Eric Sieverts, Onno Mastenbroek en Natalia Grygierczyk

Een uniform retrieval-systeem voor de Universiteit Utrecht

De Universiteitsbibliotheek Utrecht zet een algemeen toepasbaar information retrieval-systeem op, waarmee alle door de UBU aangeboden informatie op efficiënte wijze ontsloten zal worden. Het systeem bevindt zich nog in de introductie- en opbouwfase, maar enkele conclusies zijn al te trekken.

STEEDS MEER INFORMATIE komt digitaal beschikbaar. Grote bibliotheken bieden hun gebruikers daardoor een steeds onoverzichtelijker wordend scala aan informatiebronnen: de bibliotheekcatalogus, bibliografische databases, full-text tijdschriften, collecties www-links, lokale documentatie enzovoort. Meestal is een deel ervan in eigen beheer, deels worden bronnen op cd-rom of tape aangeschaft, deels zijn ze via netwerken elders toegankelijk. Vrijwel elke bron heeft daardoor een eigen interface met eigen zoekmogelijkheden. Weliswaar is retrieval intussen een 'commodity' geworden – het 'komt gewoon uit de kraan' – maar daarmee ontstaat nog niet automatisch een systeem waarmee gebruikers eenvoudig in al die informatie kunnen zoeken of waarmee ze zelfs maar de weg kunnen vinden in al die bronnen. Bij de Universiteitsbibliotheek Utrecht (UBU) wordt daarom een algemeen toepasbaar Information Retrieval-systeem opgezet, waarmee alle door de UBU aangeboden informatie op efficiënte wijze ontsloten zal worden. Dit is gestart als onderdeel van het project 'Elektronische Bibliotheek Utrecht' en uitgegroeid tot speerpunt binnen de innovatieve projecten van de UBU (zie kader).

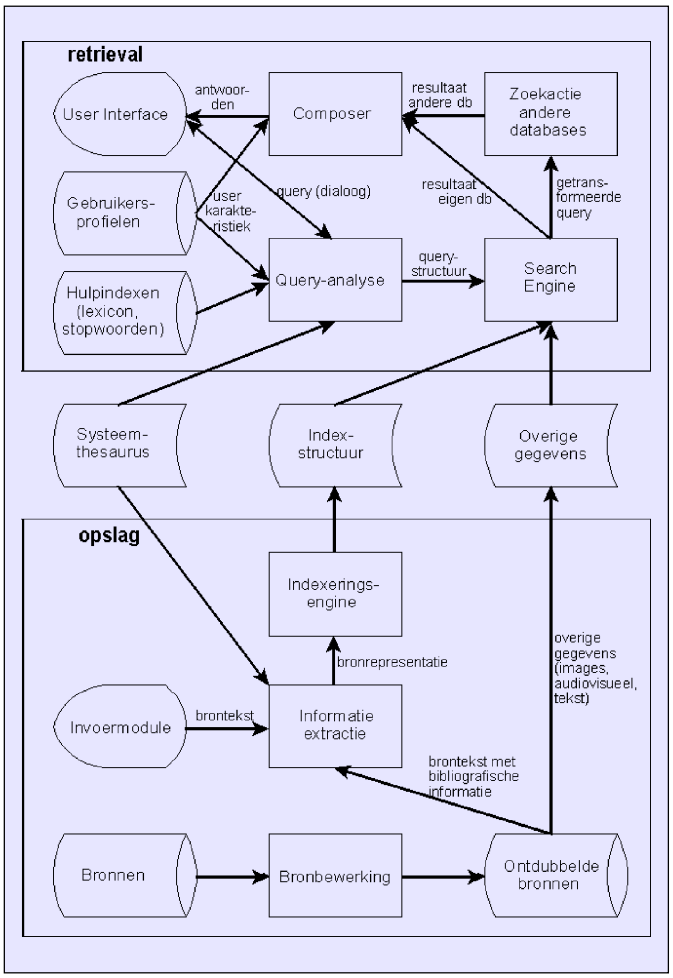
Een functioneel model

Het opzetten van een informatiesysteem vergt een projectmatige aanpak, waarin een aantal te doorlopen stappen wordt onderscheiden. In een klassieke opzet zijn dat onder meer: definitiestudie, functionele en gegevensanalyse, softwarekeuze, implementatie en evaluatie. Vooral in een situatie waarbij er nog geen commercieel verkrijgbare software is, die al volledig voldoet aan alle eisen die uit de definitiestudie voortkomen, moet veel aandacht worden besteed aan een functioneel model voor het te bouwen systeem. Voor de Utrechtse situatie was dat zeker het geval.

