Wouter Mettrop
Bibliotheek CWI
Postbus 94079
1090 GB Amsterdam
wouter.mettrop@cwi.nl

Ay-ling Ong
Bibliotheek CWI
Postbus 94079
1090 GB Amsterdam
ay.ong@cwi.nl

Online publishing

Wiskunde-informatie in een handomdraai

Sinds eind 2002 hebben internetgebruikers de beschikking over de zoekmachine EULER. Deze machine is speciaal bedoeld voor het verkrijgen van wiskundige informatie. EULER (EUropean Libraries and Electronic Resources in mathematical sciences) is ontstaan uit een Europees samenwerkingsverband. Wouter Mettrop, informatiespecialist van de bibliotheek van het Centrum voor Wiskunde en Informatica (CWI), en Ay-ling Ong, bibliothecaris van het CWI, waren bij het ontstaan betrokken.

Om wiskunde te kunnen bedrijven is informatie nodig. Je wilt als wiskundige weten wat je collega's doen en al gedaan hebben. Je zoekt in de eerste plaats naar publicaties. Van daaruit probeer je informatie over ander onderzoek en andere wiskundigen te traceren. Volledigheid, precisie, correctheid, snelheid en kosten spelen hier een rol. In deze eigenschappen begint de traditionele informatievoorziening te kort te schieten. Een wiskundige, die gewend is met Google direct en zonder kosten het adres van een collega te vinden, wil ook direct alle publicaties vinden die voor zijn werk relevant zijn. Dit gaat echter minder makkelijk.

Een Europees samenwerkingsverband tussen het Centrum voor Wiskunde en Informatica in Amsterdam, het Fachinformationszentrum in Berlijn en de universiteitsbibliotheken van Göttingen en Florence, onder de naam EULER, is de wiskundige te hulp geschoten. EULER biedt via het *Portal to Mathematics Publications* [1] op het internet toegang tot de wereldwijde wiskundige informatie.

De situatie tot nu toe

Van oudsher staan de wiskundige een aantal informatiebronnen ter beschikking: de referaattijdschriften en bibliotheekcatalogi. Recentelijk zijn daar de internetzoekmachines aan toegevoegd.

Referaattijdschriften delen, in gedrukte of electronische vorm, artikelen in naar onderwerp. Voorbeelden zijn de *Mathematical Reviews* en het *Zentralblatt für Mathematik* die aan de hand van de MSC (*Mathematical Subject Classification* [2]) onderwerpscodes toekennen aan nieuw verschenen artikelen en in de gedrukte versie periodiek publiceren wat er binnen elke code verschenen is. Dit levert clusters van artikelen op die min of meer over hetzelfde onderwerp gaan. Deze clustering kan ook anders tot stand komen. De ISI (*Institute for Scientific Information* [3]) koppelt in de *Science Citation Index* artikelen door te kijken naar referenties. Twee artikelen worden aan elkaar gekoppeld wanneer de ene refereert naar de ander. Al jaren lang zijn deze bestanden bereikbaar via internet: *MathSciNet* (Mathematical Reviews [4]), *Zentralblatt MATH* (Zentralblatt für Mathematik [5]) en *Web of Science* (Science Citation Index [6]). Het grote voordeel van dergelijke bestanden is dat ze redelijk volledig en precies zijn, en dat de inhoud gecontroleerd is. Maar ideaal zijn ze zeker niet. Ze kosten geld, werken langzaam en lopen achter bij de hedendaagse stijl van wiskunde beoefening. Heb je een artikel in een referaattijdschrift gevonden, dan heb je het geschrift zelf nog niet. De documentleverantie is verre van ideaal. Via de internetversies is het wel mogelijk kopieën van tijdschriftartikelen te bestellen, vaak tegen hoge kosten, maar boeken verkrijgt men niet. Er zijn ook inhoudelijke bezwaren: zo zijn de rapporten nauwelijks of niet opgenomen en komen webpagina's al helemaal niet voor.

