDL 2003, Trondheim, August 2003, pp 126-139, http://www.cs.kun.nl/peking/ecdl03.pdf

- P. Biebricher, N. Fuhr, G. Knorz, G. Lustig, M. Schwantner - The Automatic Indexing System AIR/PHYS; from Research to Application - 11th International Conference on Research and Development in Information Retrieval, 1988, 333-342

- C. Binding, D. Tudhope - KOS at your Service: Programmatic Access to Knowledge Organisation Systems Example - Journal of Digital Information 4 (2004) nr 4, http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/

- Martin Braschler - Combination Approaches for Multilingual Text Retrieval - Information Retrieval 7 (2004) nr 1, 183-204

- Martin Braschler, Bärbel Ripplinger - How Effective is Stemming and Decompounding for German Text Retrieval? - Information Retrieval 7 (2004) nr 3, 291–316

- Vanda Broughton, Heather Lane - Classification schemes revisited: applications to web indexing and searching - Journal of Internet Cataloging 2 (2000) nr 3/4, 143-155

- Michael Buckland - Mapping entry vocabulary to unfamiliar metadata vocabularies - DLib Magazine 5 (1999) 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 - A faceted approach to subject data in the Dublin Core metadata record - Journal of Internet Cataloging 4 (2001) nr 1/2, 35-47

- L.M. Chan, M.L. Zeng - Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis - 68th IFLA Council and General Conference 2002, IFLA Journal 28 (2002) nr 5/6, 323-27, http://www.ifla.org/IV/ifla68/papers/008-122e.pdf

- A. Chen, F.C. Gey - Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding - Information Retrieval 7 (2004) nr 1, 149-182

- Stephen Paul Davis - HILCC: A Hierarchical Interface to Library of Congress Classification - Journal of Internet Cataloging 5 (2002) nr 4, 19-49

- Martin Doerr – Semantic problems of thesaurus mapping – Journal of Digital Information 1 (2001) nr 8, http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

- R. Dolin, D. Agrawal, A. El Abbadi, J. Pearlman - Using automated classification for summarizing and selecting heterogeneous information sources - DLib Magazine 4 (january 1998) nr 1, http://www.dlib.org/dlib/january98/dolin/01dolin.html

- R. Dolin, D. Agrawal, A. El Abbadi - Summarization and selection of information sources using automated classification - Eighth International World Wide Web Conference, Toronto, 1999, http://pharos.alexandria.ucsb.edu/publications/query_anal.ps

- Rosemary A. Franklin - Re-inventing subject access for the semantic Web - Online Information Review 27 (2003) nr 2, 94-101

- Joop van Gent, Onno Makor - Automatische verrijking in de praktijk - Informatie Professional 6 (2002) nr 7/8, 28-31

- Alan Gilchrist - Thesauri, taxonomieën en ontologieën: overeenkomsten en verschillen - Informatie Professional 6 (2004) nr 10, 24-27

- Carol Jean Godby, Ray Reighart - Using Machine-Readable Text as a Source of Novel Vocabulary to Update the Dewey Decimal Classification - 1998 ASIS Classification Workshop, http://orc.rsch.oclc.org:5061/papers/sigcr98.html

- Carol Jean Godby, Ray R. Reighart - The WordSmith Indexing System - Annual Review of OCLC Research, 1998, http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003487

- Jean Godby, Eric Miller, Ray Reighart - Automatically Generated Topic Maps of World Wide Web Resources - Presentation at Association for Computational Linguistics Conference, College Park, Maryland, June 1999, http://www.cs.umd.edu/users/traum/ACLDemo/Abs/godby.word

- Jean Godby, Jay Stuler - 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

- Carol Jean Godby, Ray Reighart - Terminology Identification in a Collection of Web Resources - Journal of Internet Cataloging 4 (2001) nr 1/2, 49-65

- C. Jean Godby, Ray R. Reighart - The Wordsmith indexing system - Journal of Library Administration 34 (2001) nr 3/4, 375-384

- Jean Godby, Devon Smith - 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

