Van PVE via RFI langs POC en RFQ naar finish

Bij het woord zoekmachines denken veel mensen aan Google of Yahoo. Zoekmachines vormen echter ook het instrumentarium om bijvoorbeeld lokale documentcollecties of intranetten te doorzoeken. Software kiezen voor zo'n toepassing is niet altijd een triviale zaak. Anders dan bij zoeken op internet zijn er vaak grote investeringen mee gemoeid. Deze bijdrage beschrijft een best-practice voor dit proces op basis van de ervaring met een keuzetraject bij de UB Utrecht.

Eric Sieverts, Monique Teubner

Bij de Universiteitsbibliotheek Utrecht is al geruime tijd een zoekmachine in gebruik voor het doorzoeken van metadata (titels, auteursnamen, trefwoorden, samenvattingen) van wetenschappelijke artikelen. Dit betreft artikelen waarvoor op basis van licenties of anderszins toegang is verkregen tot de full-text op de sites van diverse uitgevers of leveranciers. Op deze manier biedt het zoeksysteem via één uniform interface toegang tot de full-text artikelen bij al die uitgevers. Het Omega-systeem, waarbinnen deze zoekmachine functioneert, is overigens meer dan alleen dit zoeksysteem. Het biedt gebruikers ook via tijdschriftenlijsten en inhoudsopgaven toegang tot al het elders toegankelijke materiaal en het omvat een beheerssysteem waarin ook de administratie van de digitale collectie kan worden bijgehouden.
De tot dusverre gebruikte zoekmachinesoftware werd door de leverancier helaas niet meer onderhouden of verder ontwikkeld. Omdat de software ook wat functionele tekortkomingen bleek te hebben, moest naar vervanging worden uitgezien. Hiervoor is een zorgvuldig keuzetraject opgezet.

Het hele traject

De stappen in het keuzetraject werden deels opeenvolgend en waar mogelijk ook parallel doorlopen. Het hele traject, zoals geïllustreerd in de figuur op p. 30, zag er globaal als volgt uit.
PvE: op basis van de beoogde toepassing, de ervaring met de tot dusverre gebruikte software en algemene kennis van de ontwikkelingen op retrieval-gebied, is een uitgebreid programma van eisen opgesteld.
Longlist: met dit programma van eisen in het achterhoofd is een marktverkenning gedaan, welke pakketten in principe voor deze toepassing in aanmerking kwamen.
RFI: de leveranciers van de pakketten die in de resulterende *longlist* voorkwamen, is een *Request for Information* toegestuurd. De leveranciers dienden aan te geven aan welke onderdelen van het programma van eisen hun software standaard of onder speciale condities kon voldoen.
Shortlist: op basis van de ontvangen antwoorden, in combinatie met een eerste prijsindicatie, is een selectie gemaakt van pakketten die in aanmerking kwamen voor meer diepgaand onderzoek.
PoC: met de leveranciers van drie pakketten die op de *shortlist* waren overgebleven, is een *Proof of Concept* afgesproken. Op basis van het Programma van Eisen en met gebruik van een representatief deel van de gegevens uit het bestaande informatiesysteem moesten zij zo goed mogelijk vergelijkbare prototypes voor het zoeksysteem inrichten.
Test: wederom op basis van het Programma van Eisen zijn deze systemen door een groep informatiespecialisten, volgens standaard protocollen, op hun functionele eigenschappen uitgetest. Daarnaast keken ict-ontwikkelaars naar de technische kant van de systemen.
RfQ: parallel hieraan ontvingen de drie leveranciers een *Request for Quotation* – een verzoek om een offerte – en werden op basis hiervan al prijsonderhandelingen gevoerd.
Keuze: op basis van resultaten van de *Proof of Concept* en deels ook additionele overwegingen, is voor de stuurgroep van het Omega-project een keuzevoorstel opgesteld en is een pakket aangeschaft.

Op een aantal van deze stappen gaan we hier in meer detail in. Tot slot zullen we ook al enige ervaringen bij de implementatie beschrijven.

Programma van eisen

Onder 'Programma van Eisen' kunnen in de praktijk nogal uiteenlopende soorten documenten worden verstaan. In dit geval ging het concreet om wat de zoek-

bijschrift

machinesoftware functioneel beslist moest kunnen en aan welke technische randvoorwaarden deze moest voldoen. Zoals een deel van die randvoorwaarden wordt bepaald door de lokale ict-omgeving, zo komen de functionele eisen direct voort uit de aard van het materiaal waarvoor en uit de manier waarop het zoeksysteem moet worden gebruikt.