Misschien wel de oudste informatiebron is de bibliotheekcatalogus. Deze bron, die dikwijls goed ontsloten is, kent uiteraard een vorm van vergaande documentleveran-

tie: boeken kunnen geleend worden en tijdschriftartikelen gekopieerd. Nadeel is dat in verreweg de meeste gevallen artikelen uit tijdschriften en proceedings niet zijn opgenomen.

Nieuwe bronnen maken opgang: de 'gewone internetzoekmachines', zoals *Google* [7] en *AltaVista* [8]. Men kan zonder kosten zoeken naar webdocumenten. Een dergelijke machine streeft ernaar zo veel mogelijk te indexeren. Qua zoekdomein is een zoekmachine als Google dan ook helemaal niet te vergelijken met MathSciNet, maar ze vullen elkaar wel aan en ook via Google worden soms reguliere publicaties gevonden. Auteurs bieden hun publicaties immers dikwijls aan via hun homepages. Ook de opkomst van de Institutional Repositories (waarbij een instituut al haar publicaties electronisch aanbiedt) speelt de internetzoekmachine in de kaart. Helaas zijn de resultaten van deze gewone internetzoekmachines verre van precies — ze bevatten uiteraard veel irrelevant materiaal — en zijn ze ook zeker niet volledig. Bovendien zijn de gegevens waar deze machines uit putten ongecontroleerd op het internet gezet. De inhoud van databases die doorzoekbaar zijn op het web (zoals bibliotheekcatalogi) is meestal niet terug te vinden met een gewone zoekmachine. Slechts een klein gedeelte van de reguliere wetenschappelijke publicaties is langs deze weg te vinden. Er bestaan wel gespecialiseerde zoekmachines, die een grotere precisie bereiken. Dikwijls doorzoeken zij slechts een beperkt stuk internet. Een zoekmachine die op dit moment redelijk in staat is om relevante wetenschappelijke informatie van internet te halen met een vrij grote precisie, is *Scirus* van Elsevier [9].

Tot op dit moment bestond er geen informatiebron, die de voordelen bundelt van referaattijdschriften, internetzoekmachines en bibliotheekcatalogi. De internetzoekmachine is ongecontroleerd, onvolledig en meestal niet precies, de bibliotheekcatalogus mist ontsluiting op artikelniveau en is qua inhoud afhankelijk van de eigen collectie, en de commerciële databases brengen kosten met zich mee, zijn traag en zijn ook niet volledig.

Het project EULER

Het FIZ (Fachinformationszentrum Karlsruhe) in Berlijn, het CWI en de universiteitsbibliotheken van Göttingen en Florence hebben deze onbevredigende situatie onderkend en zijn een Europees project begonnen met de naam EULER (EUropean Libraries and Electronic Resources in mathematical sciences). Het project is gestart in 1998. In de periode tussen 1998 en 2000 heeft het eerste deel van het project plaatsgevonden, waarin gebruikerswensen in kaart zijn gebracht en waarin een daadwerkelijk prototype is gebouwd. Tussen 2001 en 2002 heeft de tweede helft plaatsgevonden (onder de naam EULER-TAKEUP), waarin het prototype omgebouwd is tot een reële service. Het project is inmiddels afgesloten. Het eindproduct is een 'one-stop shopping site' voor iedereen met interesse in de wiskunde [1]. Het systeem kenmerkt zich vooral door geen kosten te berekenen aan de eindgebruiker terwijl het volledigheid nastreeft.

Alvorens in te gaan op specifieke eigenschappen van EULER, tonen we een eenvoudig voorbeeld dat aangeeft dat EULER tegemoet komt aan de eerder beschreven bezwaren. We zoeken naar een artikel van een van de meest geciteerde auteurs van dit moment: Saharon Shelah. We kiezen een artikel dat redelijk bekend is (volgens de ISI 22 keer geciteerd in de afgelopen negen jaar) en dat zo oud is (uit 1993) dat al onze bronnen ruim de tijd hebben gehad het artikel op te nemen. We krijgen het volgende zoekresultaat:

Auteurs: Hodges W., Hodkinson I., Lascar D., Shelah S.

Titel: The small index property for omega-stable omega-categorical structures and for the random graph.

