data & research

Jaargang 1 nummer 3 / 1 december 2006

Kwartaalblad over data en onderzoek in de alfa- en gammawetenschappen

Pagina 3 • TNO blij met remote access • Als een van de eersten onder de gebruikers van CBS-cijfers maakte TNO gebruik van de nieuwe *remote access* faciliteit, met een vijftal projecten. Tot tevredenheid.

Pagina 4 • Initiatieven rond webarchivering • Webpagina's moeten bewaard blijven, vinden steeds meer mensen. Een update van lopende initiatieven.

Pagina 5 • W.F.Hermans digitaal compleet • ind 2005 verscheen het eerste van vierentwintig delen waarin tot 2016 de volledige werken van schrijver W.F. Hermans worden verzameld. Bij deze wetenschappelijke editie speelde automatische tekstvergelijking een belangrijke rol.

EN VERDER

Kans voor Europees dataproject

De Europese erkenning van het DARIAH-voorstel voor een internationale datainfrastructuur leidt tot 'goede vooruitzichten' voor het daadwerkelijk realiseren ervan. Dat zegt DANS-medewerkster Ellen Willemse nu het plan is opgenomen in de *European Roadmap for Research Infrastructures*.

De Digital Research Infrastructure for the Arts and Humanities (DARIAH) moet onderzoekers en onderzoekinstellingen in de alfa- en gammawetenschappen op Europese schaal ondersteuning bieden op het gebied van digitalisering en hergebruik van wetenschappelijke data. Resultaat van het project, waarvan de aanloopkosten tien miljoen euro zouden bedragen en de jaarlijkse kosten vier miljoen, zou een grootschalige elektronische infrastructuur moeten zijn waarvan de alfa- en gammawetenschappen en de cultureel-erfgoedsector kunnen profiteren. Advies, voorzieningen, opslagcapaciteit en daardoor onderzoekresultaten en erfgoed worden makkelijker bereikbaar dan in de huidige, sterk versnipperde situatie.

Initiatiefnemers van DARIAH zijn behalve DANS drie andere instituten die in hun eigen land belangrijk zijn op het gebied van archivering en digitalisering: het Franse Centre National de la Recherche Scientifique (CNRS), de Duitse Max Planck Gesellschaft (MPG) en de Arts and Humanities Data Service AHDS in het Verenigd Koninkrijk. De basis van de samenwerking moet echter zo snel mogelijk worden verbreed door andere instellingen met een nationaal netwerk op te nemen.

FIGUUR: EDITA-KNAW

Infrastructuur volgens de informatiebrochure

De opname in de Roadmap is een belangrijke ruggesteun voor het voorstel. Willemse: 'Daarmee wordt het belang van DARIAH erkend door een Europees lichaam dat speciaal bedoeld is om een communautaire strategie voor infrastructuren uit te stippelen. Het European Strategy Forum on Research Infrastructures heeft als taak de Europese Commissie op dit gebied te adviseren'. In de Roadmap staan 35 'baanbrekende' projecten op het gebied van research infrastructuur. 'De volgende stap is nu,' aldus Willemse, 'dat de nationale gedelegeerden van ESFRI in hun land de onderzoekorganisaties gaan opzoeken om hen aan te sporen de Roadmap-projecten te steunen. De Nederlandse gedelegeerden klopten al vrij snel na het ESFRI-besluit bij ons aan voor informatie en dat zal in de andere Europese landen niet anders zijn'. Vanuit verschillende landen is overigens ook positief gereageerd door data-instituten en verwante organisaties. Vertegenwoordigers uit onder meer Slovenië, Hongarije, Italië, Cyprus, Roemenië en Duitsland zullen vrijwel zeker aanschuiven bij een door DANS te organiseren strategieoverleg in december.

Een tweede weg waarlangs middelen vrij kunnen komen is het Zevende Kaderprogramma van de Europese Unie. Dit programma van ruim vijftig miljard euro moet vanaf 1 januari 2007 voor een periode van zeven jaar een nieuwe impuls geven aan de z.g. Lissabonstrategie voor het Europese onderzoek. Binnen het Kaderprogramma wordt een bedrag van 135 miljoen al volgend jaar vrijgemaakt voor projecten uit de Roadmap. (MdG)

Nieuw offensief NWO voor extra onderzoekgeld

Het nieuwe kabinet moet alsnog tegemoet komen aan de wens van NWO om meer middelen vrij te maken voor toponderzoek in Nederland. Dat kan uit het reguliere budget of met behulp van geld uit het Fonds Economische Structuurversterking (FES). Dat fonds wordt gevoed uit aardgasbaten en is bedoeld voor het stimuleren van kennis, innovatie en onderwijs. In de toedeling van de middelen zouden organisaties als NWO en SenterNovem dan wel een structurele rol moeten krijgen. Dit in tegenstelling tot de huidige meer incidentele gang van zaken tot nu toe.

Dat standpunt heeft NWO bekend gemaakt bij het uitreiken van de Spinozaprijs op woensdag 29 november. De organisatie is van plan brede bekendheid te geven aan haar visie maar vooral om die onder de aandacht te brengen van de informateur. Daarmee geeft ze een nieuwe impuls aan de strategische visie die eerder dit jaar werd gepresenteerd.

Het tweede kabinet-Balkenende heeft geen standpunt meer willen innemen over de NWO-visie, die op 22 mei aan minister Van der Hoeven van OCW werd aangeboden met een dringend beroep om extra te investeren in de wetenschap. 433 miljoen euro per jaar zou nodig zijn voor extra steun aan excellente onderzoekers, bundeling van krachten en betere terugkoppeling naar de samenleving. Met name onder het hoofd 'bundeling van krachten' zou ook ruimte moeten zijn voor digitalisering en data-archivering, zo lichtte NWO-voorlichtster Dominique de Vet toe in *e-data&research*. Nadat de minister in verschillende stadia positief had gereageerd op de NWO-nota, kwam er op 22 september een brief aan de Kamer waarin duidelijk werd gemaakt dat het kabinet toch niet wilde ingaan op de budgettaire wensen van NWO in verband met 'de veranderde politieke situatie'.

Het volgende kabinet moet besluiten, aldus de kabinetsreactie. Het zelfde lot trof de voorstellen van de commissie Dynamisering die door de minister was ingesteld om de effectiviteit en efficiency van het onderzoekbestel te evalueren.

(vervolg op pagina 3)

Onderzoek naar meganetwerk GRID

DANS (Data Archiving and Networked Services) en het Amsterdamse fysica-instituut NIKHEF gaan samen de mogelijkheden onderzoeken om het meganetwerk GRID te gebruiken voor de mens- en maatschappijwetenschappen.

Het wereldwijde netwerk GRID is ontstaan in de natuurwetenschappen, waar vaak met zeer grote hoeveelheden data gewerkt moet worden. Het is bedoeld om reken- en opslagcapaciteit vrij te maken voor grootschalige experimenten. NIKHEF (Nationaal Instituut voor Kernfysica en Hoge Energie Fysica) is de beheerder van het Nederlandse deel van GRID. Het instituut kreeg van het Innovatieplatform een subsidie toegewezen van 25 miljoen euro voor een uitbreiding van de bestaande infrastructuur, onder meer voor zijn deelname aan de Europese deeltjesversneller CERN. Bij het plan daarvoor was inbegrepen dat ook andere wetenschappen gebruik kunnen maken van de GRID-capaciteit. Volgens DANS-medewerker Laurents Sesink kunnen de sociale wetenschappen en humaniora baat hebben bij zo'n faciliteit omdat zij zelf vaak niet over voldoende grootschalige reken- en opslagruimte beschikken.