Voor het opstellen van zo'n PvE hoeft meestal niet helemaal met een schone lei te worden begonnen. In onze situatie konden we al uitgaan van een soortgelijk document bij een eerdere zoekmachinekeuze. Dat kon worden aangevuld met eisen uit een lijst van een zusterorganisatie, punten die voortkwamen uit problemen met de tot dusverre gebruikte software, en eisen en criteria op grond van ervaringen van Utrechtse informatiespecialisten met moderne zoeksystemen, vooral op internet.

Eisen die voortkomen uit gebruikersonderzoek of algemeen onderzoek aan informatiezoekgedrag, laten zich lang niet altijd direct vertalen in concrete eisen aan een zoekmachine. De meeste moderne zoeksystemen zijn namelijk zodanig flexibel te configureren dat voor de gebruikers, op basis van algemene functionaliteit, technische werking en mogelijkheden, nog allerlei interfaces en functionaliteiten gerealiseerd kunnen worden. Ervaring met een eerdere keuze had ook geleerd dat een bepaalde functionele eis vaak op technisch verschillende manieren gerealiseerd kan worden. Functionele eisen moeten dus niet te sterk in termen van technische oplossingen worden geformuleerd. Het uiteindelijke PvE bestond uit ruim tweehonderd eisen en beoordelingscriteria (zie kader).

Bij de formulering van het programma van eisen bleek vaak meer sprake te zijn van wensen en beoordelingscriteria dan van harde eisen. Dat illustreert dat in een PvE in principe bij elk punt nog de status en het belang daarvan moeten worden aangegeven. Als alle punten werkelijk 'hard' zouden zijn, zou zeer waarschijnlijk nooit enig product overblijven dat aan het PvE voldoet. Er dient dus zorgvuldig en tamelijk selectief te worden bekeken welke punten harde eisen zijn, waarop pakketten zonder meer afvallen als ze daaraan niet voldoen. De wensen die overblijven moeten dan in principe *nice to have* zijn, liefst met een weegfactor die het belang aangeeft. En er blijken zelfs punten over te blijven die alleen maar *nice to know* zijn.

De genoemde weegfactoren zijn ook van belang voor het proces waarin wordt geselecteerd welke van de pakketten uit

Traject om tot keuze van software te komen

de oorspronkelijke lijst in meer detail onderzocht moeten worden. Die selectie kan gebaseerd zijn op een volstrekt kwantitatieve berekening, bijvoorbeeld via een spreadsheet, maar de wegingen kunnen daarbij ook wat losser als kwalitatieve indicaties gehanteerd worden.

Van longlist naar shortlist

Bij het begin van een keuzetraject is het handig als binnen de organisatie al wat kennis aanwezig is van de belangrijkste producten die op de markt zijn. Vaak is dat ook wel het geval. Anders zijn op internet wel overzichten te vinden. Een andere aanpak is om een gespecialiseerd adviesbureau in te huren. In ons geval is een combinatie van deze drie aanpakken gevolgd. Het door ons ingehuurde bedrijf, *Search Engine Consultancy (SEC)*, bezat zelf al een database met de belangrijkste karakteristieken van de meeste in onze eerste lijst genoemde producten. Op basis van hun expertise hebben zij ook nog enkele leveranciers aan onze longlist toegevoegd.
Uiteindelijk bleek ons PvE een aantal dermate gedetailleerde criteria te bevatten, dat die niet of onvoldoende in de database van SEC voorkwamen. Daarvoor en ook voor de producten waarvan SEC nog niet over de meest recente gegevens beschikte, heeft het bedrijf door middel van een *Request for Information* de ontbrekende gegevens opgevraagd. Zo verzamelde gegevens dienen vervolgens nog kritisch tegen het licht te worden gehouden, omdat veel leveranciers geneigd zijn bijna elke vraag met JA te beantwoorden. Vergelijking met antwoorden op andere vragen of met technische specificaties uit standaard meegestuurde documentatie, kan dergelijke antwoorden vaak enigszins relativeren. Ook moet soms expliciet worden nagevraagd of het antwoord wel precies betrekking heeft op datgene wat in het PvE werd bedoeld.
Op basis van een vergelijking van ons PvE met de zo verwerkte gegevens, adviseerde SEC met welke drie producten verder te gaan. Opmerkelijke redenen waarom – buiten het PvE om – ook nog twee producten waren afgevallen: de leverancier van FAST achtte ons budget onvoldoende om de vragen uit ons RFI zelfs maar te beantwoorden; de geavanceerde zoekmogelijkheden van Collexis zouden alleen tot hun recht komen als voor het hele onderwerpsdomein – in ons geval 'alles' – een voldoende gedetailleerde taxonomie beschikbaar was.
Uiteindelijk is besloten verder te gaan met de bekende marktleiders *Verity* en *Autonomy* en de Nederlandse producent *Irion*. Voor beide eerstgenoemde was dat op basis van hun totale score op het hele PvE, voor Irion op basis van vooral het functionele deel van de eisen, in combinatie met te verwachten lagere kosten.