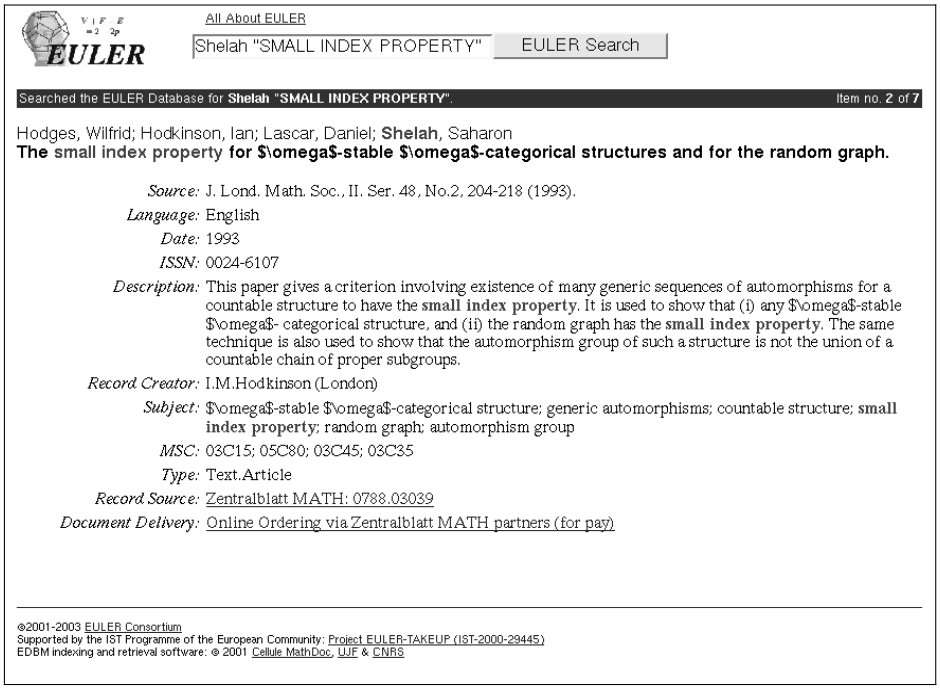
Bron: Journal of the London Mathematical Society-second series 48: 204-218 Part 2 Oct 1993.

Eerst zoeken we, op auteursnaam en titel, met Google. Resultaat: 7 webdocumenten, elk met overzichten van publicaties van één van de auteurs. Meer gegevens dan we al hadden vinden we niet. Dan zoeken we in EULER, ook weer op auteursnaam en titel. Het artikel wordt gevonden in het Zentralblatt (zie figuur 1).

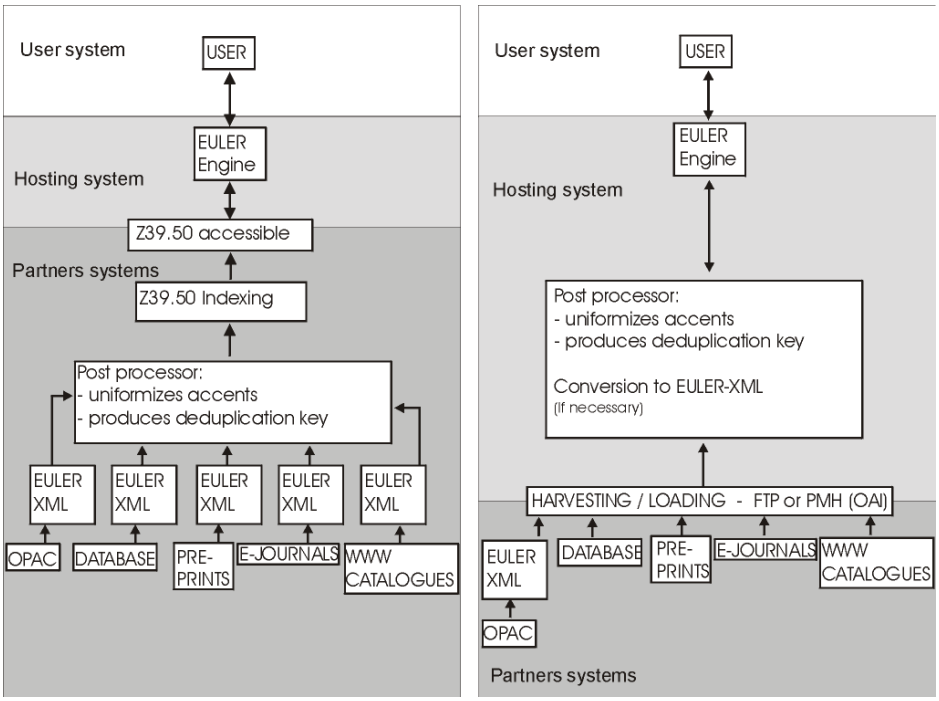
We beschikken nu al over extra informatie: een abstract met subject aanduiding en classificatiecodes. Vervolgens biedt EULER ook de mogelijkheid het artikel daadwerkelijk onder ogen te krijgen. De manier waarop dat gebeurt, hangt af van de bron waarin de publicatie gevonden is. Via het Zentralblatt (de bron waaruit in dit geval geput wordt) kan het artikel tegen betaling aangevraagd worden. Maar ook kan de gebruiker kijken of het tijdschrift in een van de deelnemende bibliotheken aanwezig is. Kopieën kunnen vervolgens aldaar aangevraagd worden.

