Pieter Parmentier

Kennis binnen vinger(afdruk)bereik

Hoe bouw je een wereldwijd netwerk en hoe financier je dat? Waarom sluiten 350 'Bibliotheca Virtual' in Latijns-Amerika en de belangrijkste milieuorganisaties in Zuidelijk Afrika zich bij SHARED en ENF aan? Barend Mons en Pieter Parmentier ontwikkelden onafhankelijk van elkaar een internetwerk om (wetenschappelijke) kennis uit te wisselen. Nu beschikken ze over de technologie om hun doel gezamenlijk te realiseren. Hun verhaal over het bouwen van een kennisnetwerk, het vinden van deskundigheid en het terugbrengen van documenten tot hun essentie.

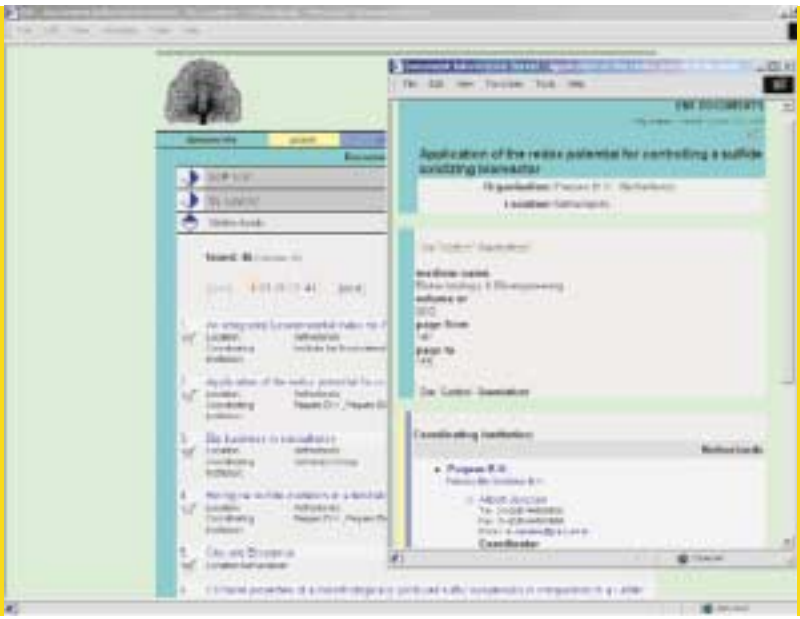
In 1997 ontwikkelen twee idealisten onafhankelijk van elkaar de filosofie dat het mogelijk moet zijn om, gebruik makend van internet, op grote schaal kennis toegankelijk te maken door kennis over kennis uit te wisselen. De voordelen lijken vooral groot voor deskundigen in ontwikkelingslanden. Bescheiden starten de twee ieder met één vakgebied, gezondheid en milieu; maar wel meteen wereldwijd.
Vier jaar later komen hun wegen bij elkaar en zijn de ambities bijgesteld: naar boven. Waarom niet alle goede literatuur, alle deskundigen en alle toporganisaties op het gebied van gezondheid, milieu en landbouw verenigen? Velen hebben de twee in de afgelopen jaren welwillend te woord gestaan en geluisterd naar de enthousiaste verhalen. Ondertussen kregen de twee onafhankelijke initiatieven langzaam maar zeker handen en voeten. Voor milieu heet dat netwerk ENF en de naam van het gezondheidsnetwerk is SHARED.