Voor opslag wordt GRID al aangesproken, bijvoorbeeld in een project van de Rijksdienst voor Archeologie, Cultuurlandschap en Monumentenzorg (RACM) om opgravingstekeningen digitaal op te bergen (zie ook p.6). 'Maar in het rekenwerk ligt de echte uitdaging. Daar kunnen veel nieuwe mogelijkheden voor innovatief interdisciplinair onderzoek ontstaan. Die zijn we op het ogenblik samen aan het onderzoeken'.(MdG)

'Launch' voor Virtual Knowledge Studio

Nadat de Virtual Knowledge Studio al enige tijd had gedraaid was het donderdag 11 oktober tijd voor een officiële 'Launch' in het Amsterdamse Trippenhuis. Dat gebeurde onder leiding van science quizmaster Bart Peeters met een inleiding van programmaleider Paul Wouters, een forum en een *Research Gallery*. Wouters noemde in zijn speech de studio 'a critique of the dominant model of e-science'. We moeten misschien voorzichtig zijn met blind en grootschalig digitaliseren van cultureel erfgoed en dataverzamelingen, opperde hij. 'Elke keer moeten we ons afvragen waarvoor, voor wie, (...) wat halen we naar de voorgrond en wat krijgt in zulke activiteiten juist geen accent'. De officiële opening werd verricht door KNAW-president Frits van Oostrom. Er werken momenteel elf mensen bij de VKS als (gast-)onderzoeker of fellow. (zie ook pagina 3)

AGENDA

December 4-6
Amsterdam – KIT
2nd International conference on e-science and grid computing
The conference brings together developers and users of e-Science applications and enabling IT technologies from leading international and interdisciplinary research communities. Results of the latest research and product/tool developments will be presented, and related activities around the world highlighted. www.escience-meeting.org/eScience2006/

8 december, vanaf 14.00 uur
Gent – Koninklijke Academie voor Nederlandse Taal- en Letterkunde
Studiedag Digitale Beeldcollecties
Vereniging voor Geschiedenis en Informatica (VGI), Vlaams Centrum voor Volkscultuur (VCV), Koninklijke Academie voor Nederlandse Taal en Letterkunde (KANTL), Centrum voor Tekstediti e en Bronnenstudie (CTB). www.vgi-online.org

12 en 13 december
Rotterdam – Doelen
De digitaal erfgoedconferentie 2006
Digitaal Erfgoed Nederland organiseert deze derde conferentie over trends en technieken. Centraal staat de digitale dienstverlening van erfgoedinstellingen. De conferentie is er voor musea, archieven, bibliotheken, monumentenzorg en archeologische instellingen.www.den.nl

17 januari vanaf 12.15 uur
Tilburg – Universiteit Tilburg
DANS Symposium: Van twee kanten, moeilijkheden en mogelijkheden van multi-actor data In de sociale wetenschappen worden steeds meer data verzameld bij aan elkaar gerelateerde individuen. Op dit symposium voor wetenschappelijk onderzoekers, aio's en studenten gaan de presentaties over inhoudelijke dataverzameling. www.dans.knaw.nl/nl/dans_symposia/1_2007/

14 February, 2007
Manchester – University of Manchester, Cathie Marsh Centre for Census & Survey Research
Computerised qualitative analysis This workshop covers the computerised annotation and coding of qualitative data. We use Atlas TI qualitative coding software (and NVIVO and NUDIST demo software). Your existing knowledge of qualitative interpretation techniques is integrated with an awareness of the possibilities for computerised manipulation and annotation of data.
www.ccsr.ac.uk/courses/external/2006-2007/index.html

26 February - 27 March 27, 2007
Cologne, Germany
Spring seminar: Topics in advanced categorical data analysis Central Archive for Empirical Social Research. A training course for social scientists interested in advanced techniques of data analysis and in the application of these techniques to data. Participants must have a sound basic knowledge of statistics as well as experience in the handling of PCs and of working with SPSS www.gesis.org/Veranstaltungen/ZA/FS/index.htm

28 February, 2007
University of York
GIS in historical research, a free one day workshop Geographical Information Systems (GIS) are becoming increasingly used by scholars with an interest in the geographies of the past. To date take-up has been hampered by a lack of understanding of what GIS is and what it has to offer. This free workshop, sponsored by the ESRC Research Seminars Competition and hosted by the Arts and Humanities Data Service (AHDS), will provide a basic introduction to GIS both as an approach to academic study and as a technology. http://ahds.ac.uk/history/hgis/seminar-york.htm

8 maart 2007 (voorlopige datum)
Den Haag – Ministerie van VROM
DANS Workshop: Wegwijs in WoON: voorkom omrijden De data van WoOn 2006, de voortzetting van de Woningbehoefteonderzoeken, zijn onlangs beschikbaar gekomen voor secundaire analyse. Voor gebruikers van deze data organiseert DANS samen met VROM een gratis workshop. Informatie (w.o. def. datum): www.dans.knaw.nl/nl/dans_symposia/

Speeddaten om subsidiegelden

In de laatste begroting van het kabinet heeft de filmsector 173 miljoen euro gekregen voor het conserveren en digitaliseren van het audiovisuele erfgoed. Zeker een verstandige uitgave. Toch zullen de beheerders van andere erfgoedcollecties zich hebben afgevraagd: hebben wij niet ook belangrijk materiaal dat moet worden gedigitaliseerd?

Het verwerven van geld stond centraal op de workshop *Subsidies voor Digitaal Erfgoed*, op 26 september jl. in Utrecht. Organisator was DEN, Digitaal Erfgoed Nederland. Erfgoedinstellingen konden luisteren naar mensen met geld, goede raad en/of ervaring met digitalisering.

Fondsen werven is een specialisme met eigen websites, vakbladen, en consultants. Maarten de Vries, één van die consultants, schatte dat gemiddeld 25% van de verkregen subsidies moet worden besteed aan de verwerving van de subsidie. De 'terugtredende' overheid heeft een grotere rol voor particuliere geldgevers tot gevolg. Aanvragers moeten in zijn visie het sociale profijt van de gevraagde investering benadrukken, bijvoorbeeld door het berekenen van een 'schaduwprijs' of een voorstel voor een 'sociale onderneming'.

ELSE LAURA RADEMAKER (DEN)

Wachten op een date bij de DEN subsidieworkshop

Ellen Fleurbaay, hoofd publieksdiensten van het Amsterdamse gemeentearchief, liet een aantal projecten zien waarvoor succesvol geld was geworven, zowel bij vermogensfondsen als bij bedrijven. Bedrijven geven het liefst geld aan een voor het publiek aantrekkelijke site als de *Schatkamer van Amsterdam*, die de schatten van het gemeentearchief toont. Marjan Scharloo, directeur van het Teylers Museum, vertelde over een aantal educatieve projecten en de financiering daarvan. Voor (middel-) grote projecten wordt vaak een beroep op vermogensfondsen gedaan.

Van de overheid heeft de erfgoedwereld op het vlak van digitalisering weinig te verwachten. Kees Somer (OCW) lichtte de subsidieregeling 'digitaliseren met beleid' toe. Daarbij gaat het om kleine bedragen, te besteden aan informatieplannen en soortgelijke zaken. 'Innovatie' is voor deze regeling een sleutelwoord dat moeilijk precies te beschrijven bleek. Voor basisdigitalisering, hoe noodzakelijk ook, geeft de overheid niet thuis.

De workshop werd afgesloten met een sessie *speeddaten*. Subsidievragers stonden in de rij voor korte gesprekjes met subsidieverstrekkers. Dit licht chaotische programmaonderdeel (*'O, stond jij ook in de rij voor het VSB fonds?'; 'In welke rij sta jij?'; 'Ik sta eigenlijk in deze twee rijen tegelijk'*) besloot een informatieve ochtend. (PB)

www.den.nl/docs/200606231111050/

Tekstontsluiting op het Grid

Te midden van de glooiende heuvels van Frankenland kwamen 5 tot 7 oktober in Würzburg de leden van TextGrid bijeen. TextGrid is een onderdeel van het Duitse D-grid, een project dat streeft naar een infrastructuur voor het delen van hardware, software en services. Daarbinnen streeft het naar een *workbench* voor tekstbezorgers en tekstwetenschappers, bestaande uit hulpmiddelen voor analyseren, indexeren, editeren, annoteren en publiceren van teksten. Die hulpmiddelen zijn nieuw of gebaseerd op eerdere programmatuur, maar zullen altijd moeten passen binnen een architectuur gericht op open standaarden en uitwisselbaarheid. In TextGrid werken bibliotheken, universiteiten en commerciële partners samen.