Volledigheid

Onder een volledig systeem verstaan we een systeem waarin in de eerste plaats alle types van informatie vertegenwoordigd zijn: boeken, artikelen uit tijdschriften en proceedings, rapporten en webpagina's, en waarin vervolgens alles van wat er per type verschenen is, is opgenomen. Misschien is het een utopie volledig in deze zin te zijn. EULER streeft naar volledigheid. EULER combineert daartoe alle mogelijke bestaande bibliografische bronnen: databases met metadata en databases met fulltext versies; databases die geld kosten en databases die geen geld kosten, databases met webpagina's en



Figuur 1 Het zoekresultaat van EULER voor Shelah "small index property"



Figuur 2 Vergelijking decentraal (links) en centraal systeem (rechts)

databases met e-prints. Verschillen in structuur, retrieval system en user interface mogen daarbij geen belemmering vormen. EULER gebruikt zowel commerciële bronnen als niet commerciële bronnen. Deze unieke bundeling van commerciële en niet-commerciële informatie is wellicht EULER's grootste kracht. Om meerdere heterogene bestanden tot één geheel samen te smelten heb je in de eerste plaats eenzelfde opmaak (format) nodig. EULER gebruikte daarvoor 'common resource descriptions' die volgende de Dublin Core (DC) [10] standaard zijn opgemaakt.

Eén van de partners is het FIZ dat de data van het Zentralblatt MATH beschikbaar stelt. Door deelname van FIZ is ook het zoeken op artikelniveau gewaarborgd. Het bestand bevat meer dan 1.8 miljoen records uit meer dan 2300 seriewerken en tijdschriften. Voor abonnees van het Zentralblatt MATH wordt automatisch de volledige versie van deze database beschikbaar gesteld. Voor niet-abonnees wordt alleen een deel beschikbaar gesteld. FIZ is een non-profit organisatie maar heeft ook inkomsten nodig om hun diensten te kunnen leveren. Ook het *Jahrbuch über die Fortschritte der Mathematik* (digitaal archief van de belangrijkste wiskundige publicaties in de periode 1868–1942) is in de database opgenomen. Ook op citaties kan worden gezocht, zij het dat dat beperkt is tot de Zentralblatt data. Op dit moment is ook al een aantal belangrijke bibliotheken partner wat met name de documentleverantie ten goede komt. Verder kunnen partner worden alle soorten commerciële en niet commerciële uitgevers en boekwinkels: iedereen die betrokken is bij de distributie van wiskundige informatie.