Proof of Concept

Uit eerdere ervaringen was gebleken dat – ook correcte – respons op een PvE nog niet altijd garandeert dat een zoeksysteem echt precies kan wat de gebruiker wil. Zo bleek in de oorspronkelijk in Utrecht gebruikte zoeksoftware de normaliter uitstekende probabilistische *relevance ranking* volledig onbruikbaar te worden zodra ook truncatie van zoektermen of fuzzy zoeken werden toegepast. Aan alle eisen afzonderlijk werd goed voldaan, maar in combinatie presteerde het systeem ineens heel slecht. Daarom, en ook omdat we zeker wilden weten hoe de systemen met ons eigen materiaal zouden presteren, besloten we de drie overgebleven leveranciers een vrij uitgebreide *Proof of Concept* te laten verzorgen.
Die hield in dat drie zoeksystemen moes-

ten worden gebouwd om te bewijzen dat de betreffende producten aan onze belangrijkste eisen konden voldoen en dat de zoekresultaten bevredigend zouden zijn voor grote hoeveelheden materiaal zoals in het Omegasysteem. Daartoe bevatte elk van de drie (dezelfde) circa tien procent van de ruim tien miljoen artikelen waaruit de totale collectie op dat moment bestond. De inhoud van die tien procent was representatief voor de totale collectie, ten aanzien van zowel de uitgevers van de artikelen als van vakgebied en jaar van uitgave. Ter vergelijking waren dezelfde een miljoen documenten door ons zelf ook doorzoekbaar gemaakt met de oorspronkelijk gebruikte zoekmachine.

De bewijssystemen moesten verder zo geïmplementeerd worden dat ze dezelfde basisfunctionaliteit boden, waarin de belangrijkste functionele eisen uit het PvE ten aanzien van het zoeken verwerkt waren. Er moest dus zowel geavanceerd probabilistisch als klassiek booleaans gezocht kunnen worden. Met het eerste wordt bedoeld dat een rijtje zoektermen een op relevantie geordend resultaat moet opleveren, waarbij liefst ook taaltechnologie wordt ingezet. Verder moesten onder meer truncatie, word-stemming, fuzzy-zoeken, veld-specifiek zoeken en inperkingen op jaren en vakgebieden mogelijk zijn. Het zoekinterface hoefde nog niet aan specifieke eisen voor gebruikersvriendelijkheid te voldoen, aangezien testvragen door ervaren zoekspecialisten uitgevoerd en geanalyseerd zouden worden en er later toch een nieuw interface gebouwd zou worden.

Er is geen moeite gedaan de drie systemen op volledig vergelijkbare hardwaresystemen te laten inrichten. Deze test was immers niet bedoeld als benchmark voor de performance, en bovendien kan bij de meeste systemen achteraf nog zoveel aangepast worden dat resultaten van deze prototypes toch niet volledig representatief zouden zijn voor de uiteindelijke systemen. De drie leveranciers hebben daarom hun systemen elk op eigen hardware geïmplementeerd, voor onze testers via internet op de eigen werkplek toegankelijk. Bij een zo grootschalige PoC is het wel nog de vraag of leveranciers die zomaar gratis willen verzorgen. Grote bedrijven zullen hiertoe meestal wel bereid zijn. Voor kleinere bedrijven zal het inrichten daarvan, zonder gegarandeerd uitzicht op latere inkomsten, een minder vanzelfsprekende investering zijn. Onderhandelingen over voorwaarden en eventuele kosten horen dan ook tot de voorbereiding.