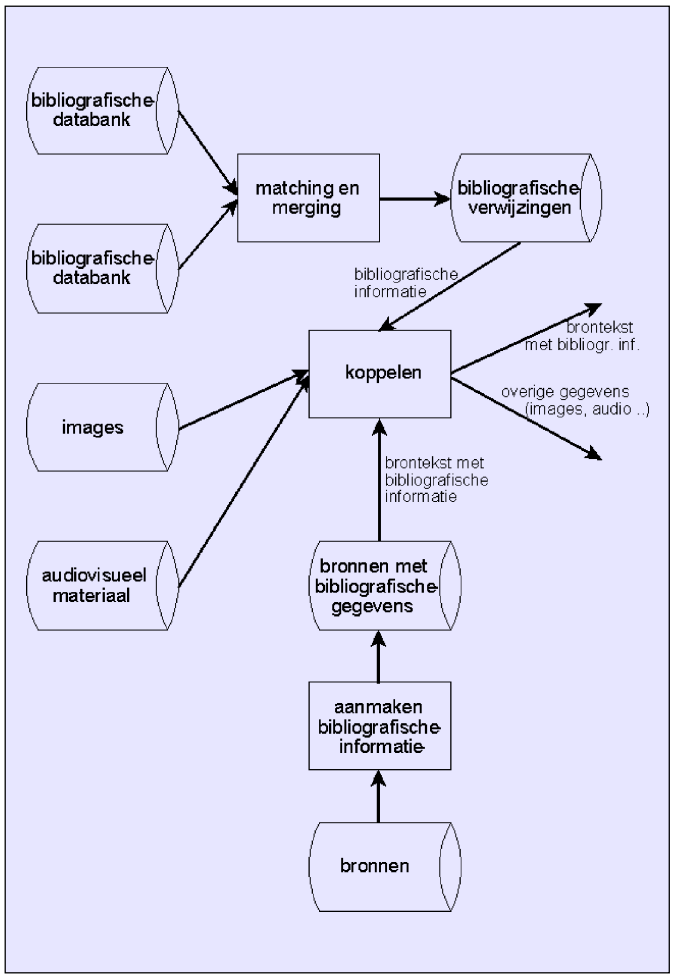
Belangrijkste uitgangspunten waren:

- informatie uit vele bronnen en in vele vormen moet in één systeem met een uniform gebruikersinterface aangeboden worden;
- individuele bronnen en groepen van bronnen moeten zo nodig ook afzonderlijk doorzocht kunnen worden;
- om gebruikers niet met ingewikkeld geachte zoekopdrachten te confronteren, moet zoveel mogelijk gebruikgemaakt worden van natuurlijke-taaltechnieken;
- daarnaast moet wel gebruikgemaakt kunnen worden van

Verloop van het project	Tijdperiode	Kader
• voorstudies algehele nieuwe dienstverlening	**febr. - nov. 1995**	Project Elektronische Bibliotheek Utrecht
• onderzoek naar gebruikersbehoeften		
• probleemanalyse huidige bibliotheekdiensten aan Universiteit Utrecht		
• onderzoek naar nieuwe informatiediensten en strategische keuzes		
• haalbaarheidsstudie IR-systeem: opstellen programma van eisen (globaal ontwerp)	**jan. - aug. 1996**	
• keuze van IR-systeem: inventarisatie bestaande producten, vergelijking met programma van eisen, opstellen contract	**dec. 1996 - dec. 1997**	
• implementatie gekozen IR-systeem op deelcollectie UU	**jan. -dec. 1998**	Innovatieve Projecten
• scholing IT-personeel UBU		
• gebruikerstests en aanpassing		
• plan voor vervolgtraject		
• stapsgewijze implementatie op volledige collectie UU en uitbreiding functionaliteiten	**jan. 1999 - heden**	



Figuur 1. Functioneel model



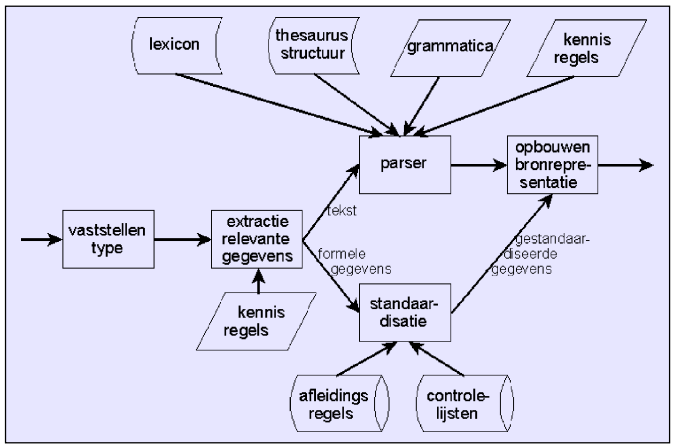
Figuur 2. Bronbewerking

eventueel aanwezige gestructureerd opgeslagen informatie (velden);
• er moet gebruikgemaakt kunnen worden van al in bepaalde informatiebronnen aanwezige gecontroleerde ontsluiting;
• het systeem moet geavanceerde methoden van relevance ranking toepassen, waarbij gegevens uit verschillende bronnen, ook al verschillen die sterk in rijkdom aan tekst, op zinnige manier gemengd worden;
• het systeem moet gebruik van multimediale informatiebronnen ondersteunen.