Eenvoudig doorzoekbaar

Het formaat van de records (genoemd EULER XML) is gebaseerd op de 15 DC elementen en maakt gebruik van de DC qualifiers om elementen beter te kunnen beschrijven of om schema's te identificeren. Zo kent het element 'title' een qualifier 'alternative title', en als je een code uit het MSC schema gebruikt wordt in het element 'subject' het betreffende schema aangeroepen. EULER gebruikt 14 DC basiselementen en voegt daar 10 eigen elementen aan toe. EULER kent een zogenaamd 'ontdubbelingskenmerk', dat opgebouwd is uit een aantal van de DC elementen. Dit kenmerk zorgt ervoor dat een publicatie die door meerdere partners wordt aangeboden en dus meerdere keren in de database voorkomt slechts eenmaal in het resultatenoverzicht aan de gebruiker getoond wordt. Pas in een later stadium krijgt hij te zien waar het document allemaal te verkrijgen is. Dit kenmerk past goed in de multilinguale context die EULER nastreeft. Bovendien gebruikt EULER geen stopwoordenlijsten; elk woord wordt letterlijk genomen. De zoekresultaten worden aangeboden in een bepaalde volgorde ('ranking'). Er wordt niet on-the-fly geranked. Dat zou teveel tijd kosten. In plaats daarvan is de gehele database van te voren geranked. Ranking hangt niet meer af van de zoekvraag, maar van de inhoud van de database, op het moment dat de zoekvraag gesteld wordt. Aan de hand van welk criterium de database gesorteerd gaat worden, is nog niet besloten. Er zijn twee mogelijkheden: ofwel de database wordt chronologisch geranked ofwel er wordt, uitgaande van de database, een impact factor berekend voor auteurs of tijdschriften. In hoeverre deze 'relevance ranking' een toegevoegde waarde betekent wordt aan de hand van gebruikersonderzoek nader onderzocht. Ook behoort het nog tot de mogelijkheden dat de gebruiker de keuze wordt gelaten uit de twee ranking mechanismen. Op dit moment wordt de relevance ranking variant gebruikt.

Extra services

De belangrijkste additionele service die EULER kan bieden is documentleverantie. Deze service is kosteloos, als het gaat om gratis aangeboden elektronische publicaties, zoals rapporten die veelal gratis door instituten worden aangeboden. Het kan echter ook kosten met zich meebrengen als het document besteld wordt bij een deelnemende bibliotheek of als het document gekocht wordt bij een uitgever. Per gevonden publicatie wordt vermeld welke mogelijkheden er zijn het document in te zien. Zo kunnen er bibliotheken genoemd worden die bereid zijn kopieën te maken van artikelen of om boeken uit te lenen. Uiteraard is het wenselijk dat een bibliotheek in de buurt genoemd wordt. Het ligt voor de hand dat een wetenschapper in Rome makkelijker een boek leent bij de wiskunde bibliotheek in Florence en een wetenschapper in Amsterdam juist bij de bibliotheek van het CWI in Amsterdam.

In principe gebeurt het lenen van boeken via de bibliotheek van de instelling waar de gebruiker werkzaam is. Maar ook wordt de gebruiker de mogelijkheid geboden om zelf een publicatie te kopen, door te vermelden hoe een uitgever of boekwinkel, waar de publicatie te koop is, bereikt kan worden. De laatste faciliteit snijdt aan twee kanten: aan de ene kant vormt het een mooie, want doelgerichte, vorm van reclame voor commerciële boekverkopers, aan de andere kant kan het waardevolle informatie bevatten voor diegenen die de publicatie zelf in bezit willen krijgen. Op de advertentiemogelijkheden komen wij later terug. Een andere service is de 'content awareness service': een zoekprofiel wordt desgevraagd onthouden en om de zoveel tijd herhaald waarna de resultaten aan de aanvrager worden toegezonden. Deze service draait nog niet. Het is de bedoeling dat ze weinig geld gaat kosten.