Proef op de som

De praktijktest was vooral bedoeld om de drie systemen te vergelijken en de belangrijkste zoekeisen te testen met behulp van standaard zoekvragen. Of een Booleaanse combinatie goed wordt uitgevoerd en of bij zoeken in het titelveld inderdaad alleen in titels wordt gezocht, is betrekkelijk eenvoudig te controleren. Daarbij kan bovendien meestal worden volstaan met ja/nee-antwoorden. Beoordeling van de relevantievolgorde bij "best-match" zoekvragen – in ons geval een belangrijk criterium – ligt aanzienlijk lastiger. Bij dergelijke zoekvragen, in onze tests telkens rijtjes van vier zoekwoorden, zoals *verbal memory children dyslexia*, werd de testers gevraagd de relevantie in te schatten van de resultaatdocumenten op bepaalde rangnummers (1-20, 51-55, 101-105, 501-505). Daarnaast moesten zij bij elk van die documenten aangeven welke van de zoektermen in welk onderdeel (titel, samenvatting), hoe vaak voorkwamen. Zo hoopten we ook een meer objectieve indruk te krijgen van de factoren die de software bij het ranken van de zoekresultaten liet meespelen.

Voor het uitvoeren van de test is een vrij grote groep informatiespecialisten ingezet. Dat was in de eerste plaats om het vele werk over voldoende mensen te spreiden: dezelfde vraag in drie systemen uitvoeren en elke keer de zoekresultaten analyseren kan op den duur wat saai worden. Een minstens zo belangrijke reden was dat zo al vroeg draagvlak voor het nieuwe systeem gecreëerd kon worden. Deze testers spelen namelijk ook een belangrijke rol om gebruik van het systeem door studenten en medewerkers te stimuleren en te ondersteunen. Ondanks het gesignaleerde gevaar van saaiheid bleek iedereen het in de praktijk spannend te vinden met dergelijke proefsystemen aan de gang te gaan. Men kreeg zo ook het gevoel aan de uiteindelijke beslissing bij te dragen (meer over de testprocedure in bijgaand kader).

Voor de technische eisen uit het PvE was een test lastiger te realiseren. De ict-ont-

Categorieën criteria uit het PvE

Functionele eisen

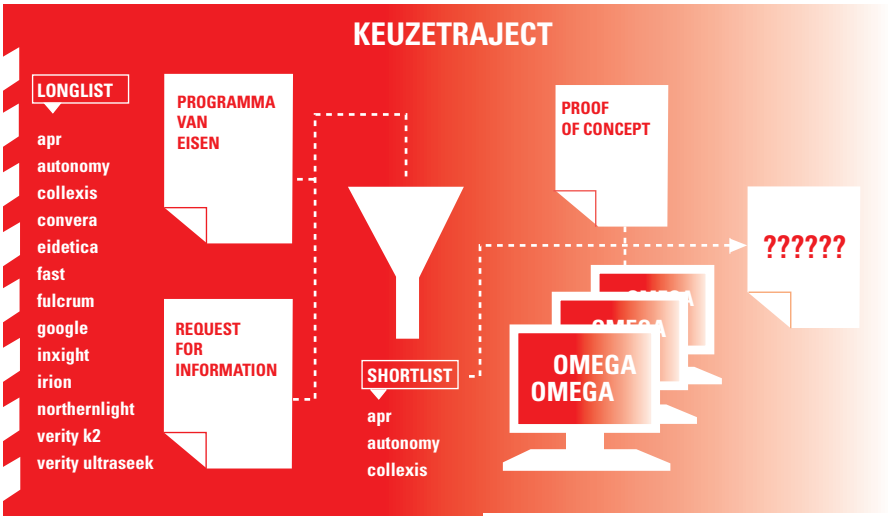
> Indexeringsfunctionaliteit
> Spider functionaliteit
> Retrieval functionaliteit
  - Zoeken
  - Feedback mechanismen
  - Overige zoekvraagverbetering
  - Presentatie zoekresultaten
> Personalisatie
  - Algemeen
  - Attenderingsdiensten
> Overige functies

Leverancierscriteria

> Leverancier
> Softwarelicentie
> Financiën
> Support en training

Technische eisen

> Software
  - Algemeen
  - Performance
  - Configureerbaarheid
  - Toegang en beveiliging
  - Gebruikersinterface
> Hardware
  - Algemeen
  - Analyse & monitoring
  - Application tuning & modelling



Van een lijst van 13 tot dat ene pakket

wikkelaars die verantwoordelijk worden voor het systeem hebben zich uiteindelijk een indruk kunnen vormen door:

- bestudering van technische documentatie van de leveranciers,
- gesprekken, uitleg en technische demonstraties door degenen die de PoC's technisch gerealiseerd hadden,
- bezoek aan andere gebruikers van de betreffende software.