Specifieke aandacht ging in Würzburg uit naar het coderen van het *Wörterbuch der Deutschen Sprache* van Campe, uit 1807-1813. Zes delen met in totaal zes duizend bladzijden, waarvan het de bedoeling is de lemma's tot op detailniveau te coderen. De workshop over de codering van het woordenboek stond onder leiding van Laurent Romary, op dit gebied werkzaam binnen de ISO (International Organization for Standardization) en het TEI (Text Encoding Initiative), en tegenwoordig directeur van de Max Planck Digital Library (Berlijn). Elk woordenboek is anders, en Romary liet zien hoe het mogelijk is de standaards te respecteren en tegelijk de karakteristieken van de bron recht te doen.

Wanneer hoogwaardig gedigitaliseerd materiaal voor wetenschappers beschikbaar komt, zijn annotatiehulpmiddelen nodig. Ook daaraan werd een deel van de workshop gewijd. Een demonstratie van de Linux-variant Ubuntu besloot de workshop. Belangstellenden kunnen kennismaken met TextGrid tijdens de IEEE eScience 2006 conferentie, begin december (zie Agenda). (PB)

www.textgrid.de/

Open Access: sleutelrol voor archieven

Welke taak is weggelegd voor de Europese sociaal-wetenschappelijke data archieven om toegang en hergebruik te stimuleren? Dat was op 11 en 12 oktober in Athene de vraag in de tweedaagse workshop 'Open Access to data: anonymisation, data protection & confidentiality', georganiseerd door de Council of European Social Science Data Archives (CESSDA).

Wetten en regels over privacybescherming belemmeren de toegang tot data in de meeste Europese landen. Daarnaast krijgen onderzoekers moeilijk toegang door de versnippering van autoriserende instanties. Zo is het in Nederland niet altijd of slechts met veel moeite mogelijk om onderzoek te doen aan CBS-bestanden. Ook andere Europese landen kennen dat probleem. Sociale wetenschappers willen juist steeds vaker onderzoeksbestanden aan elkaar koppelen waardoor innovatief onderzoek mogelijk wordt en nieuwe vragen beantwoord kunnen worden. Dat vraagt om een nieuwe, deels gemeenschappelijke, Europese data-infrastructuur. Een gezamenlijke strategie van de Europese sociaal-wetenschappelijke data archieven om de toegankelijkheid van dergelijke data te vergroten, is daarom volgens CESSDA belangrijk want zijn zij de belangrijkste nationale schakels om de juridische, organisatorische en technische drempels te verlagen. CESSDA streeft daarom naar een Europese infrastructuur waarin, met inachtneming van nationale verschillen, data beter toegankelijk worden en het koppelen van data wordt gefaciliteerd. De blauwdruk voor deze infrastructuur is door het Europese Strategie Forum voor Onderzoeks Infrastructuren (ESFRI) opgenomen in de Europese *Roadmap* voor onderzoeksinfrastructuren (zie ook pagina 1). (Laurents Sesink, Heleen van Luijn)

ECDL: goed eten maar slechte verhalen

De 10de 'European Conference on Digital Libraries' vond dit jaar plaats in Alicante van 17 tot 22 september. De term *Digital Libraries* kan ruim worden genomen: er komen ook veel onderzoekers en vertegenwoordigers van archieven op deze conferentie. De organisatie is op zijn Zuid-Europees: goed verzorgde uitstapjes, zeer smakelijk eten maar problemen met de verstaanbaarheid van de sprekers. De akoestiek is slecht en de geluidsinstallatie is daar niet op ingesteld. De meeste sprekers van niet-West-Europese of -Noord-Amerikaanse origine hebben wel een goed verhaal maar een slechte intonatie, uitspraak en presentatie: de opbouw is vaag, de toon is monotoon en men is te weinig op het publiek gericht. Je zou dit in 2006 niet meer verwachten.

ECDL is een peer-reviewed congres, met een acceptatiegraad van 28%. De lezingen worden in de serie *Research and Advanced Technology for Digital Libraries* door Springer uitgegeven. Een van de beste bijdragen kwam van open access pionier *Carl Lagoze*. Als elke Amerikaan kan hij zijn verhaal goed verkopen. Hij is de afgelopen jaren enkele keren de mist in gegaan met voorspellingen. Maar ook dit weet hij subliem te verkopen: het hoort bij het pionierschap, zegt hij. Waar herken je de pionier aan? Aan het aantal pijlen in zijn rug. Ook *Michael Keller*, bibliothecaris van Stanford University en 'uitgever' van het repository Highwire, hield een interessante speech over de status van het Google Book Search project en de hierbij horende nieuwe diensten.

Aansluitend op de conferentie waren er workshops. In de workshop over webarchivering (IWAW) was het niveau van de bijdragen vergeleken met ECDL hoog: verstaanbaar en goed gepresenteerd. De tools voor webarchivering, kort samengevat: Heritrix (als crawler), NutchWax (als zoekmachine, een Web Archive eXtension op Nutch) en New Generation Wayback Machine (als presentatiemodule). (Henk Harmsen)

www.ecdl2006.org/

Teleblik wint VGI Innovatieprijs

Het project Teleblik, waarin historische radio en televisie voor het onderwijs beschikbaar wordt gemaakt, heeft de VGI Innovatieprijs 2006 gewonnen. Dat werd op 22 september in Amsterdam bekend gemaakt.

De prijs werd op 22 september toegekend aan de meest vernieuwende ict-toepassing van het jaar. Een 'deskundige jury' koos het project tijdens een speciale studiedag van de VGI uit een drietal genomineerden onder de 25 inzendingen. Naast Teleblik waren dat het project Volkstellingen en E-thesis. 'Een belangrijke innovatie op het gebied van publieksbereik', aldus het juryrapport (www.vgi-online.org) dat in Teleblik 'een zeer professioneel gemaakt product' ziet. Teleblik is een samenwerkingsproject van het Nederlands Instituut voor Beeld en Geluid, Teleac/NOT en Kennisnet. Sinds april van dit jaar zijn tienduizend uren radio en televisie uit de archieven van de publieke omroepen – vanaf het begin van de twintigste eeuw tot nu – via internet toegankelijk gemaakt voor het primair en voortgezet onderwijs in Nederland. Docenten kunnen de audiovisuele bronnen van Teleblik gebruiken voor hun lesmateriaal. De formule slaat aan: een halfjaar na de invoering staan er al meer dan 330.000 Teleblikgebruikers geregistreerd.