- Jane Greenberg - Automatic query expansion via lexical-semantic relationships - Journal of the American Society for Information Science and Technology 52 (2001) nr 5, 402-415

- Jane Greenberg - Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology - Journal of the American Society for Information Science and Technology 52 (2001) nr 6, 487-498

- Jane Greenberg - User comprehension and searching with information retrieval thesauri - Cataloging & Classification Quarterly 37 (2004) nr 3/4, 103-120

- Martin Grummann - Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? - Bibliothek 24 (2000) nr 3, 297-318

- Widad Mustafa el Hadi - Human language technology and its role in information access and management - Cataloging & Classification Quarterly 37 (2003) nr 1/2, 131-151

- T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, K. Järvelin - Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000–2002 - Information Retrieval 7 (2004) nr 1, 99-119

- Magda Heiner-Freiling - Die DDC in der Deutschen Nationalbibliografie - Dialog mit Bibliotheken 15 (2003) nr 3, 8-13, http://www.ddc-deutsch.de/literature/Heiner-Freiling_3_2003.pdf

- Djoerd Hiemstra - Using language models for information retrieval - Proefschrift Universiteit Twente, 19 januari 2001

- Linda Hill, Olha Buchel, Greg Janée, Marcia Lei Zeng - Integration of Knowledge Organization Systems into Digital Library Architectures - Proceedings of the 13th ASIST SIG/CR Workshop on "Reconceptutalizing Classification Research" (2002) 62-68, http://www.alexandria.ucsb.edu/~gjanee/archive/2002/kos-dl-paper.pdf

- Vera Hollink, Jaap Kamps, Christof Monz, Maarten de Rijke - Monolingual Document Retrieval for European Languages - Information Retrieval 7 (2004) nr 1, 33–52

- Michèle Hudon - Structuration du savoir et organisation des collections dans les répertoires du Web - Bulletin des bibliothèques de France 46 (2001) nr 1, 57-62

- C. Jenkins, M. Jackson, P. Burden, J. Wallis - Automatic classification of Web resources using Java and Dewey Decimal Classification - 7th International World Wide Web Conference, 1998, http://decweb.ethz.ch/WWW7/1846/com1846.htm

- Cornelis H.A. Koster, Marc Seutter - Taming Wild Phrases - Proceedings 25th European Conference on IR Research (2003), pp 161-176, http://www.cs.kun.nl/peking/ecir03.pdf

- Cornelis H.A. Koster, Marc Seutter, Jean G. Beney - Multi-Classification of Patent Applications with Winnow - Proceedings PSI 2003, pp 545-554, http://www.cs.kun.nl/peking/psi2003.pdf

- Wessel Kraaij - Variations on language modeling for information retrieval - Proefschrift Universiteit Twente, 18 juni 2004

- Martin Kunz - 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 - Multilingual subject access: the linking approach of MACS - Cataloging & Classification Quarterly 37 (2004) nr 3/4, 177-191

- Raija Lehtokangas, Eija Airio, Kalervo Järvelin - Transitive dictionary translation challenges direct dictionary translation in CLIR - Information Processing and Management 40 (2004) nr 6, 973-988

- Jens-Erik Mai - The future of general classification - Cataloging & Classification Quarterly 37 (2003) nr 1/2, 3-12

- Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore - A Machine Learning Approach to Building Domain-Specific Search Engines - The Sixteenth International Joint Conference on Artificial Intelligence, 1999, http://www.kamalnigam.com/papers/cora-ijcai99.pdf

- Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore - Automating the Construction of Internet Portals with Machine Learning - Information Retrieval 3 (2000) nr 2, 127-163

- Jessica L. Milstead - Metadata: the content issue - Proceedings of the National Online Meeting, New York, May 1999, 313-319

- Dennis Nicholson, Susannah Neill - Interoperability in subject terminologies: The HILT project - The New review of Information Networking 7 (2001) 147-157

- Dennis Nicholson, Gordon Dunsire, Susannah Neill - Moving towards interoperability in subject terminologies - Journal of Internet Cataloging 5 (2002) nr 4, 97-111

- Louisa Nigg - 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 - 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-ONeill_Mai-Chan.pdf