SHARED

De oorspronkelijke betekenis van deze afkorting is 'Scientists for Health and REsearch for Development'. SHARED is ontwikkeld op kosten van de EU om kennis over gezondheidsonderzoek beter te kunnen uitwisselen tussen deskundigen in Afrika en Europa. Vanaf het begin is intensief samengewerkt met diverse (internationale) instanties. Nu heeft SHARED een brede opzet en bevat meer dan 2000 gezondheidsprojecten in 161 landen waarbij 1966 organisaties zijn betrokken en 3894 wetenschappers. De projecten worden gefinancierd door 380 verschillende organisaties.
In Afrika wordt het netwerk ondersteund door een aantal regionale steunpunten (Focal Points) die individuele instituten en wetenschappers helpen om het netwerk optimaal te benutten. In Latijns-Amerika is aansluiting gevonden bij Bireme, een internationaal bibliotheeknetwerk, dat al beschikte over een indrukwekkende infrastructuur.
Met de additionele data van NIH in de Verenigde Staten (meer dan 60.000 records) en de projecten uit Latijns-Amerika zal SHARED dit jaar naar meer dan honderdduizend projecten groeien, gepaard met een grote stijging van het aantal wetenschappers en instituten die via SHARED gematched en gevonden kunnen worden op basis van hun interesse en kennisprofielen.
Voor Azië zijn de plannen even ambitieus maar nog niet zover gevorderd als in de rest van de wereld. In tegenstelling tot Latijns-Amerika ontbreekt het in Azië aan een operationeel netwerk en in tegenstelling tot Afrika is het gebruik van internet in een aantal landen zeer intensief. Landen als Zuid-Korea en Japan zijn daarin zelfs wereldleiders.

ENF

Environmental Networking Facilities is in 1997 ontstaan uit een particulier initiatief met als doel om een specifieke kennisuitwisselingsbijdrage te leveren aan het oplossen van milieuproblemen in de wereld. De eerste doorbraak werd bereikt toen een achttal illustere deskundigen op het gebied van milieu, IT en internationaal onderzoek (Jacqueline Cramer, Wouter van Dieren, Marius Enthoven, Hans van Ginkel, Jón Kristinsson, Karel J. Samson, Leo Smits, Eckart Wintzen) bereid bleken om toe te treden tot de Board of Patrons. Dat een dergelijk netwerk voor kennisuitwisseling van grote waarde zou kunnen zijn, stond bij hen direct als een paal boven water.
Door het rechtstreeks benaderen van duizenden milieudeskundigen via internet en ze in staat te stellen via de community van ENF in contact te komen met andere deskundigen, ontstond binnen enkele jaren een netwerk verspreid over 75 landen.
In aanvulling op deze individuele methode om deelnemers te werven, is kortgeleden in Nederland een prototypeproject gestart waaraan in eerste instantie vier organisaties meedoen – Alterra, Cordaid, Globus en Novib – die gezamenlijk de terreinen milieu, natuur, duurzame ontwikkeling en ontwikkelingssamenwerking bestrijken.
Doel is om de nieuwste software te testen met de 'content' van de participanten en om vervolgens alle relaties van de



Schermvoorbeeld uit de ENF-applicatie

participerende organisaties uit te nodigen om ook deel te nemen. Op 1 januari 2002 komt de vernieuwde ENF-site in de lucht en is vanaf dat moment wereldwijd toegankelijk voor alle deskundigen en instellingen op de bovengenoemde terreinen.
Een aantal andere Nederlandse organisaties heeft al laten weten zich te willen aansluiten en hetzelfde geldt voor milieuorganisaties in Latijns-Amerika en Zuidelijke Afrika. Wat begon als het particulier initiatief van een eenling, wordt per die datum omgezet in de Stichting ENF die tot doel krijgt het netwerk verder uit te bouwen en te onderhouden.

Focal Points

Tussen SHARED en ENF en het vergelijkbare, maar nu nog naamloze, initiatief voor de landbouwsector (waaraan onder meer het Duitse GTZ en de FAO meewerken) zal een goede coördinatie worden gewaarborgd. De secretariaten van de drie initiatieven zullen regelmatig overleg hebben en bovendien zijn de drie zusterinitiatieven alle op dezelfde technologie gebaseerd. Organisatorisch worden de drie netwerken dus niet gekoppeld, maar technisch vormt het geen enkel probleem. Zoekopdrachten (we praten liever over 'vindopdrachten') in de ENF-omgeving kunnen dus heel goed leiden tot een 'hit' in de agrarische of de gezondheidssector en wellicht is het juist deze 'kruisbestuiving' die tot de kansrijkste ontwikkelingen leidt.