Teleblik (www.teleblik.nl) is op verschillende manieren toepasbaar. De bronnen (die in verband met auteursrecht alleen online te bekijken zijn) kunnen worden gebruikt door docenten bij oriëntatieopdrachten of bij instructie en uitleg, maar ook door leerlingen in werkstukken, presentaties of webquests. De gebruiker kan bronnen eenvoudig selecteren, bekijken en bewaren. Bovendien kunnen docenten en leerlingen zelf fragmenten uit programma's knippen. Hierdoor biedt Teleblik naast een receptieve ook een actieve, constructieve en creatieve manier van leren. Tom Brink, marketingcoordinator van Teleblik, is blij met de *innovation award*. 'De prijs betekent dat de mogelijkheden van het project ook worden onderkend door partijen buiten het onderwijsveld. Het is een bevestiging dat een van de uitgangspunten van Teleblik breed wordt gedragen: namelijk dat het belangrijk is Nederlands audiovisueel erfgoed publiek beschikbaar te maken.' (Tom Brink)

Tom Brink, marketingcoördinator van Teleblik neemt de prijs in ontvangst van VGI-voorzitter Yola de Lusenet

Balkenende II schoof NWO-claim door

(vervolg van pagina 1)

Hein Meijers, directeur communicatie van NWO, licht desgevraagd toe dat NWO hoop put uit de *Kennis-investeringsagenda 2006-2016* die begin november werd gepubliceerd. Die agenda is opgesteld door een werkgroep van het Innovatieplatform onder voorzitterschap van premier Balkenende. Meijers: 'Als het CDA weer in de regering komt dan zal het deze agenda, die ook nog eens door 22 landelijke koepelorganisaties wordt aanbevolen, toch moeilijk naast zich neer kunnen leggen'. De agenda pleit voor een structurele verhoging van de overheidsinvesteringen voor onderzoek en onderwijs met zes miljard euro. Eerder had de Raad van Economische Adviseurs, een adviesraad van de Tweede Kamer, al gepleit voor extra investeringen van 10 tot 15 miljard euro per jaar voor het hele kennissysteem. NWO zal er volgens Meijers bij de informateur op aandringen in 2007 'in elk geval' 250 miljoen euro structureel uit te trekken voor toponderzoek en in de jaren daarna 433 miljoen, zoals in de strategienota werd gevraagd. (MdG)

Erasmus gaat VKS-subcentrum hosten

De Erasmusuniversiteit richt samen met de KNAW een eigen Virtual Knowledge Studio op. Dat maakte ze bekend bij de opening van het academisch jaar. De Erasmus Studio moet nieuwe vormen van onderzoek en kennis mogelijk helpen maken in de sociale en geesteswetenschappen. Het wordt een 'subcentrum' van de reeds in Amsterdam gevestigde VKS. De universiteit heeft voor de komende vier jaar bijna 2 miljoen euro beschikbaar gesteld. Er staan zes onderzoekprojecten op de rol met onderzoekers uit verschillende disciplines.

MESS-panel voor wetenschappelijk onderzoek van start

Op 27 oktober is het project MESS officieel gelanceerd met een symposium in de thuisstad Tilburg. MESS (een geavanceerde faciliteit voor Metingen en Experimenten in de Social Sciences) bouwt voort op het reeds bestaande CentERpanel van het Tilburgse instituut CentERdata. Dat internetpanel omvatte twee duizend huishoudens. MESS, dat met een NWO-subsidie van bijna 13 miljoen euro van de grond wordt geholpen, zal er vijfduizend hebben. Het streven naar representativiteit, door met een random steekproef te werken en faciliteiten te verlenen aan moeilijk bereikbare groepen, verenigt beide panels. CentERdata wil veel aandacht besteden aan de duurzaamheid van de deelname aan het panel, onder meer door vergoedingen te geven. Daarvoor is een belangrijk deel van het budget gereserveerd. Het is de bedoeling dat wetenschappelijke onderzoekers kosteloos gebruik kunnen maken van het panel dat na een pretest eind 2006 in 2007 moet gaan werken.

www.uvt.nl/centerdata/nl/mess/

TNO tevreden over eerste remote access faciliteit

Een van de eerste Remote Access werkplekken van het CBS is onlangs in gebruik genomen door de divisie Arbeid van de Nederlandse Organisatie voor Toegepast natuurwetenschappelijk onderzoek (TNO). De ervaringen zijn positief.

Externe onderzoekers kunnen voor onderzoeks- en beleidsdoeleinden toegang krijgen tot de microdata van het Centraal Bureau voor de Statistiek (CBS). Vanwege de geheimhoudingsplicht is die toegang beperkt tot enkele goed controleerbare wegen. Er kan *On Site* worden gewerkt bij het CBS, via *Remote Execution* of via de nieuwe *Remote Access* faciliteit. TNO koos in de analysefase van diverse projecten voor de laatste, meest innovatieve mogelijkheid. 'Er bestaat al sinds de jaren tachtig een prima relatie met het CBS. De samenwerking in het kader van de nationale enquête arbeidsomstandigheden is daar een prima voorbeeld van' licht projectleider Van den Bossche toe. 'In dit project voert TNO de enquête uit in opdracht van het ministerie van Sociale Zaken en Werkgelegenheid, terwijl het CBS zorgt voor het steekproefkader en de verrijking van de data. De NEA data verschijnen ook in tabelvorm op CBS StatLine'.

Op de keuze om met Remote Access te gaan werken was een aantal overwegingen van invloed. Van den Bossche: 'TNO wil graag voorop lopen bij dit soort innovaties. Daarnaast werkten we al regelmatig On Site bij het CBS'. Naast besparing op de reistijd zijn er andere voordelen. 'Remote Access geeft onze onderzoekers de mogelijkheid om direct intern overleg te voeren over de analyse-resultaten. Ook is het veel eenvoudiger onze eigen methodoloog mee te laten denken en zijn we niet meer afhankelijk van de openingstijden van de On Site werkplekken bij het CBS'. Na een snelle inventarisatie bleken er zeker vijf TNO projecten voor Remote Access in aanmerking te komen.

De ervaringen zijn tot nu toe positief. Eind oktober werd de eerste werkplek geïnstalleerd, daarna kon er meteen worden gewerkt aan een onderzoek dat al On Site liep. De overige projecten zijn in de loop van november van start gegaan. Van den Bossche; 'Na de divisie Arbeid gaan op termijn wellicht ook andere divisies met Remote Access werken'. Overigens is TNO niet onbekend met dezemanier van werken. 'Iedere TNO-medewerker kan via een beveiligde Remote Desktop verbinding waar ook ter wereld inloggen op het TNO netwerk, en dan over alle informatie beschikken zonder dat er sporen van bestanden op de gebruikte pc achter blijven'. Uiteraard is de CBS Remote Access faciliteit alleen vanuit de beveiligde omgeving in Hoofddorp benaderbaar. (Luuk Schreven)

Interactieve tijdbalken

Hoe visualiseer je, van minuut tot minuut, de gebeurtenissen rond de dood van John F. Kennedy? Of, op een andere tijdschaal, het ontstaan en uitsterven van de verschillende soorten dinosaurussen? Binnen het SIMILE project aan het Massachusetts Institute of Technology (MIT) is daarvoor de applicatie Timeline ontwikkeld.

Timeline kan, op basis van een XML-bestand, een combinatie van een aantal tijdbalken tonen. De weergave maakt deel uit van een webpagina. Met de muis kan in de weergave gesynchroniseerd vooruit en achteruit worden gebladerd. Hierbij een voorbeeld van ontwikkelingen op het vlak van 'campus conflict resolution'. De bovenste balk geeft de details, de onderste balk een chronologisch overzicht. De icoontjes geven het type van de gebeurtenis (een symposium, rapport, onderzoek, etc).

Proefschriftensite gaat Europees

De Nationale Proefschriftensite, die half september door DARENet is gelanceerd, heeft ervoor gezorgd dat het bezoek aan de DARE-site meer dan verdubbeld is. Het succes heeft moederorganisatie SURF geïnspireerd om een project te beginnen voor een internationale proefschriftensite.

Het bezoek ligt op circa twee duizend per dag; dat is twee en een half keer de gemiddelde score van 750 die DARENet daarvoor haalde.

De groei in het bezoek is deels te danken aan de extra publiciteit die het project opleverde, zegt projectmedewerker drs. Maurice Vanderfeesten van DARENet. 'Maar het aanbod is ook sterk verbeterd. De universiteiten hebben keihard gewerkt om ervoor te zorgen dat nieuwe proefschriften hoe dan ook online komen, en om oudere proefschriften te digitaliseren'. Op veel universiteiten is nu een digitale versie van het proefschrift een voorwaarde bij de promotie. Half november waren er ruim 12 duizend dissertaties beschikbaar, waarvan de meeste van recente datum. Jaarlijks verschijnen er zo'n 2,5 duizend. In 2005 kwam 60% van alle proefschriften beschikbaar via de nationale site. DARENet hoopt dat dat percentage opgevoerd kan worden naar 90.