Hiervan uitgaande is, in samenwerking met een externe adviseur, een globaal schema opgezet (Van Berkel & Mastenbroek, 1996). Figuur 1 geeft een overzicht van het systeem als geheel. Daarin zien we onderin een indexeerdeel, bovenin een gedeelte dat het zoeken en de interactie met de gebruiker afhandelt en als verbindingslaag daartussen de indexen en bijbehorende hulpbestanden. Exacte invulling van alle componenten in dit model was om een aantal redenen nog niet mogelijk of zelfs niet gewenst:
• omdat nog onbekend is welke soorten nieuwe bronnen op termijn moeten worden opgenomen, dient de opzet zo generiek en open mogelijk te zijn;
• modules voor de verschillende gewenste functionaliteiten dienen zo min mogelijk zelf te worden ontwikkeld, terwijl nog onvoldoende bekend was wat voor modules met welke precieze werking al commercieel beschikbaar waren of in een experimentele omgeving al voldoende waren uitontwikkeld om te worden opgenomen;
• introductie van nieuwe technieken en inzichten moet zo min mogelijk ingrijpende gevolgen hebben voor het systeem als geheel;
• voor een aantal elementen werd voorzien dat tijdrovend aanvullend onderzoek, bijvoorbeeld naar gebruikersgedrag, nodig zou zijn.

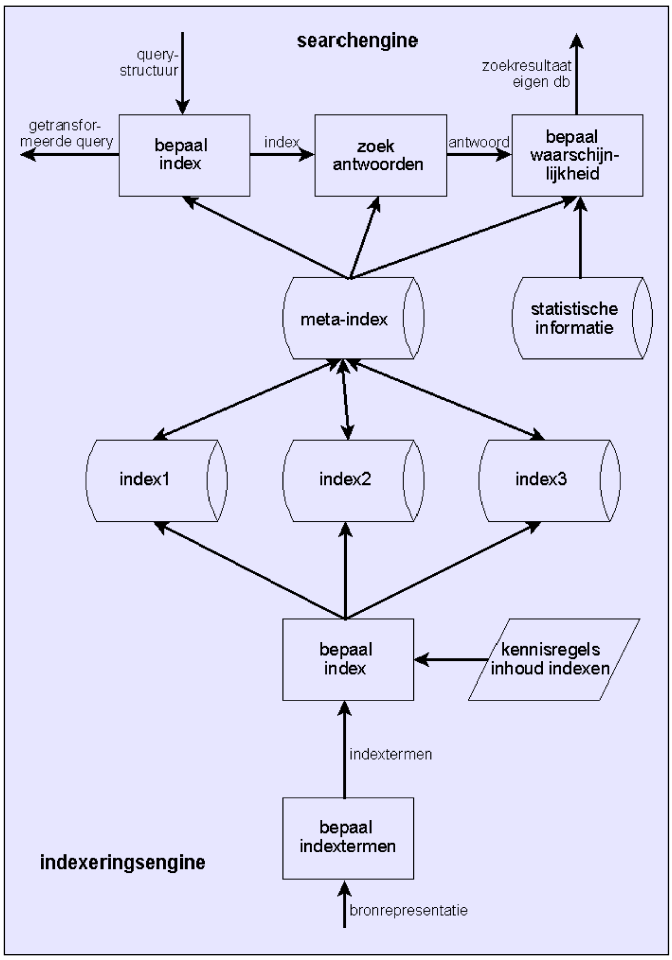
Enkele details van het functionele model

Toch heeft voor een aantal onderdelen wel al nadere invulling plaatsgevonden. Zo toont figuur 2 een nadere uitwerking van het onderdeel 'bronbewerking'. Voor elke bron worden indien nodig bibliografische gegevens aangemaakt. Dit proces dient zoveel mogelijk geautomatiseerd te zijn. Dit zou kunnen gebeuren op basis van vormkenmerken in de tekst, bijvoorbeeld dat in bepaald materiaal bovenaan altijd de titel staat, daaronder de auteur, dan een abstract enzovoort. Van de documenten in een bestand moeten verder documentdefinities gemaakt worden waarin het formaat beschreven is; dit is onder andere van belang bij het presenteren van de informatie. Overigens zal natuurlijk zoveel mogelijk gebruikgemaakt worden van al aanwezige bronbeschrijvingen. Per bron moeten voorts koppelingen worden aangebracht met bibliografische gegevens in andere bij de UBU beschikbare databanken en eventuele andersoortige informatie zoals images en AV-materiaal.