Gesteund door Wereldbank, Bireme, Alterra, Globus, Novib, Cordaid, NWO, GTZ, WAO, WHO, EC, Nature, Elsevier en nog vele andere organisaties mikken de initiatiefnemers op een wereldwijd stelsel van Focal Points ter ondersteuning en uitbouw van de netwerken. Waar deze FP's via dezelfde infrastructuur verschillende vakgebieden kunnen ondersteunen zal zeker van een gezamenlijke infrastructuur gebruik worden gemaakt. Hoe het stelsel van FP's wordt opgebouwd zal sterk afhankelijk zijn van de lokale/regionale situatie.
Uit ervaring blijkt dat het aantal FP's in Afrika groter moet zijn dan in bijvoorbeeld Latijns-Amerika. Ook de taken van de FP's moeten worden afgestemd op de situatie ter plaatse. Als deskundigen op afgelegen plaatsen niet kunnen beschikken over een betrouwbare internetverbinding kan het nodig zijn dat het FP een cd-rom service opzet (binnenkort komt de software en de netwerkinformatie ook beschikbaar op cd-rom) of zelfs een postservice.

Hoe houd je een wereldwijd netwerk operationeel en hoe financier je dat? Waarom sluiten 350 'Bibliotheca Virtual' in Latijns-Amerika en de belangrijkste milieuorganisaties in Zuidelijk Afrika zich bij SHARED en ENF aan? Waarom hebben tot nu toe ruim 1800 organisaties en 4300 deskundigen zich aangesloten bij ENF en SHARED en wat is de rol van andere partners ?

Een deel van de antwoorden is terug te voeren op de gedrevenheid van de initiatiefnemers, op het concept, de technologie en op het feit dat belangrijke partijen met grote internationale netwerken als NWO en GTZ (Gesellschaf für Technische Zusammenarbeit) aan de wieg stonden van SHARED.
Maar het succes staat of valt met de gemotiveerde deelname van organisaties in ontwikkelingslanden en juist op dat punt is recent grote vooruitgang geboekt. Mede daardoor is SHARED (dat vanaf het begin een samenwerkingsverband met Afrikaanse partners had) volledig opgenomen in de formele politieke besluitvorming in Afrika en Latijns-Amerika, met brede internationale steun.
Zeer recent hebben de 'Southern African Network for Training and Research on the Environment' (SANTREN) en het 'Institute of Environmental Studies' (IES) zich gemeld om te participeren in het milieunetwerk. Ook in de westerse landen is de motivatie groot om mee te doen aan dit letterlijke 'global network'. Door medewerking van Europese en Amerikaanse wetenschappelijke koepelorganisaties kan binnen enkele jaren een groot deel van de aangesloten wetenschappers toegang hebben tot het netwerk.

Problemen

Een belangrijke reden voor de positieve reacties uit Afrika en Latijns-Amerika is dat het beroemde 'ownership of data'-probleem, dat normaal gesproken een groot struikelblok vormt, door de gebruikte technologie goed wordt ondervangen. Niet de documenten of projectinformatie worden op de centrale server gezet, maar alleen de fingerprints met hun verwijzing naar de originele bron. De beheerder van de bron kan de toegang tot het origineel vervolgens zo goedkoop of duur maken als hem/haar belieft. Ook wordt door het opzetten van 'mirror servers' voorkomen dat politieke of technische problemen voor (groepen) deelnemers leiden tot beperking van de toegang tot de informatie.
'Normale' projectproblemen doen zich bij de implementatie en het gebruik van de technologie natuurlijk wel voor. Zo blijkt eens te meer dat heel veel afhangt van de persoon van de projectleider ter plaatse[1]. Ook blijkt dat het bijhouden van de informatie een veel minder populaire bezigheid is dan de eerste implementatie en dat het tot stand brengen van een simpele internetverbinding niet overal in de wereld zo simpel is.
Daarnaast blijkt af en toe dat het niet voor iedereen vanzelfsprekend is dat iedereen toegang krijgt tot meer kennis. Sommige managers en organisaties vinden het ronduit bedreigend dat meer medewerkers kunnen beschikken over veel meer informatie.

Kwaliteit

Een ander voor de hand liggend probleem is wie de kwaliteit moet bewaken van de aangeleverde informatie. De eenvoudigste oplossing is om die taak neer te leggen bij de deelnemende organisaties, die daar dan zelf een werkgroep of een persoon voor aanwijzen. Ook de Focal Points kunnen daarin een rol vervullen, maar dan wordt de vraag nog belangrijker bij welke organisatie in welk land het Focal Point moet worden ondergebracht.
Gelukkig werkt de grote doorzichtigheid van het systeem in de hand dat de deelnemende deskundigen ook zichzelf beperkingen opleggen. Het is immers voor iedereen zichtbaar welke kwaliteiten of publicaties een ander zichzelf toedicht en overdrijving wordt dan spoedig aan de kaak gesteld.