- Edward T. O'Neill, Eric Childress, Rebecca Dean, Kerre Kammerer, Diane Vizine-Goetz, Lois Mai Chan, Lynn El-Hoshy - 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

- François Paradis - Un modèle d'indexation pour les documents textuels structurés - Proefschrift Université Joseph Fourier - Grenoble 1, 7 novembre 1996

- A. Stephen Pollitt, Amanda J. Tinker, Patrick A.J. Braekevelt - 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 - Verzoek om milieu-informatie: kaarten en teksten vraaggericht begrepen; toepassing van automatische samenvattingen en trefwoordenzoekers bij het ontsluiten van milieu-informatie volgens het verdrag van Aarhus - Informatie Professional 8 (2004) nr 11 (in press)

- Jonathan Rothman - Bridging the Gap Between Materials-Focus and Audience-Focus: Providing Subject Categorization for Users of Electronic Resources - Journal of Internet Cataloging 5 (2002) nr 4, 67-80

- Gerda Ruge, Sebastian Goeser - Information Retrieval ohne Liguistik? - Nachrichten für Dokumentation 49 (1998) 361-369

- H. Saeed, A.S. Chaudry - 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 (2001) nr 1, 17-26

- Jacques Savoy - Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval - Information Retrieval 7 (2004) nr 1, 121-148

- K. Shafer - Evaluating Scorpion Results , 1997, http://purl.oclc.org/scorpion/eval-sc.html

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

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

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

- Srividhya Subramanian, Keith E. Shafer - Clustering - Annual Review of OCLC Research, 1997,
http://www.oclc.org/oclc/research/publications/review97/shafer/clustering/clustering.htm

- R. Thompson, K. Shafer, D. Vizine-Goetz - Evaluating Dewey concepts as a knowledge base for automatic subject assignment - 2nd International Conference on Digital Libraries, 1997, http://orc.rsch.oclc.org:6109/eval_dc.html

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

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

- Douglas Tudhope, Ceri Binding, Dorothee Blocks, Daniel Cunliffe - Compound Descriptors in Context: A Matching Function for Classifications and Thesauri - Proceeding of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (2002),
http://www.glam.ac.uk/soc/research/hypermedia/publications/jcdl02.pdf

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

- D. Vizine-Goetz, C. Hickey, A. Houghton, R. Thompson - Vocabulary Mapping for Terminology Services - Journal of Digital Information 4 (2004) nr 4,
http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

- H.-J. Wätjen, B. Diekman, G. Möller, K.-U. Carstensen - Bericht zur DFG-Project GERHARD: German Harvest Automated Retrieval and Directory (16-6-1998), http://www.gerhard.de/info/dokumente/dokumentation/gerhard/bericht.pdf

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

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

Een selectie van relevante websites en projecten

ARION
- http://www.dl-forum.de/Foerderung/Projekte/ARION/
- http://www.cultivate-int.org/issue4/arion/

Architecture for Accessing Scientific Collections

Bibliography on Automatic Text Categorization
- http://faure.iei.pi.cnr.it/~fabrizio/

zoek- en browse-bare collectie van 509 artikelen (okt 2004)

BINDEX
- http://www.hltcentral.org/projects/detail.php?acronym=BINDEX

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

CERES (California Environmental Resources Evaluation System)
- http://ceres.ca.gov/thesaurus/

CERES/NBII Thesaurus Partnership Project

CROSSWALK STW-SWD
- http://www.zbw-kiel.de/projekte/konkordanz-e.html

project afgesloten 2003

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

introductie van DDC in de Duitse Nationale bibliografie

DESIRE
- http://www.desire.org/

project afgesloten 2000
- 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

ET-MAP (University of Arizona)
- http://ai3.eller.arizona.edu/ent/

project afgesloten 1998

FACET (University of Glamorgan)
- http://www.comp.glam.ac.uk/~FACET/

faceted thesauri for retrieval from multimedia collections

GERHARD (Universität Oldenburg)