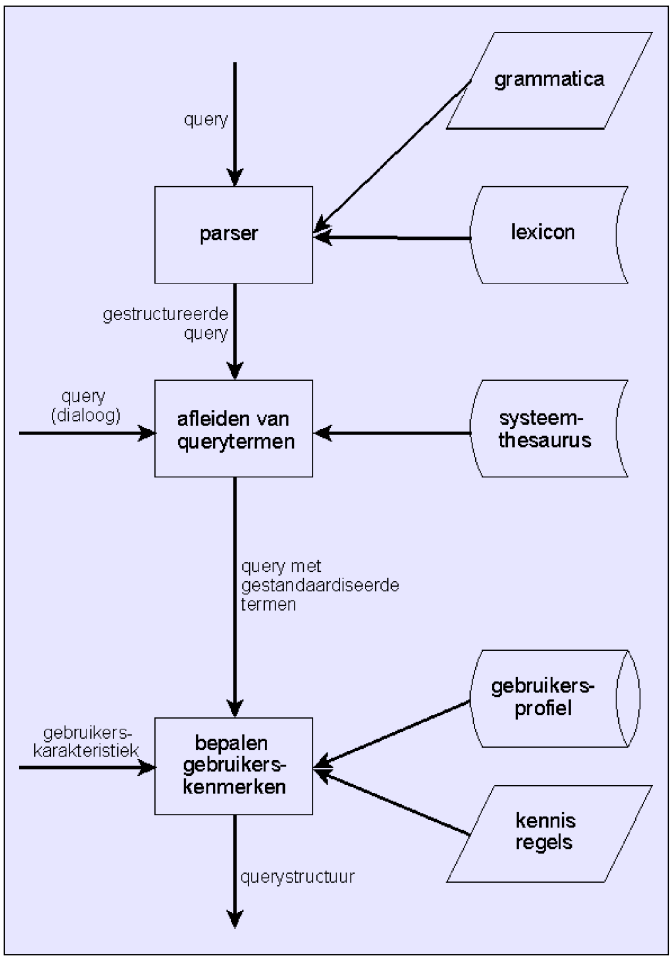
Figuur 3. Informatie-extractie

In figuur 3 wordt aangegeven hoe informatie-extractie uit aangeleverd bronmateriaal ten behoeve van indexering zou kunnen plaatsvinden. Nadat ze bij bronverwerking zijn verwerkt, analyseert het systeem documenten en kent er trefwoorden aan toe. Op grond van de vanuit bronbewerking doorgezonden informatie (brontekst met bibliografische informatie) wordt bepaald om wat voor soort bron het gaat (full-text, bibliografisch, hypertekst et cetera). Vervolgens wordt aan de hand van dtd's (documenttypedefinities) en kennisregels (hoe herken je auteur, discipline, jaar van uitgave, maar vooral ook niet-bibliografische gegevens) verdere relevante informatie aan de bron onttrokken.
De tekst van een full-text bron, van een abstract en de trefwoorden worden onderworpen aan een parser die met behulp van een lexicon, een systeemthesaurus, een grammatica en kennisregels een semantische representatie van de bron geeft. Hierbij spelen morfologische, syntactische, semantische en lexicale analyses een rol. Automatische thesaurering van documenten, met gebruik van uit bestaande thesauri afkomstige termen, kan een onderdeel van dit proces zijn. Uiteindelijk zal dit ook tot multilinguale verwerking moeten leiden.
Voor zaken als auteur, jaar van uitgave en dergelijke wordt tevens presentatie binnen de bronpresentatie gestandaardiseerd aan de hand van controlelijsten (Czechow ——> Tjechov) en afleidingsregels (X ——> 10). Eindresultaat van dit proces dient te zijn dat alle documenten standaard gecodeerd zijn en beschrijvende trefwoorden aanwezig zijn.
Nadruk wordt gelegd op gebruik van metadata voor karakterisering van individuele documenten of hele broncollecties. Enerzijds omdat ook gewone webpagina's in het systeem worden opgenomen, waarvan formele en inhoudelijke kenmerken op gestructureerde wijze moeten worden vastgelegd. Anderzijds omdat het voor het samenbrengen van materiaal uit verschillende bronnen noodzakelijk is formele kenmerken van die bronnen aan individuele records te koppelen. Dergelijke gegevens kunnen nodig zijn voor het onderscheiden van deelcollecties, voor het op juiste wijze naar relevantie mengen van materiaal uit verschillende bronnen, voor het oproepen van voor presentatie van gegevens benodigde hulpprogramma's, voor het correct koppelen van multimediaal materiaal, enzovoort.



Figuur 4. Search-engine en indexerings-engine

Figuur 4 toont een voorlopige invulling voor de onderdelen indexering en search engine, alsmede de tussenliggende indexen. Een gestandaardiseerde ingevulde querystructuur wordt doorgegeven aan de 'search engine'. Deze bepaalt waar de zoekvraag het best heen gestuurd kan worden aan de hand van:
• in de representatie van de vraag aanwezige gegevens, (statistische) informatie over de bronnen (bijvoorbeeld frequenties van het voorkomen van termen);
• informatie uit de meta-index;
• een voorbeeld: het trefwoord 'lever' in combinatie met vakgebied 'medisch' stuurt de vraag door naar medische bestanden, in combinatie met 'keuken' wordt voor heel andere databases gekozen.
Na bepaling van te gebruiken databases komt het zoeken zelf. In de gekozen index(en) worden de antwoorden opgezocht en, aan de hand van daartoe in het systeem voorhanden zijnde regels, wordt de mate van overeenkomst tussen vraag en antwoord bepaald (ranking).
Voor het bouwen van de indexen worden vanuit de bronrepresentatie de indextermen bepaald (bijvoorbeeld losse woorden, maar ook zinsdelen en uitdrukkingen). Met behulp van kennisregels wordt vastgesteld in welke van de aanwezige indexbestanden verwijzingen naar de desbetreffende bron worden opgenomen. De benodigde kennisregels kunnen betrekking hebben op de *aard* van de bron (medisch, juridisch et cetera), op het *type* bron (full-text, bibliografisch, beeldmateriaal et cetera) en het *niveau* van