Voorbeeld van een zoekresultaat binnen een specifieke 'community'

BVSA

Bibliotheca Virtual en Salud Y Ambiente is het netwerk van virtuele bibliotheken voor gezondheid- en milieuinformatie in Latijns-Amerika en het Caraïbisch gebied. Dit initiatief van de World Health & Pan American Health Organisation omvat nu al 350 centra in 21 landen gericht op het verzamelen, verwerken en verspreiden van informatie over milieuvervuiling die bedreigend is voor de gezondheid van de mens.
De initiatiefnemers hebben besloten om met de BVSA aan te sluiten op het SHARED-netwerk en daarmee op het milieunetwerk, gebruikmakend van dezelfde technologie. Zij mikken uiteindelijk op wereldwijde toepassing van hun concept. Op de kortere termijn moeten/zullen in ieder Latijns-Amerikaans land steunpunten opgericht worden ter versterking van de regionale coördinatie met eigen personeel, budget en apparatuur.

Financiering

De financiering van dit alles rust op drie pijlers. Een deel van de omzet bij verkoop van de technologie aan commerciële klanten, vloeit in het SHARED-fonds. Dit zal onder meer worden gebruikt om mee te betalen aan het SciDevNet-initiatief, maar ook wordt daaruit de participatie aan het gezondheidsnetwerk mogelijk gemaakt voor organisaties die dat zelf niet kunnen betalen. Daarnaast wordt een beroep gedaan op subsidies van grote (overheids-)organisaties in de westerse wereld om de samenwerkende netwerken (milieu, landbouw en gezondheid) en hun focal points in het zuiden te ondersteunen.
Tot slot betalen deelnemende organisaties die daartoe wel in staat zijn een beperkte onderhoudsbijdrage die recht geeft op gebruik van de help desk en updates van de software en die de administratieve ondersteuning mogelijk zal maken.

SciDevNet

De hier beschreven initiatieven worden nog eens verder versterkt door (en andersom) het 'Science for Development' (SciDev) plan van Nature Magazine. Deze gerenommeerde wetenschappelijke uitgever was een van de eerste commerciële klanten voor de Collexis®-technologie. Zij sponsort, samen met verschillende internationale donoren en de Third World Academy of Sciences, de bouw van een 'vrij toegankelijk internetinformatienetwerk, gericht op het onderzoeken van de verbindingen tussen wetenschap, technologie, innovatie en ontwikkeling'. De kern van dit netwerk wordt een website die vooral dienst doet als elektronisch nieuwsblad. Met nieuwtjes en artikelen geschreven door een professionele staf van wetenschapsjournalisten en de vindmachine die toegang geeft tot informatie en contacten op wetenschappelijke en technologische deelterreinen.
Een belangrijke doelgroep voor dit SciDevNet (Science & Development Network) is de tot op heden slecht bereikte groep van beleidsmakers en beslissers van overheden (met name in de Derde Wereld). Ook voor hen zullen verteerbare artikelen geschreven worden. Bovendien zullen wetenschappers uit diezelfde landen veel eerder dan tot op heden mogelijk was, toegang krijgen tot nieuwe publicaties. Het kan bij publicatie in tijdschriften immers lang

Recept voor het maken van een fingerprint

Men neme een digitale tekst en een thesaurus*. De tekst, dat wil zeggen: een bestand dat digitaal is vastgelegd, kan worden onderworpen aan de IKA-procedure (zie kader op p. 45). Wat overblijft, is een setje concepten dat de eigenaar ter controle krijgt aangeboden. Dit hele proces duurt minder dan een seconde.

Gaat hij/zij akkoord, dan worden deze concepten, maar in feite de getallen waaraan ze gekoppeld zijn, opgestuurd naar de centrale 'fingerprintdatabase'. Natuurlijk wordt aan de fingerprint eveneens meegegeven waar het originele document gevonden kan worden.