Op basis hiervan schatten zij het gemak in waarmee de systemen aan specifieke eigen wensen kunnen worden aangepast en extra zaken 'bijgebouwd' kunnen worden, alsook de eventuele zwakheden in de daarvoor te gebruiken technieken en tussenlagen. De gesprekken met andere gebruikers gaven bovendien een indruk van de ondersteuning die bij implementatie van de leveranciers verwacht kon worden.

Selectie

De resultaten van de functionele tests vormden een belangrijk element voor de uiteindelijke keuze. De analyse van de resultaten van de 'probabilistische' testvragen was daarbij het meest complex. Moest de relevantie worden beoordeeld op grond van vergelijking met de oude zoekmachine, op grond van de eerder genoemde objectieve factoren of op de persoonlijke, per definitie meer subjectieve beleving van de testers? Uiteindelijk heeft dat laatste – ook met het oog op het beoogde draagvlak – de doorslag gegeven. Eén zoekmachine werd door meerdere personen echt afgewezen, ook al zouden de zoekresultaten volgens een meer objectieve analyse gemiddeld niet significant slechter horen te zijn.

Alles bij elkaar genomen bleken de wat recenter ontwikkelde systemen van Irion en Autonomy voor probabilistisch zoeken de meest bevredigende resultaten op te leveren. Als word-stemming, fuzzy zoeken en verder toegepaste taaltechnologie daar ook bij werden betrokken was er enige voorkeur voor het systeem van Irion. Voor de andere functionaliteit bleken de drie geteste systemen elkaar uiteindelijk niet veel te ontlopen, met dien verstande dat echt 'harde' Booleaanse combinaties door Irion nog apart geprogrammeerd zouden moeten worden.

Andere belangrijke elementen bij de definitieve keuze waren:

- Technische kwaliteiten van het systeem, zoals in de vorige paragraaf al kort besproken.
- Projectmatige realiseerbaarheid van een geheel aan onze eisen aangepast systeem. Daarbij onder meer de afweging hoeveel functionaliteit standaard leverbaar was, hoeveel nog ontwikkeld moest worden, hoe makkelijk en hoe snel het ontwikkelen van extra functionaliteit en integratie met eigen modules zou gaan en welke support bij implementatie en verdere ontwikkeling gegeven zou kunnen worden.
- De leverancier, waaronder degelijkheid van de organisatie, geboden support, helpdesk en dergelijke.
- Kosten van het systeem (waarbij inbegrepen implementatie, onderhoud en training van medewerkers).

Bij dat laatste punt speelden prijsonderhandelingen ook een rol. De aanvankelijk duurder lijkende producten bleken nog een flinke onderhandelingsruimte te hebben, vooral wat betreft de aard van de licentie, meegeleverde extra modules en training. Bij het goedkoper lijkende Irion lieten extra te ontwikkelen functies de prijs juist wat oplopen. Dat leidde ertoe dat de uiteindelijke offertes elkaar nog maar weinig ontliepen. Prijs vormde dus geen beslissende factor in de eindbeslissing.

Hoewel Irion op basis van de PoC functioneel wat betere papieren had, heeft de Omega-stuurgroep op basis van de andere bovengenoemde factoren uiteindelijk voor Autonomy gekozen. Verity, op dat moment nog niet overgenomen door

Autonomy, toonde zich erg verbaasd niet als winnaar uit de bus te zijn gekomen.

Ervaringen

De hier beschreven zeer uitgebreide keuzeprocedure bleek een langere doorlooptijd te hebben dan we zelf al hadden ingeschat. Dat speelde in vrijwel alle fases – zeker ook in die waar we van anderen afhankelijk waren. Het verzamelen en verwerken van de gegevens van de pakketten uit onze longlist en het opstellen van het advies door SEC nam meer tijd dan verwacht. En ook de voorbereiding van de PoC kostte langer en leverde meer praktische problemen dan – kennelijk ook door de leveranciers – was voorzien. Of twee maanden echt te kort was om de PoC-systemen te realiseren, of dat men te laat was begonnen weten we natuurlijk niet. In elk geval leek men zich op het indexeren van een miljoen records te hebben verkeken. Uiteindelijk waren alle drie testsystemen pas een week later beschikbaar dan de met de testers afgesproken datum. En ook toen bleken er af en toe nog problemen met toegankelijkheid en responsetijden.