- http://www.gerhard.de/info/index_en.html
- http://www.gerhard.de/gerold/owa/gerhard.create_index_html?form_language=99
  1ste fase van project afgesloten 1998
- http://www.gerhard.de/info/gerhard2.html
- http://www-is.informatik.uni-oldenburg.de/forschung/forschung_1657.htm
  beschrijving vervolgproject vanaf 2001

HEREIN

- http://www.european-heritage.net/sdx/herein/thesaurus/introduction.xsp
  European Heritage Information Network

HILT (JISC)

- http://hilt.cdlr.strath.ac.uk/
  high-level thesaurus for cross-searching and browsing

INTERSPACE

- http://www.canis.uiuc.edu/interspace/
  prototype environment for semantic indexing

MACS (Multilingual access to subjects)

- http://infolab.kub.nl/prj/macs/
  project afgesloten 2001
- http://infolab.kub.nl/prj/macs/pub/MACSreport3.pdf
  eindrapport

MEANING

- http://www.lsi.upc.es/~nlp/meaning/meaning.html
  Multilingual Web-scale Language Technologies

MILOS (Universitäts- und Landesbibliothek Düsseldorf)

- http://www.ub.uni-duesseldorf.de/projekte/milos/mil_home
  project afgesloten 2001

NKOS

- http://nkos.slis.kent.edu/
  Networked Knowledge Organization Systems/Services

OASIS

- http://www-ti.informatik.uni-tuebingen.de/oasis/
- http://www.ucd.ie/ofrss/html/science/compute/1368.html
  project afgesloten 1999

OCLC: SCORPION, WORDSMITH, CORC, FAST

- http://purl.oclc.org/scorpion/
  Scorpion Archive (afgesloten 2000)
- http://www.oclc.org/research/software/scorpion/
  Scorpion project van OCLC
- http://www.oclc.org/research/projects/auto_class/
  Automatisch classificeren binnen Scorpion
- http://purl.oclc.org/oclc/wordsmith/
  Wordsmith Archive (afgesloten 2000)
- http://www.oclc.org/research/projects/fastac/
  FAST as a knowledge base for automatic classification

PEKING
- http://www.cs.kun.nl/peking/
project afgesloten 2003

PHAROS
- http://pharos.alexandria.ucsb.edu/
project afgesloten 1998

RENARDUS
- http://www.renardus.org/
integrating subject gateways across Europe

Geraadpleegde experts

In dit onderzoeksrapport zijn gegevens verwerkt uit antwoorden op mondeling of schriftelijk gestelde vragen aan de volgende personen:

- Dr. Gerhard Riesthuis (Universiteit van Amsterdam)
- Drs. Joop van Gent (Irion Technology)
- Dr. Paul van der Vet - met inbreng van Prof. Dr. Franciska de Jong en Prof. Dr. Theo Huibers (allen: Universiteit Twente)
- Martin Kunz (Die Deutsche Bibliothek)
- Magda Heiner-Freiling (Die Deutsche Bibliothek)
- Caroline Brazier (British Library)

Gebruikte afkortingen

AAT	Art & Architecture Thesaurus
AI	artificial intelligence
ALA	American Library Association
BLISS	(geen afkorting maar de eigennaam van de ontwerper van de gelijknamige facet-classificatie)
BT	Broader Term (in thesaurus)
CLEF	Cross-Language Evaluation Forum
CLIR	Cross-Language Information Retrieval
CORC	Cooperative Online Resource Catalog
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
GOO	Gemeenschappelijke onderwerpsontsluiting
IMIX	Interactieve Multimodale Informatie Extractie
ISO	International Organization for Standardization
LCC	Library of Congress Classification
LCSH	Library of Congres Subject Headings
MeSH	Medical Subject Headings
MILOS	Maschinelle Indexierung zur erweiterten Literaturerschließung in Online-Systemen
NBC	Nederlandse Basis Classificatie
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
RAMEAU	Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié
RT	Related Term (in thesaurus)
SWD/RSWK	Schlagwortnormdatei/Regeln für den Schlagwortkatalog
STW	Standard-Thesaurus Wirtschaft
TREC	Text REtrieval Conference
UDC	Universele Decimale Classificatie
XML	Extensible Mark-up Language