Figuur 5. Query-analyse

de bron (wetenschappelijk, studenten, populair et cetera). De boven de indexen aanwezige meta-index bevat ook dit soort informatie, bijvoorbeeld index 1 is medisch, index 2 is voor studenten geschikt en dergelijke. Bovendien bevat de meta-index gegevens met betrekking tot elders toegankelijke indexen.
Uitgangspunt bij het indexeer- en zoekproces is dat, voor optimaal gebruik van probabilistische of andere partial-match zoektechnieken en daaruit voortvloeiende relevantieordening van zoekresultaten, alle te doorzoeken gegevens aan eenzelfde indexeringsmethode onderworpen moeten worden. Anderzijds werd voorzien dat eigen indexering niet in alle gevallen mogelijk zou zijn, omdat een via een netwerk toegankelijke externe informatiebron alleen via een zoekinterface benaderd kan worden en de brongegevens zelf niet rechtstreeks bereikt kunnen worden. Een protocol als Z39.50 biedt dan weliswaar gestandaardiseerde toegang, maar het zal waarschijnlijk heel moeilijk zijn om zo verkregen gegevens op relevantie geordend samen te voegen met resultaten uit het eigen systeem. Niettemin vormt een Z39.50-interface een noodzakelijke functionaliteit.

Figuur 5 toont hoe door gebruikers gestelde vragen worden bewerkt en aan de zoekmachine doorgegeven. De gebruiker kan de vraag op verschillende manieren invoeren, zowel met trefwoorden als in de vorm van een 'natuurlijke zin'. Daarnaast moet het mogelijk zijn eerder gevonden resultaten (primair of secundair; geheel of gedeeltelijk) als input te geven ('query by example') of een al eerder gedefinieerd zoekprofiel. De vraag wordt opgesplitst in losse betekenisvolle elementen. Dit kunnen ook samengestelde begrippen zijn, bijvoorbeeld *aandoeningen van de lever*. Hierbij kunnen verschillende NLP-technieken, zoals morfologische, lexicale en syntactische analyse, worden gebruikt, terwijl tevens interactie met de gebruiker dient plaats te vinden. Deze query-elementen worden gekoppeld aan thesaurustermen, hetgeen kan leiden tot verrijking van de query, onder andere door het toevoegen van synoniemen (fiets – rijwiel), specifiekere termen (fiets – vouwfiets) en dergelijke.
De query is nu omgevormd tot een reeks gestandaardiseerde termen. Hieraan wordt informatie over de gebruiker toegevoegd (bijvoorbeeld *vakgebied*, zoals geneeskunde of letteren, *ervaring*, zoals onderscheid student, onderzoeker, informatiespecialist, *eerder gestelde vragen* et cetera). Deze informatie zou al per gebruiker in het systeem aanwezig kunnen zijn. Ook zou de gebruiker pas bij invoer van zijn vraag eventuele informatie over zichzelf kunnen meegeven.

Softwarekeuze

Uit het functioneel model volgen uiteraard eisen waaraan te kiezen software moet voldoen. Gezien het nog tamelijk globale karakter van dit model, waren dat vaak nog weinig gedetailleerde eisen. Bovendien zijn ook andere overwegingen van belang voor een verantwoorde keuze. In de eerste plaats valt te denken aan zaken die met de leverancier samenhangen, zoals de opgebouwde klantenkring en te verwachten ondersteuning, mede gebaseerd op ervaringen van die klanten. Ook is gekeken naar de financiële positie van de leverancier en de te verwachten continuïteit van de dienstverlening. Een essentieel punt was natuurlijk de te verwachten verdere ontwikkeling van het product en de mogelijkheid die te beïnvloeden. Verder zijn er zaken die met de technische opzet van het systeem samenhangen, zoals een programmeerinterface (API), client-server-architectuur, mate van openheid voor integratie van elders ontwikkelde modules, overdraagbaarheid, platformonafhankelijkheid, beschikbare ontwikkeltools, mogelijkheden zelf gebruikersinterfaces te ontwikkelen, performance bij veel gelijktijdige gebruikers en dergelijke. Uiteindelijk is een lijst van circa honderd beoordelingscriteria opgesteld (List, 1997). Op grond daarvan is een zogenaamde 'tool-evaluatie' uitgevoerd.

Na globale voorselectie waren aanvankelijk drie potentieel interessante producten overgebleven, te weten *Search97* (van Verity), *RetrievalWare* (van ExCalibur) en het Engelse *Muscat*. In een later stadium werd *AltaVista* nog toegevoegd en is een verkorte analyse van *Orion ScienceServer*-software uitgevoerd. In samenwerking met externe adviseurs zijn de met de beoordelingscriteria samenhangende vragen voor deze potentiële 'tools' zo goed mogelijk beantwoord. Daarbij is gebruikgemaakt van documentatie van de leveranciers, van indrukken opgedaan tijdens hiervoor georganiseerde productdemonstraties, van door de leveranciers op specifieke vragen gegeven antwoorden en van door de leveranciers zelf ingevulde vragenlijsten.