Het totale traject, van het bedenken van de procedure tot het tekenen van het contract, nam uiteindelijk vijftien maanden in beslag. Het besluit bij welke drie leveranciers een PoC zou worden aangevraagd lag ongeveer halverwege deze periode. Dat die eerste stap al betrekkelijk lang geduurd heeft, kwam door een aanvankelijk andere opzet om tot een shortlist te komen. Een aantal erkende zoekmachine-experts in binnen- en buitenland was gevraagd om als externe adviseurs een rijtje favorieten te noemen voor het type toepassing als de onze. Dit bleek in de praktijk totaal niet van de grond te komen.

Na de definitieve keuze heeft de implementatie van het systeem ook nog meer tijd gekost dan voorzien. Een van de redenen was de grootte van de te doorzoeken collectie. Tussen een PoC op een miljoen records en een werkelijk systeem met intussen meer dan twaalf miljoen records bleek nog een wezenlijke technische grens te worden overschreden. Voor het soort veldgestructureerde gegevens waarmee wij werken, bleek het onderste uit de kan gehaald te moeten worden op het laagste niveau van hardware en operating system, om ook voor complexe zoekopdrachten nog snelle responsetijden te garanderen. De keuze van de hardware moest daar op het laatste moment nog aan worden aangepast. Omdat geen van de gesproken gebruikers in een vergelijkbare situatie had verkeerd, waren wij voor deze problematiek niet gewaarschuwd.

Bij een eventuele volgende keer zouden we het keuzetraject niet wezenlijk anders inkleden. Wel realiseren we ons nu dat zelfs een goed doorlopen keuzetraject met een vrij uitgebreide PoC nog geen garantie hoeft te zijn dat de uiteindelijke implementatie in een tamelijk complexe toepassing, geheel zonder problemen verloopt. Dit keuzetraject bleek zich wel heel goed te lenen om vrij veel medewerkers als testers een actieve rol te laten spelen, ten behoeve van het draagvlak voor het project als geheel. <

Eric Sieverts is consultant sector Innovatie & Ontwikkeling bij de Universiteitsbibliotheek Utrecht.
Monique Teubner is projectmanager Sector Innovatie & Ontwikkeling bij de Universiteitsbibliotheek Utrecht.

Puntsgewijze de testprocedure

- Opzetten testvragen op basis van de belangrijkste zoekeisen uit het PvE: alle vragen zijn tevoren op de oude zoekmachine uitgetest (niet door de vragenmaker) en aangepast als ze niet eenduidig bleken.
- Samenstelling van de testgroep was gericht op kennis van informatie zoeken, maar ook op het creëren van draagvlak in de organisatie.
- Er was voorlichtingsbijeenkomst voor testers vooraf en evaluatiebijeenkomst (met borrel) achteraf.
- Vijftien testers werd gevraagd tijdstippen aan te geven (24 uur in twee weken) waarop men kon testen. In de praktijk bleken er aanzienlijke verschillen in werkelijk benodigde tijd per tester. Dat kon van toebedeelde vragen afhangen, maar was ook individueel bepaald.
- Op elke vraag werden minstens twee testers gezet. Aanmoediging onderling te overleggen bleek niet te worden opgevolgd.
- Elke tester stelde zijn vragen aan alle drie zoekmachines: deze moesten in dezelfde periode voor de testers beschikbaar zijn.
- Er werd getest op de eigen werkplek. Bij evaluatie bleek men in het vervolg liever in een apart testlokaal te testen, weg van de werkplek, met een direct aanspreekbare begeleider.
- Vooraf waren standaard antwoordformulieren ontworpen.
- Elke tester kon gebruik maken van de (oude) Google toolbar in Firefox om zoektermen en woordstammen in zoekresultaten in afzonderlijke kleuren te laten oplichten. Dit scheelde heel veel tijd.
- Tijdens de testperiode werd enkele malen het testdocument aangevuld en rondgestuurd.
- Achteraf werden de testresultaten door enkele mensen geanalyseerd. Bij twijfel aan een gegeven antwoord werd die vraag nogmaals gecontroleerd.