Het blijkt mogelijk om op deze wijze enorme aantallen fingerprints per dag te produceren en op te slaan. Zonodig de hele bibliotheek op één laptop.

Het vinden van informatie werkt vanzelfsprekend met dezelfde technologie. De vindopdracht (hoe meer tekst hoe beter) wordt omgezet in een fingerprint die vervolgens vergeleken wordt met de inhoud van alle andere opgeslagen fingerprints. Als het zoek profiel van een persoon eenmaal bekend is, kunnen nieuwe documenten in het systeem die passen bij dat profiel desgewenst automatisch worden aangeboden.

Van iedere persoon en organisatie en van alle documenten en/of projecten die worden aangeboden, worden volgens bovenstaande methode fingerprints gemaakt. Aangezien het systeem dynamisch is, zal iedere wijziging onmiddellijk zichtbaar zijn.

Wel is het zo dat alle aangeboden informatie eerst voor controle een wachtkamer moet passeren. Wie de controles uitvoert, is een kwestie van goede afspraken maken.

* Thesauri zijn er in vele soorten en maten. Van de meeste vakgebieden is inmiddels wel een thesaurus beschikbaar en anders wel samen te stellen of in korte tijd op te bouwen. De grootste is de Unified Medical Language System (UMLS) met ruim 900.000 concepten en 2,8 miljoen termen in de laatste versie.

duren voordat de publicatie een feit is en de klant bereikt. Belangrijk is verder dat het hier niet alleen gaat om publicaties over afgeronde onderzoeken, maar ook over lopende projecten waardoor informatie jaren eerder op de plek van bestemming komt dan tot nu toe mogelijk was.

7.000 artikelen per dag

Om duidelijk te krijgen waarom bovengenoemde initiatieven nu kunnen doorbreken, is het nodig om even de bibliotheek binnen te gaan. Als het waar is dat we wereldwijd kunnen kiezen uit 24 miljoen boeken en dat er jaarlijks 75.000 bij komen. En als alleen Japanse medische onder-

zoekers al bijna 110.000 artikelen per jaar schrijven. Dan is het geen wonder dat de 100.000 tijdschriften in de wereld 7.000 artikelen per dag publiceren. Vreemd is het dan evenmin dat in de gemiddelde bibliotheek 10 procent van de tijdschriften zorgt voor 80 procent van het gebruik en dat 50 procent van de tijdschriften nooit geraadpleegd wordt. Er is gewoon veel te veel.
Reeds in 1963 in het boekje: 'Automatie, industriële en culturele revolutie; mens en computer' wordt de computer grote mogelijkheden toegedicht voor het terugzoeken van literatuur. Ir G. Nielen stelt in zijn hoofdstuk over de toekomst van de elektronische apparatuur: 'Wij zullen dankbaar zijn als wij het (information retrieval) met machines kunnen, want veel ervan is verschrikkelijk werk.' Hij is ervan overtuigd dat de oplossing nabij is en over de vertaalmachine zegt hij zelfs dat die binnenkort beschikbaar zal komen. Bijna veertig jaar later begint het erop te lijken.

De opkomst van het internet heeft de situatie van 'te veel' in eerste instantie niet verbeterd, integendeel. Met de 100 beschikbare zoekmachines kun je 2,1 miljard pagina's vinden en dagelijks komen er nog 7 miljoen bij. De doorsnee trefwoord- (Booleaanse) zoekmachine levert als het tegen zit bij een zoekopdracht 10.000 wel plausibele maar niet gewenste hits.
We hebben dus te maken met drie problemen: de productie is enorm, door de digitalisering wordt de bereikbaarheid sterk uitgebreid, de huidige zoekmachines leveren te veel hits.