Waar gewenste functionaliteit door een product (nog) niet werd geboden, is bovendien onderzocht of daarvoor redelijke alternatieven aanwezig waren, of dat de gewenste functionaliteit als externe module gekoppeld of geïntegreerd zou kunnen worden. Ook is gekeken of voor de betreffende software TREC-testresultaten beschikbaar waren (TREC, 1998). Deze bleken overigens weinig uiteen te lopen. Alle resultaten zijn verzameld in een vertrouwelijk intern rapport dat als uitgangspunt diende voor de uiteindelijke keuze.

De ervaringen met beantwoording van evaluatievragen door de leveranciers waren wisselend. Uiteraard zal een leverancier bij de interpretatie van functionaliteitscriteria al snel aangeven dat een bepaalde functionaliteit voorhanden is, ook al verzorgt de software dat op een door de gebruiker als omslachtig ervaren, niet gewenste of beslist niet bedoelde manier. Met voldoende kritische zin waren dergelijke interpretatieverschillen vaak te achterhalen. Bijvoorbeeld waar een leverancier er stilzwijgend van uitging dat eerst handmatig een topics-database met gerelateerde begrippen zou worden opgebouwd, teneinde zoekvragen automatisch met aanvullende zoektermen te kunnen uitbreiden. Bij criteria met betrekking tot het automatisch toekennen van thesaurustermen aan documenten of het genereren van zogenaamde noun-phrase indexen hadden we dezelfde ervaring. In sommige gevallen bleek interpretatieverschil in details pas achteraf bij praktische implementatie. In de paragraaf 'knelpunten' is het voorbeeld van fuzzy-zoeken bij Muscat verder uitgewerkt.

Uit de vergelijking kwam als conclusie naar voren dat:
• Muscat beter aansloot op de wensen van de UBU dan de andere pakketten;
• de Muscat-software een open structuur heeft, waardoor het mogelijk is nieuwe ontwikkelingen vorm te geven;
• door Muscat aangegeven toekomstige ontwikkelingen de functionaliteit nog dichter bij de wensen van de UBU brengen, terwijl bij andere pakketten nog niet duidelijk was wat in de nabije toekomst aandacht zou krijgen;
• de communicatie met Muscat waarschijnlijk makkelijker en directer zou verlopen dan met de andere bedrijven; de contractonderhandelingen bleken daar inderdaad een voorbeeld van;
• er bij Muscat, door afkomst uit de academische wereld en daarmee nog bestaande nauwe banden, meer (inhoudelijk) begrip zou zijn bij besprekingen over functionaliteit, leidend tot betere mogelijkheden tot inhoudelijke samenwerking.
Voorts is Muscat in 1996/97 overgenomen door MAID en maakt nu dus deel uit van de Dialog Corporation, waardoor de aanvankelijke vrees voor onvoldoende stabiliteit van een betrekkelijk klein bedrijf werd weggenomen. Het feit dat volop aan het product ontwikkeld wordt, geeft garanties voor continuïteit in de toekomst.

Implementatie

Elsevier Allereerst is de software gebruikt om een zoeksysteem voor de wetenschappelijke tijdschriften van Elsevier op te zetten (februari 1999). Hoewel Muscat filters heeft voor het full-text indexeren van pdf-bestanden en Elsevier naast pdf-bestanden ook sgml (dus ASCII-versies) van de volledige teksten van de artikelen aanlevert, is besloten voorlopig alleen de bibliografische gegevens en abstracts doorzoekbaar te maken. Reden hiervoor is dat het zoeken in full-text bestanden geheel eigen problemen kent en we die niet tegelijk met een nieuwe opzet van het zoeksysteem wilden introduceren.

Sabine Na implementatie op de Elsevier-tijdschriften, is Muscat toegepast op een bibliografisch bestand: Sabine, de bibliografie voor de provincie Utrecht. Een punt van aandacht hierbij was de noodzaak een invoermodule voor de Muscat-software te ontwikkelen, waarmee records toegevoegd, verwijderd, of gewijzigd konden worden. Dit is geen standaardonderdeel, maar is in een bibliotheekomgeving wel noodzakelijk en maakt natuurlijk deel uit van het onderdeel bronbewerking in het besproken functionele model.

Standaardmodel Als voorbereiding op de toevoeging van andere bronnen is een algemene veldenstructuur opgezet die ook voor die bronnen toereikend moet zijn. In de zomer van 1999 hebben de ervaringen met Muscat geresulteerd in een eerste prototype van het standaard IR-systeem. Full-text en bibliografische bronnen zijn gemengd in het systeem ondergebracht, de bijbehorende indexstructuur is opgezet en er is een eerste algemeen zoekinterface ontwikkeld. In samenwerking met de faculteit Psychologie van de Universiteit van Maastricht wordt onderzoek verricht naar de reactie van de gebruikers (studenten en wetenschappelijk medewerkers) op de zoekmachine. Met de Muscat-software wordt nu een zoeksysteem voor de diverse overige bronnen gerealiseerd (zie kader).