Architectuur

Het initiële project EULER ging uit van een systeem waarbij de gedistribueerde data providers zelf hun data aanboden aan de EULER engine via het Z39.50 protocol. (Het Z39.50 protocol definieert een internationale standaard (ISO 23950) voor het communiceren tussen computers op het gebied van de information retrieval [11].) Dit model levert nogal wat werk op voor de data providers. Behalve het omzetten van hun data naar het EULER XML formaat, moet er nog werk in een 'post processing fase' gedaan worden, zoals het bepalen van het ontdubbelingskenmerk. Tenslotte dient de data provider een Z39.50 server te implementeren en te onderhouden en ze moeten zelf hun data invoeren op die server. EULER-TAKEUP daarentegen kiest voor een centraal systeem. Data providers kunnen hun data aanleveren of laten 'harvesten'. Alles wat de data provider dan hoeft te doen is het neerzetten van zijn data op een bepaalde, vooraf-gesproken plek. Het werk dat nodig is om de data vindbaar te maken voor de EULER engine wordt dan gedaan door het centrale systeem. De decentrale optie wordt wel opengelaten voor het geval er partners zijn die, om welke reden dan ook, decentraal willen blijven. Een centraal systeem is stabieler. In het gedistribueerde geval moeten meerdere systemen tegelijk werken — valt er één uit dan verkrijgt men incomplete resultaten — en er kan vertraging op treden bij het zoeken (zie figuur 2).

In zowel het centrale als het decentrale systeem wordt gebruik gemaakt van de EDBM software die beschikbaar wordt gesteld door Cellule MathDoc [12]. EDBM nam deel aan het initiële EULER project.

Geen winstoogmerk

Het consortium draait op non-profit basis. De enige inkomsten die vereist zijn, zijn die welke nodig zijn om het systeem draaiend te houden. Eindgebruikers moeten de EULER zoekmachine gratis kunnen gebruiken. Wat zijn de kosten voor een systeem als EULER? In het gekozen centrale systeem moet allereerst de centrale server onderhouden worden. Daarnaast zijn er administratieve zaken zoals de onderhandelingen met potentiële sponsors en adverteerders, als ook de ledenadministratie. Bovendien zullen er altijd technische aanpassingen verricht moeten worden. De inkomsten bestaan uit lidmaatschapsgelden van aangesloten leden, uit advertenties van commerciële partners en uit bijdragen van sponsors en wiskundige organisaties. De EMS (European Mathematical Society) sponsort EULER. Commerciële partners kunnen als data provider optreden, met als gevolg dat ook hun publicaties getoond worden. Ze kunnen ook een zoekterm sponsoren. Als een gebruiker zoekt op die term verschijnt er een reclameblokje in beeld. Deze vorm van reclame is effectief omdat de doelgroep heel nauwkeurig bepaald is. Reclameblokjes worden overigens terughoudend aangeboden.
Er kunnen altijd nieuwe leden toetreden. Het is zelfs de bedoeling dat het consortium steeds blijft groeien, om uiteindelijk zoveel mogelijk wiskundige literatuur te omvatten. Belangrijk voor de inkomsten zijn zowel lidmaatschapsgelden als inkomsten door reclame. Hoe hoger de laatstgenoemde categorie, hoe lager de eerste kan zijn. Wij denken dat EULER inderdaad voor een groot deel zal kunnen drijven op reclame inkomsten.

EULER streeft ernaar de voordelen van bestaande hulpmiddelen te combineren in één systeem. Iedereen kan kennis nemen van de nu al omvangrijke inhoud. En die inhoud zal groeien. Hoe meer partners er deelnemen, hoe volledi-ger EULER zal zijn. Ook de documentleverantie zal verbeteren als het netwerk van deelnemende bibliotheken groter en fijner wordt. Het is onze hoop en verwachting dat EULER uit zal groeien tot de informatieleverancier bij uitstek voor de wiskunde.

Referenties

1 http://www.emis.de/projects/EULER
2 http://www.ams.org/msc/
3 http://www.isinet.com/isi/
4 http://www.ams.org/mathscinet/search
5 http://www.emis.de/ZMATH/
6 http://www.isinet.com/isi/products/citation/wos/index.html
7 http://www.google.com/
8 http://www.altavista.com/
9 http://www.scirus.com/
10 http://dublincore.org/
11 http://www.niso.org/standards/resources/Z3950_Resources.html
12 http://www-mathdoc.ujf-grenoble.fr/