Semantisch web

Wat we zouden willen, is een vindsysteem waarmee we alleen krijgen wat we willen hebben. Het moet dus tegemoet komen aan ieders specifieke wensen en tegelijkertijd moet alle goede en correcte informatie gelijkwaardig zijn; een gelijke kans hebben om gevonden te worden.
En dat moet niet alleen gelden voor officieel gepubliceerde artikelen, maar ook voor de zogenaamde 'grijze literatuur' die nu nog zelden onze 'grijze massa' kan bereiken.
Om dit mogelijk te maken, zou er zich een 'semantisch web' moeten ontwikkelen of ontwikkeld moeten worden. Tim Berners-Lee (de uitvinder van WWW, html en http-protocols) bedoelde met deze term een extra internetinformatielaag die het mogelijk moet maken om met machines zonder centrale controle en regels een veelheid aan toepassingen (zoals zelfdenkende huizen bijvoorbeeld) mogelijk te maken. Maar ook en vooral het op de persoon toegesneden geautomatiseerd zoeken, vinden en afleveren van informatie.
Het semantische web maakt het onnodig om grote hoeveelheden informatie centraal op te slaan. Als maar duidelijk is waar de bestanden onder de oppervlakte drijven om op het juiste moment als een soort magma naar boven te komen. Het wordt van groot belang dat de bovenlaag redelijk homogeen is (gestandaardiseerde metadata) en voldoet aan een (ISO) standaard zoals de Dublin Core of de 'Digital Object Identifier' (DOI). DOI is ontwikkeld omdat andere standaarden zoals ISBN niet bedoeld en derhalve ongeschikt zijn voor het doorzoeken van elektronische publicaties en de Uniform Resource Locator (URL) die wel de plaats maar niet de inhoud aanduidt.

IKA

Dr. Erik van Mulligen, een van de uitvinders van de Interactieve technologie en CTO van Collexis BV, beschrijft de techniek als volgt:
Van de op het scherm geselecteerde tekst worden door de software allereerst de zinnen geselecteerd. Van alle woorden in de zin wordt vervolgens de basisvorm geïdentificeerd en de stopwoorden worden verwijderd. Van de lijst die overblijft, zoekt de software de mogelijk bijbehorende concepten. Vervolgens worden de basisvormen van de woorden geclusterd en clusters vervangen individuele woorden.
Tot slot worden de beste concepten op basis van deze cluster geselecteerd door het nemen van de volgende stappen:
• Per cluster wordt de *verspreiding* berekend, dat wil zeggen: de gemiddelde afstand in woorden tussen woorden die aan hetzelfde concept refereren.
• De *frequentie* geeft het aantal woorden of clusters die refereren naar een concept.
• De *omvang* van de cluster geeft het aantal concepten/alternatieven in een cluster.
• De *specificiteit* wordt gebaseerd op het totaal aantal concepten waaraan gerefereerd wordt door een basiswoord in de tekst.
• De *overeenkomst* geeft aan welk deel van het concept letterlijk is teruggevonden in de tekst.
• De *combinatie* stap geeft aan hoe vaak een concept voorkomt in combinatie met andere concepten uit dezelfde tekst.

Als deze stappen zijn doorlopen (in een fractie van een seconde) krijgt de opdrachtgever een overzicht te zien van de concepten die de betreffende tekst het beste dekken, de 'vingerafdruk'. Wordt deze selectie akkoord bevonden, dan zal de vingerafdruk, compleet met verwijzing naar de eigenaar van de tekst, toegevoegd worden aan een centrale 'Collexion' naar keuze.

Plagiaat

Wat tot nu toe ontbrak was de krachtige software die het mogelijk zou maken om te identificeren en te beschrijven alsmede te koppelen aan de wensen van de gebruiker en vervolgens te verbinden met gestandaardiseerde metadata en 'locator technology' zoals DOI. Met dit laatste puzzelstukje op z'n plaats wordt het mogelijk om in korte tijd diverse gerelateerde artikelen met elkaar te vergelijken[2] of virtuele communities op te zetten die gebaseerd zijn op een gezamenlijke belangstelling en behoefte.

Noten

1. Een actieve netwerkmanager in Zimbabwe zag kans om in enkele maanden tijd honderden projecten op het netwerk te krijgen.
2. Artikelen die het hoogst scoren bij een vindopdracht zijn (naast het artikel dat gebruikt is voor het zoeken) waarschijnlijk artikelen die grotendeels uit plagiaat bestaan.

Met dank aan Jan Velterop (directeur DocDemon en uitgever van de BioMed Central Group), Erik van Mulligen (CTO van Collexis BV) en Barend Mons (initiatiefnemer SHARED).

Pieter A. Parmentier is directeur van ENF.