Aan een Europese proefschriftensite, *European doctoral e-thesis*, wordt momenteel gewerkt met zusterorganisaties uit Duitsland, Denemarken, Zweden en het Verenigd Koninkrijk. 'In januari komen we in Nederland bijeen om te proberen ter plaatse iets te bouwen waarvan we dan in de daarop volgende maanden kunnen leren', aldus Vanderfeesten. 'De bedoeling is dat het resultaat in juli beschikbaar is'. (MdG)

PETER BOOT

Beweging rond webarchivering

Webpagina's moeten bewaard blijven voor later onderzoek, vinden steeds meer mensen. *e-data&research* rapporteert over drie hoopgevende initiatieven op dat gebied.

In ruim een decennium is het World Wide Web uitgegroeid tot een medium dat elk aspect van onze levens beïnvloedt. Het web is informatiemedium, publicatieomgeving, winkel, manier van leven en huiskamer. Webpagina's zijn daarom een belangrijke bron voor: historici, letterkundigen, politicologen, en eigenlijk iedereen die zich bezighoudt met menselijk gedrag. Het is belangrijk dat webpagina's bewaard blijven voor latere onderzoekers. Er zijn op dat gebied in elk geval drie hoopgevende initiatieven: van de Koninklijke Bibliotheek (KB), Virtual Knowledge Studio (VKS) en het European Archive.

Het European Archive (EA) presenteerde zichzelf op het symposium 'Avoiding the Digital Memory Loss' op 27 september in Amsterdam. Het EA is een zusterinstelling van het (Amerikaanse) Internet Archive. De oprichter daarvan, Brewster Kahle, sprak op het symposium over zijn visioen van 'Universal access to human knowledge'. Het Internet Archive archiveert niet alleen webpagina's, maar ook boeken, films en muziek. De *Wayback Machine* van het Internet Archive kan laten zien hoe een website er in het verleden uitzag.

EUROPEAN ARCHIVE

Het archief van het European Archive bij xs4all

Het EA zal een eigen archief creëren, maar ziet zich ook als een katalysator voor ontwikkelingen elders. De directeur van het EA, Julien Masanès, zegt tegen *e-data&research* zowel een *competence center* als een archief te willen zijn. Voor het eigen webarchief zal de oppervlakte van het web worden doorzocht, zonder specifieke voorkeuren; samen met anderen zullen diepgaande deelarchieven worden gemaakt. Op het moment is het EA gevestigd in Nederland, onder meer vanwege het begrip voor internetkwesties. Als bewijs van dat begrip kan gelden dat web provider *xs4all* de technische behuizing voor het archief levert. De Koninklijke Bibliotheek is, vertelt Marcel Ras, aan een inhaalrace begonnen (*e-data&research* juni 2006). Ras is projectleider webarchivering bij de KB. De KB heeft veel tijd besteed aan het inrichten van een e-depot voor elektronische publicaties. In de toekomst zal het depot ook gebruikt worden voor het archiveren van internetsites – in eerste instantie selectief: zo'n duizend tot drieduizend sites in vooral de overheidssector. Later zullen ook volledige 'crawls' van het Nederlandse deel van het net worden gedaan.

De KB wil zorgvuldig omgaan met de juridische aspecten hiervan. Volgens een onderzoek van de Leidse universiteit maakt de Nederlandse wet het lastig om websites van derden te archiveren en toegankelijk te maken, bijvoorbeeld op grond van auteursrechten en bescherming van persoonsgegevens. 'Wat het Internet Archive doet is in Nederland strikt genomen misschien illegaal', bevestigt Ras. 'Maar als particuliere instelling kunnen zij natuurlijk anders opereren dan een overheidsinstelling zoals wij zijn'. De KB overweegt contracten af te sluiten met de te archiveren instellingen, of misschien een *opt-out* optie te gaan hanteren. Een andere mogelijkheid is de gearchiveerde data toegankelijk te maken binnen het KB-gebouw.

Een ander probleem bij het archiveren van websites zijn de vele bestandsformaten, voor afbeeldingen: Flash, pdf, Officebestanden en zo meer. Pdf is voor het e-depot goed onderzocht, maar voor andere bestandstypen moet nog veel gebeuren. Daarbij werkt de KB samen met andere instellingen. Ras ziet sowieso graag een nationaal samenwerkingsverband voor webarchivering tot stand komen. Sommige instellingen, zoals het Documentatiecentrum Nederlandse Politieke Partijen, archiveren al delen van het Nederlandse web. Ras werkt ook graag samen met de Virtual Knowledge Studio (VKS), als spreekbuis van de onderzoekersgebruikers. De VKS is een *e-research* onderzoeksgroep van de KNAW. Webarchivering is een VKS-speerpunt. De VKS is, uit hoofde van haar missie, geïnteresseerd in science studies, en heeft ervaring opgedaan met het archiveren van wetenschappelijke websites. Paul Wouters, hoofd van VKS, meldt aan *e-data&research* te willen werken aan de opzet van een breder toepasbare webarchiveringsservice. Onderzoekers zouden dan gebruik kunnen maken van een interface dat op specifieke onderzoeksvragen en onderzoeksstijl is aangepast. Wouters wil daarbij samenwerken met de KB, met DANS, en met het European Archive. Bij de workshop die de VKS in september organiseerde zijn verschillende bruikbare hulpmiddelen besproken.

Wat opvalt is dat alle partijen het belang van samenwerking benadrukken. Ook dat stemt hoopvol. Het gaat uiteindelijk om een klus die het vermogen van elke organisatie afzonderlijk te boven gaat.

www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering.html
www.europarchive.org/
www.virtualknowledgestudio.nl/

Beperkingen van vrijwillige onderzoekpanels

MARTIJN DE GROOT

'Vraag je eerst af wat je wil weten'

'Als je met een panel van vrijwilligers werkt, kijk dan of zo'n panel wel de goede weg is om je onderzoeksvragen te beantwoorden'. Dat zegt Ineke Stoop, hoofd van de afdeling Informatievoorziening en Automatisering van het Sociaal-Cultureel Planbureau.

Stoop geeft daarmee de belangrijkste uitkomst weer van het symposium *Access panels en Online Onderzoek* over internetpanels, op 16 september op initiatief van DANS gehouden in Amsterdam. Het moment van de bijeenkomst was goed getimed. Een maand eerder waren de resultaten gepresenteerd van het Nederlands Online Panel Vergelijkingsonderzoek (NOPVO) dat, georganiseerd vanuit de marktsector, tot kritische conclusies kwam over de kwaliteit van onderzoek met internetpanels. Ineke Stoop zelf leverde op het symposium een *keynote* af waarin ze benadrukte dat er veel verschillende soorten marktonderzoek, en daarbinnen panelonderzoek bestaan - ieder met hun eigen voor- en nadelen, ook wat betreft de betrouwbaarheid van de uitkomsten en het mogelijke effect van de non-respons. 'Voor de kritische conclusies van het NOPVO en van het hele symposium is niet de webbasis van de panels beslissend,' zegt ze dan ook terugkijkend, 'maar het vrijwillige karakter. Je kan ook panels organiseren met *face-to-face* interviews of schriftelijke enquêtes. Maar met een steekproef van vrijwilligers krijg je altijd een afwijking doordat het om een specifieke groep gaat. En daarvoor kan je niet controleren omdat je niet precies weet in welk opzicht die groep afwijkt'.