	Successievelijk aan het systeem toe te voegen informatiebronnen
Elsevier full-text tijdschriften	410 full-text wetenschappelijke tijdschriften op eigen server
Sabine	ruim 20.000 bibliografische records over de provincie Utrecht op eigen server
Core Collection in Physics	ca. 90 full-text natuurkundige tijdschriften op servers van een aantal uitgevers
Internet-bronnen	verzameling links naar externe www-pagina's
Internet Law Library	verzameling juridische www-links
Swets full-text	ca. 250 full-text tijdschriften op server van Swets
Ebsco full-text	ca. 1000 full-text tijdschriften op server van Ebsco
UU-Proefschriften	via WWW full-text aangeboden proefschriften
Bijzondere collecties	gedigitaliseerde eigen primaire bronnen (o.a. oude drukken)
UBU-catalogus	ca. 1,8 miljoen bibliografische records*
Diverse bibliografische databases	nu nog via ERL of op afzonderlijke cd-rom aangeboden

* Voor administratieve bibliotheektaken als catalogiseren, uitlenen, bestellen e.d. blijft het Aleph-bibliotheeksysteem gebruikt worden.

Functionaliteit Een volgende stap zal zijn meer functionaliteit aan de zoekmachine toe te voegen. Zo is het mogelijk bronnen in categorieën in te delen en die doorzoekbaar te maken. Hierbij moet de gebruiker een zoekactie tot één of meer categorieën kunnen inperken (voor- of achteraf), en moet het mogelijk zijn met hoofd- en deelcategorieën te werken. Die structuur moet binnen Muscat zo worden gedefinieerd, dat wijzigingen in de structuur gemakkelijk en dynamisch kunnen worden aangebracht. Voor automatische attendering van gebruikers op nieuw toegevoegde documenten en bronnen die voor hen interessant kunnen zijn, zal gebruikgemaakt worden van 'agent technology'. In het najaar van 1999 zullen de eerste experimenten hiermee plaatsvinden.
De gebruiker zal zijn zoekactie niet willen beperken tot bij de UBU aanwezige (of geïndexeerde) databestanden. Zoekacties moeten ook geëxporteerd kunnen worden naar externe bestanden. Hiertoe zou het Z39.50-protocol gebruikt kunnen worden, maar ook andere oplossingen (xml) zijn denkbaar. Dit is de komende maanden een punt van aandacht.

Knelpunten

Natuurlijk zijn we bij de praktische implementatie nog tal van knelpunten tegengekomen. Enkele daarvan willen we kort aanstippen.

De techniek Vanuit het functioneel model is gekozen voor een oplossing waarbij alle beschikbare digitale bronnen (ook die welke niet op het lokale systeem beheerd worden) door de Muscat-software worden geïndexeerd. Waar het openbaar toegankelijke webpagina's betreft is dat een logische keuze. Voor full-text wetenschappelijke tijdschriften die op grond van licenties via een webinterface met gecontroleerde toegang worden aangeboden, is het dat in principe ook. De benodigde on line toegang tot de brongegevens blijkt echter soms een knelpunt te vormen. Zo bleek een indexeerrobot na enige tijd, waarschijnlijk uit beveiligingsoverwegingen, verdere toegang geweigerd te worden tot een website waartoe we op grond van een licentieovereenkomst gewoon toegang hadden.

Licenties Licentieregelingen blijken niet expliciet te voorzien in het toestaan van indexeeractiviteiten op afstand. In de praktijk moeten daarom min of meer formele, aanvullende regelingen worden getroffen, op grond waarvan volledige gegevens, bijvoorbeeld met ftp, kunnen worden opgehaald. De vervolgens gegenereerde indexen moeten dan wel naar de oorspronkelijke locaties (url's) op de website van de externe aanbieder blijven verwijzen. Of de nu ook lokaal aanwezige volledige gegevens vervolgens kunnen worden weggegooid (afgezien van wat daarover formeel is afgesproken), moet nog worden onderzocht. In principe hoeven ze na het indexeerproces niet bewaard te worden. Maar als voorzien wordt dat een index nog herbouwd moet worden of dat proeven met alternatieve indexeerstrategieën nodig zijn, verdient het aanbeveling de gegevens tenminste tijdelijk te bewaren, om te voorkomen dat opnieuw toestemming gevraagd moet worden de hele collectie nogmaals te ftp'en.
In de praktijk blijkt dat licenties voor gebruik van elektronische data vaak nog slecht geregeld zijn. Het kost veel moeite tijdschriftagenten en uitgevers uit te leggen wat de wensen van de gebruiker zijn en hoe de bibliotheek daarmee omgaat. Dat betekent tijdverlies bij het aanbieden van elektronische bronnen, om nog maar te zwijgen van ergernissen over en weer. Is het maken van goede licentieafspraken voor full-text elektronische tijdschriften moeizaam, voor op cd-rom geleverde databases lijkt het vrijwel onmogelijk. De UBU biedt ruim honderdtien van dergelijke databases aan binnen de campus van de Universiteit (Hackenitz, 1998) en zou de gegevens uit al deze bestanden graag overbrengen naar het Muscat-systeem. Het blijkt echter dat dit niet wordt toegestaan door de bestandsproducenten, die het loskoppelen van data van de bijgeleverde zoeksoftware niet toestaan. Naar een oplossing voor dit probleem wordt naarstig gezocht.