Dat was inderdaad een van de belangrijkste conclusies van het NOPVO. In marktonderzoek wordt veel belang gehecht aan de respons, waarmee dan het percentage beantwoorde vragenlijsten wordt bedoeld binnen het panel. De respons heeft echter wel invloed op de efficiency van het onderzoek maar niet op de uitkomsten, concludeerde NOPVO. Die worden sterk gekleurd door de selectie die voorafgaand aan het panelonderzoek heeft plaatsgevonden: wie neemt er wel en niet deel aan het panel. 'Het grootste deel van de populatie zit niet in de panels', aldus de NOPVO-presentatie. 'En die geven die àndere antwoorden'. Wie er wel in zitten zijn bovendien vaak mensen die aan meerdere panels deelnemen, niet zelden vanwege de interessante beloning. 'Ons vermoeden van een flinke overlap werd bevestigd door het onderzoek,' zegt achteraf onderzoeker Robert van Ossenbruggen (ProCression), die de NOPVO-presentatie in Amsterdam mede voor zijn rekening nam. Hij voerde ook het onderzoek uit, samen met Ted Vonk (Onderzoekpaleis) en Pieter Willems (Millward Brown).

Lastige conclusies, zo wordt algemeen ingezien in de marktsector, waar uit kostenoogpunt veel met de z.g. *volunteer panels* wordt gewerkt. Al maakt het natuurlijk verschil of je 'onderzoek doet naar toetjes en wasmiddelen dan wel naar technische producten zoals digitale camera's', relativeert Van Ossenbruggen. Maar het is ook een zelfkritische dus dappere conclusie, benadrukt Stoop die naar eigen zeggen 'naar adem zat te happen' toen ze de presentatie van het NOPVO een maand eerder bijwoonde.

Zijn er ook oplossingen? NOPVO komt tot een soortgelijke conclusie als Stoop: 'Vraag je af of je dat wat je wil weten nauwkeurig genoeg kan meten met volunteer panels'. En: 'Gebruik online panels vooral voor trendmetingen en *concept testing*'. Van Ossenbruggen pleit desgevraagd ook voor beter panelmanagement: 'Er moet een standaard komen. Daar wordt internationaal al aan gewerkt, maar onze studie laat weer zien dat het echt nodig is'.

Stoop wijst nog eens op het vrijwillige karakter als oorzaak van de problemen: 'Als je een *nette steekproef* trekt kan je daarmee via internet hele betrouwbare resultaten bereiken – ook in een panel'. Aan zo'n opzet wordt volgens de methodologe door sommige onderzoekers ook hard gewerkt. De steekproef voor CentERdata van de Universiteit Tilburg bijvoorbeeld, dat ook op het symposium werd gepresenteerd als een 'voor academische onderzoekers vrij toegankelijk platform' dat drijft op een panel van vijfduizend huishoudens. Die wordt 'op traditionele wijze' getrokken samen met het CBS, met extra aandacht voor moeilijke doelgroepen zoals ouderen en allochtonen die bijvoorbeeld apparatuur in bruikleen krijgen en toegang tot het internet. Een hele vooruitgang ten opzichte van de *volunteer* opzet, aldus Stoop. Maar toch: 'Als je ouderen aan de computer gaat zetten wéét je dat dat *een bepaald soort* ouderen is. En zijn allochtonen thuis op internet, dan zijn het hoog opgeleide allochtonen'.

www.onderzoekpaleis.nl/NOPVO presentaties.htm
www.dans.knaw.nl/nl/presentaties_access_panles/
ww.scp.nl/english/publications/books/9037702155.shtml

Automatische tekstvergelijking maakt wetenschappelijke editie mogelijk

PETER KEGEL EN BERT VAN ELSACKER

Hermans blijvend leesbaar

Ruim een jaar geleden verscheen het eerste deel van de Volledige Werken van Willem Frederik Hermans, met daarin de romans *Conserve* (1947) en *De tranen der acacia's* (1949). Dat was de officiële start van een grootschalig editieproject, waarin vierentwintig verzamelbanden verschijnen van zo'n 800 pagina's, tweemaal per jaar tot in 2016.

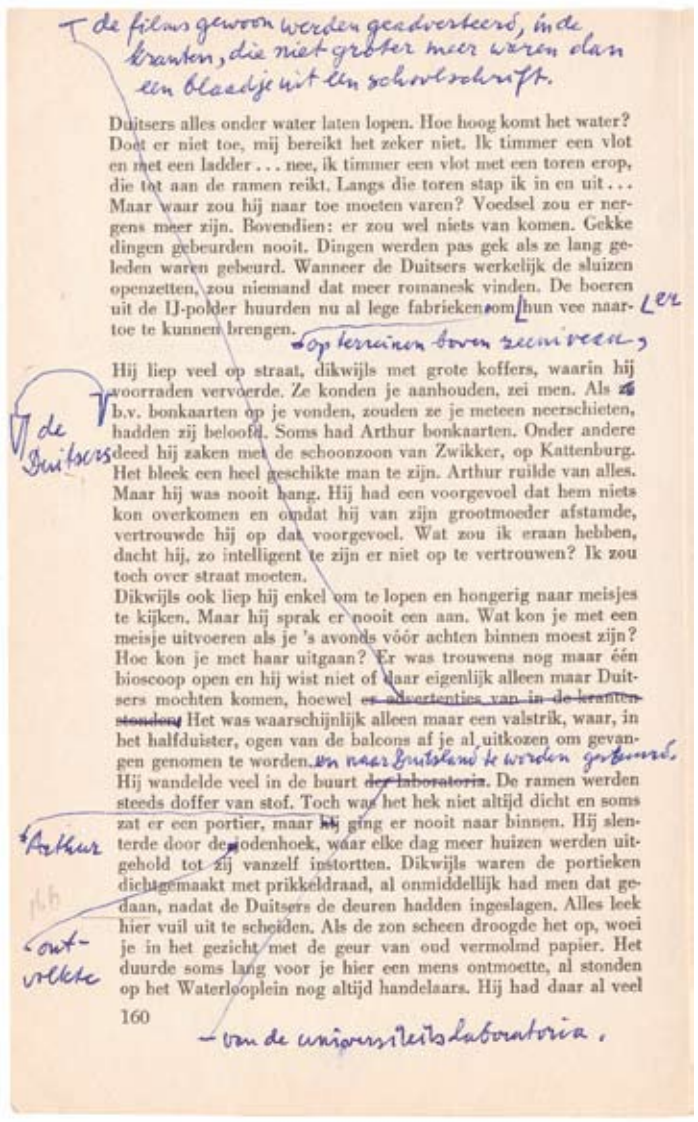
Inmiddels zijn drie delen gepubliceerd: afgelopen voorjaar deel 7, met de verhalenbundels *Moedwil en misverstand* (1948), *Paranoia* (1953) en *Een landingspoging op Newfoundland en andere verhalen* (1957); deze maand kwam daar deel 12 bij, dat *Boze Brieven van Bijkaart* (1977) en *Houten leeuwen en leeuwen van goud* (1979) bevat.

De wetenschappelijke editie is een samenwerkingsproject van het Huygens Instituut – verantwoordelijk voor de feitelijke bezorging van de tekst – het Willem Frederik Hermans Instituut en Hermans' vaste uitgever De Bezige Bij. De editie bevat de definitieve teksten van Hermans' werk: uitgangspunt is steeds de laatste tekstversie die Hermans zelf nog onder ogen kreeg en voor heruitgave goedkeurde.

Omvangrijke wijzigingen

Hermans was een auteur die ook lang na eerste publicatie nog aan zijn teksten bleef schaven. Zelf wilde hij het liefst dat zijn lezers alleen de laatste druk tot hun beschikking hadden: 'Ik zou willen dat alle oude drukken van boeken die in verbeterde vorm herdrukt zijn, als bij toverslag tot stof uiteenvielen, ook al gaat het maar om een komma', schreef hij in het voorwoord bij de bibliografie met zijn verspreide publicaties *Schrijven is verbluffen*. Dat het niet alleen om komma's ging maar vaak om omvangrijke wijzigingen, blijkt uit de tekstgeschiedenis van *De tranen der acacia's*, een roman die Hermans continu zou blijven herzien. De vierentwintigste druk van deze roman uit 1993, uitgangspunt voor de tekst van de editie, toonde uiteindelijk duizenden verschillen ten opzichte van de eerste (gedeeltelijke) publicatie, in 1946 en 1947 in het tijdschrift *Criterium*.

Bij het maken van een wetenschappelijke editie is het een eerste voorwaarde om de werkwijze van Hermans te kennen en inzicht te krijgen in de aard van de herzieningen die hij in zijn teksten aanbracht. Onderzoek naar de tekstgeschiedenis van *De tranen der acacia's* heeft uitgewezen dat het Hermans vooral ging om een blijvende leesbaarheid van zijn teksten. Om dat te bereiken bracht hij veelvuldig, maar niet altijd consequent, veranderingen aan in spelling, stijl en woordgebruik. Bovendien voorzag hij, waar hij dacht



Duitsers alles onder water laten lopen. Hoe hoog komt het water? Doet er niet toe, mij bereikt het zeker niet. Ik timmer een vlot en met een ladder . . . nee, ik timmer een vlot met een toren erop, die tot aan de ramen reikt. Langs die toren stap ik in en uit . . . Maar waar zou hij naar toe moeten varen? Voedsel zou er nergens meer zijn. Bovendien: er zou wel niets van komen. Gekke dingen gebeurden nooit. Dingen werden pas gek als ze lang geleden waren gebeurd. Wanneer de Duitsers werkelijk de sluizen openzetten, zou niemand dat meer romanesk vinden. De boeren uit de IJ-polder huurden nu al lege fabrieken om hun vee naartoe te kunnen brengen.

Hij liep veel op straat, dikwijls met grote koffers, waarin hij voorraden vervoerde. Ze konden je aanhouden, zei men. Als b.v. bonkaarten op je vonden, zouden ze je meteen neerschieten, hadden zij beloofd. Soms had Arthur bonkaarten. Onder andere deed hij zaken met de schoonzoon van Zwikker, op Kattenburg. Het bleek een heel geschikte man te zijn. Arthur ruilde van alles. Maar hij was nooit bang. Hij had een voorgevoel dat hem niets kon overkomen en omdat hij van zijn grootmoeder afstamde, vertrouwde hij op dat voorgevoel. Wat zou ik eraan hebben, dacht hij, zo intelligent te zijn er niet op te vertrouwen? Ik zou toch over straat moeten.

Dikwijls ook liep hij enkel om te lopen en hongerig naar meisjes te kijken. Maar hij sprak er nooit een aan. Wat kon je met een meisje uitvoeren als je 's avonds vóór achten binnen moest zijn? Hoe kon je met haar uitgaan? Er was trouwens nog maar één bioscoop open en hij wist niet of daar eigenlijk alleen maar Duitsers mochten komen, hoewel er advertenties van in de kranten stonden. Het was waarschijnlijk alleen maar een valstrik, waar, in het halfduister, ogen van de balcons af je al uitkozen om gevangen genomen te worden.

Hij wandelde veel in de buurt der laboratoria. De ramen werden steeds doffer van stof. Toch was het hek niet altijd dicht en soms zat er een portier, maar hij ging er nooit naar binnen. Hij slenterde door de Jodenhoek, waar elke dag meer huizen werden uitgehold tot zij vanzelf instortten. Dikwijls waren de portieken dichtgemaakt met prikkeldraad, al onmiddellijk had men dat gedaan, nadat de Duitsers de deuren hadden ingeslagen. Alles leek hier vuil uit te scheiden. Als de zon scheen droogde het op, woei je in het gezicht met de geur van oud vermolmd papier. Het duurde soms lang voor je hier een mens ontmoette, al stonden op het Waterlooplein nog altijd handelaars. Hij had daar al veel

160

Hermans' correcties bij *De tranen der acacia's*

dat dat nodig was, zijn teksten van toelichtende informatie. Ten slotte bracht hij, om de compositie van zijn teksten te versterken, herhaaldelijk meer verteltechnische wijzigingen aan.

Een op drie drukken

Het tekstkritische onderzoek voor de editie steunt op twee pijlers. Eén daarvan is het archiefonderzoek. De bezorgers van de editie hebben volledige toegang tot het archief-Hermans, dat zich in bruikleen in het Letterkundig Museum bevindt. Het archief bevat onder veel meer enkele manuscripten, typoscripten, talrijke correctie-exemplaren en drukproeven. Dankzij dit materiaal, aangevuld met de uitgebreide auteurscorrespondentie, is het mogelijk ten minste een gedeeltelijke reconstructie te maken van de totstandkoming van Hermans' publicaties, en de respectievelijke inbreng daarbij van auteur, uitgever, zetter, redacteur en/of corrector. Maar voor een verantwoorde wetenschappelijke editie is meer nodig. Een volledig beeld van alle veranderingen in Hermans' teksten, vereist een gedetailleerde bestudering van alle tekstversies met onderlinge verschillen. Bij Hermans is grofweg een op de drie drukken van zijn teksten herzien, zodat meer dan vijftig duizend gedrukte pagina's vergeleken moeten worden. Dergelijk onderzoek is ondenkbaar zonder het tweede fundament onder de editie: automatische tekstvergelijking.

Al bij de voorbereiding van de *Volledige Werken* was duidelijk dat de bezorging van de editie zonder geautomatiseerde collatie onhaalbaar zou zijn. Betrouwbare digitale bronbestanden waren daarvoor een vereiste. Om dat te realiseren heeft het Huygens Instituut samengewerkt met twee partners. Specialisten van de Koninklijke Bibliotheek verzorgden de microverfilming van het bronnenmateriaal, dat vervolgens door het Nederlands Instituut voor Wetenschappelijke Informatiediensten werd gedigitaliseerd en met OCR-software omgezet naar computerleesbare tekst in RTF (Rich Text Format). In een fase daarna zijn bij het Huygens Instituut de digitale bestanden grondig gecontroleerd aan de hand van een referentieversie, onder meer met gebruikmaking van het programma Araxis. Ten slotte zijn, om de teksten geschikt te maken voor automatische collatie, formele tekstkenmerken expliciet gecodeerd met behulp van speciaal voor dit doel ontwikkelde scripts. Na dit complexe voorbewerkingstraject zijn de teksten klaar voor automatische tekstvergelijking.

Verrijken met metainformatie

Vooral binnen de informatica wordt al sinds de jaren zeventig veel praktisch en theoretisch onderzoek naar automatische tekstvergelijking gedaan. Informatici maken hiervan bijvoorbeeld gebruik om verschillen in de opeenvolgende versies van een programmacode te analyseren en op die manier bugs op te sporen. Dit onderzoek heeft relatief snel geleid tot een aantal standaardalgoritmes en -toepassingen, waarvan het Unix/Linux-hulpmiddel 'diff' het bekendste is. Enigszins los van deze ontwikkelingen zijn er ook in de wereld van de editiewetenschap initiatieven geweest om computers in te zetten voor tekstvergelijking en voor de productie van edities. De meest bekende voorbeelden zijn COLLATE van Peter Robinson en TUSTEP van Wilhelm Ott. Een nieuw en interessant product op dit gebied is Juxta, een initiatief van het Amerikaanse NINES (Networked Interface for Nineteenth-century Electronic Scholarship)

Bij de voorbereidingen van de *Volledige Werken* is tot nu toe vooral gewerkt met COLLATE en diff-toepassingen. Om optimaal gebruik van de variantenoverzichten mogelijk te maken, worden deze via scripts omgezet naar XML-TEI gecodeerde documenten. Het basisidee van XML (Extensibile Markup Language) is om tekst niet te beschouwen als een reeks tekens, maar als een structuur, die expliciet gecodeerd wordt zodat ze computerleesbaar is. Op die manier is het mogelijk een tekst te verrijken met meta-informatie, die als een gegevensbank kan worden doorzocht en bewerkt. Voor het gebruik van XML-codes bestaan verschillende standaarden. Het Text Encoding Initiative (TEI) heeft uitgebreide richtlijnen opgesteld voor toepassingen in de humaniora, met specifieke modellen voor proza, toneel en poëzie, maar ook voor bijvoorbeeld transcripties van spraak, voor woordenboeken, en voor kritische edities.