Functionaliteit Al eerder kwam ter sprake dat vermelding door een leverancier dat hun software bepaalde functionaliteit ondersteunt, niet altijd hoeft te betekenen dat dat gebeurt op de manier die de gebruiker voor ogen staat. Een voorbeeld hiervan zijn problemen met fuzzy zoeken, die pas bij gedetailleerde praktijktests aan het licht traden. In de standaardimplementatie bleek Muscat alleen fuzzy te zoeken als de gevraagde term niet in de index voorkwam. Dan werd uit de index die term gekozen – slechts één – die het meest op de gevraagde term leek. In de praktijk wil je echter op alle varianten van een term kunnen zoeken, zeker ook als wel een in de index aanwezige spelling is gebruikt. Gelukkig bleek dit door de leverancier eenvoudig aan te passen. Nu werd echter automatisch gezocht op alle termen die, op grond van trigramanalyse, enige gelijkenis met de zoekterm hadden. Hoeveel gelijkenis minimaal vereist was, kon niet worden gespecificeerd; intern kon alleen worden ingesteld hoevéél termen moesten worden meegenomen. Keuzelijstjes waarin de gebruiker relevante varianten kan aanklikken en onzinnige kan overslaan zouden een veel betere oplossing zijn. In een intussen geïnstalleerde versie van de Muscat-software blijkt het geschetste probleem intussen op andere wijze bevredigend te zijn opgelost.
Een fundamenteler probleem was dat alle in de index voorkomende fuzzy-varianten van de vraagterm als individuele termen in de zoekactie bleken mee te tellen. Bij Muscats probabilistische zoekmethode, zou dan alleen een document dat alle termvarianten tegelijk bevat een honderd procent relevantiescore krijgen en dat komt natuurlijk niet voor. In een zoekactie met drie afzonderlijke zoektermen, krijgt een document met vier fuzzy-varianten van een van die termen daardoor ook een hogere relevantiescore dan een document dat alle drie termen bevat, maar slechts één variant van elk. Bovendien wordt het probabilistische begingewicht van een zoekterm bepaald door zijn zeldzaamheid. Aanwezigheid van een veel in de database voorkomende term is minder belangrijk voor de relevantie van een document dan de aanwezigheid van een zeldzame – dus specifieke – term. Doordat elke fuzzy-variant van een term afzonderlijk meetelt bij de relevantiebepaling, krijgt een zeldzame, fout-gespelde variant zo een hoger gewicht dan de correct gespelde voorkomens. Al met al een hele reeks ongewenste effecten.

Van de leverancier van de software wordt in dergelijke gevallen dus verwacht dat achteraf – vaak dieper ingrijpende – aanpassingen in de software worden aangebracht, om gewenste functionaliteiten in detail en in combinatie met elkaar, op gewenste wijze te laten werken. Aangezien dergelijke detailspecificaties veelal niet in leveringscontracten worden vermeld, is men er niet bij voorbaat van verzekerd dat zoiets kosteloos gebeurt.

Slotopmerkingen

Hoewel we ons nog in de introductie- en opbouwfase van het nieuwe retrieval-systeem bevinden, zijn al enkele conclusies te trekken. Al loopt niet alles volgens verwachting, toch is er tevredenheid over de gemaakte softwarekeuze. De zorgvuldige toolevaluatie heeft daar zeker toe bijgedragen. Zoals te verwachten, impliceert onze keuze voor flexibiliteit een lange leercurve. Vooral de implementatie van specifieke wensen vergt investering van veel tijd, inspanning en knowhow van mensen uit de eigen organisatie. Dat is een van de redenen dat nog lang niet alle uiteindelijk gewenste functionaliteit operationeel is. Een andere reden is dat veel nog niet standaard, geïntegreerd, commercieel verkrijgbaar is. Dat verdere ontwikkeling, door externe leveranciers dan wel binnen de eigen organisatie nog veel inspanning vergt, is zeker geen verrassing.

Referenties

Brigit van Berkel & Onno Mastenbroek, Globaal model IR (1996), www.library.uu.nl/proeftuin/ir/globaalir.htm.
Edu Hackenitz, 'CD-ROMs op weg naar het einde'. In: *Informatie Professional* 2(1998)7/8, p. 24.
List of question concerning IR-systems (1997), www.library.uu.nl/proeftuin/ir/criteria.htm.
TREC (1998), http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

Dr. E.G. Sieverts is werkzaam op de Hogeschool van Amsterdam en bij de Universiteitsbibliotheek Utrecht (UBU). Hij is redacteur van Informatie Professional.
Dr. O. Mastenbroek is onderbibliothecaris en hoofd van de afdeling IT bij de UBU.
Drs. N.J. Grygierczyk is projectleider Innovatieve Projecten bij de UBU.