Het eindresultaat van het digitaliseringstraject bestaat uit full-text digitale overzichten van Hermans' teksten. Deze XML-bestanden staan aan de basis van het tekstkritische onderzoek door de editeur. De verslaglegging van dat onderzoek wordt ook weer gedocumenteerd in de XML-bestanden. De editeur legt via TEI-codes zijn correcties en de verantwoording daarvan vast in de te editeren tekst. Een specifieke markering krijgen ook varianten die illustratief zijn voor bijvoorbeeld het herzienings- of productieproces van een bepaalde titel, of die om tekstinhoudelijke reden van belang zijn, net als tekstplaatsen met een bijzondere typografie en eigenaardigheden die specifiek zijn voor Hermans' werkwijze. Het oorspronkelijke variantenoverzicht groeit zo uit tot een systematische onderzoeksdocumentatie.

Juist de digitale beschikbaarheid van het onderzoeksmateriaal kan leiden tot nieuwe inzichten in de ontstaansgeschiedenis van Hermans' teksten. Zo bleek bijvoorbeeld uit zoekacties in de gecodeerde digitale gegevens dat Hermans, anders dan hij later in diverse interviews meldde, vóór publicatie van de eerste druk van *De tranen der acacia's* al uitzonderlijke veel herzieningen aanbracht ten opzichte van de eerder in *Criterium* verschenen versie.

Relatief korte periode

Doordat de XML-TEI-bestanden alleen maar bestaan uit tekst en codes, dus vrij zijn van opmaak, blijven ze ook op lange termijn computerleesbaar. De XML-data, die tevens gebruikt worden als kopij voor de editie, kunnen daardoor worden hergebruikt voor latere gedrukte uitgaven of voor digitale edities. De codering in XML-TEI schept bovendien de mogelijkheden voor nieuwe, dynamische vormen van tekstpresentatie en -analyse, die bijvoorbeeld kunnen ingaan op compositie, intertekstuele, cultuurhistorische of meer interpretatieve aspecten van de onderzochte tekst. Er is nog een laatste, belangrijk voordeel. De digitale beschikbaarheid van het onderzoeksmateriaal maakt het voor de editeur mogelijk om in een relatief korte tijd duizenden pagina's tekst nauwkeurig te onderzoeken. Daardoor komen in relatief snelle opeenvolging de teksten van Hermans in hun definitieve vorm en voorzien van een uitgebreid commentaar beschikbaar voor alle liefhebbers, docenten en studenten. Een speciale website: www.wfhermansvolledigewerken.nl bevat de wetenschappelijke verantwoording. Deze site wordt ook het platform voor digitale publicaties op basis van de XML-TEI-onderzoeksgegevens.

Volledige Werken: www.wfhermansvolledigewerken.nl
Collate: www.itsee.bham.ac.uk/software/collate/
TUSTEP: www.zdv.uni-tuebingen.de/tustep/tustep_eng.html
Juxta: www.patacriticism.org/juxta/
Diff: www.gnu.org/software/diffutils/
Araxis: www.araxis.com
TEI: www.tei-c.org

Vind uw vakvereniging

Welke vakorganisaties bestonden er in iedere bedrijfstak of regio vóór 1940? Om dat te verhelderen heeft het Internationaal Instituut voor Sociale Geschiedenis (IISG) alle organisaties waarvan een spoor te vinden was in een database laten opnemen.

Aan het eind van de negentiende en in het begin van de twintigste eeuw richtten veel arbeiders vakverenigingen, werkliedenverenigingen of verwante organisaties op. Een werkliedenvereniging richtte zich wat minder op strijd met de patroons dan de meeste vakverenigingen, maar probeerde toch de belangen van de arbeiders te behartigen - in dit geval niet de stoffelijke, maar vooral de geestelijke. Dergelijke verenigingen werden plaatselijk opgericht, door arbeiders die elkaar kenden. Vaak werden ze later de afdeling van een landelijke organisatie, die bijvoorbeeld cao's voor het hele land kon afsluiten. Het omgekeerde kon echter ook gebeuren: een landelijke organisatie besloot dat de arbeidsverhoudingen in Eindhoven er om vroegen dat de sigarenkistmakers of de kantoorbedienden zich organiseerden. De organisaties waren gevarieerd naar principieel uitgangspunt: christelijk interconfessioneel, of juist katholiek en protestant apart, socialisten van meer of minder radicaal slag, en nadrukkelijk neutrale bonden. Het organisatielandschap was dus ongemeen gevarieerd en werd pas in de loop van de twintigste eeuw geleidelijk overzichtelijker. Voor historici is het daarom moeilijk uit te zoeken welke organisaties er vóór 1940 in iedere bedrijfstak of regio bestonden. In de nieuwe database van het IISG worden tienduizenden plaatselijke organisaties en afdelingen van landelijke organisaties geïdentificeerd. Daarnaast is in kaart gebracht van welke van deze organisaties een archief bewaard is.

Christelijke Werkliedenvereniging 'De Eendracht' in Eindhoven, rond 1907.

Het project is tot stand gekomen door de financiële steun van het ministerie van Sociale Zaken en Werkgelegenheid, de vakcentrale FNV en het winstverdelingsfonds van Reaal. De resultaten zijn niet alleen interessant voor wetenschappers.

Wie haar familiegeschiedenis schrijft en wil weten van welke vakbond haar overgrootvader lid was, kan in de database ook vinden welke organisaties in aanmerking komen. Wie als vakbondslid wil weten welke organisaties er vroeger in zijn vak en woonplaats bestonden, kan dat ook in de database vinden. De database zal eind maart 2007 op het web openbaar toegankelijk worden en is dan te raadplegen via de site van het IISG. (Lex Heerma van Voss)

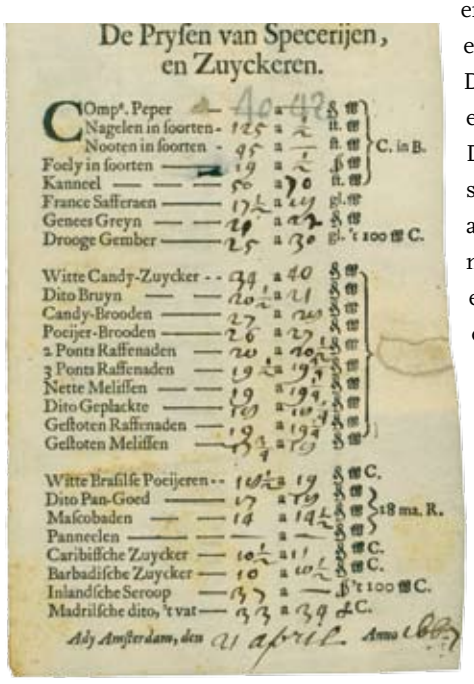
www.iisg.nl

Historische lonen en prijzen op net gezet bij het IISG

Op initiatief van prof. Jan Luiten van Zanden heeft het Internationaal Instituut voor Sociale Geschiedenis (IISG) de website *Historical prices and wages* (www.iisg.nl/hpw) opgericht.

Lonen en prijzen behoren tot de belangrijkste gegevens binnen sociaal-economisch onderzoek. Ze verschaffen informatie over de levensstandaard en het functioneren van markten; onderwerpen die een centrale plek innemen in de sociaal-economische debatten. Recentelijk is dat met name het geval binnen de discussie over het verloop van het globaliseringsproces. Om uitwisseling en hergebruik van dergelijke data te stimuleren, is het volgens het IISG van groot belang dat onderzoekers hun data digitaal beschikbaar maken op het internet.

Historical prices and wages bestaat uit vier segmenten. Allereerst zijn er de databestanden. Het project streeft ernaar bestanden aan te bieden die alle geografische en chronologische gebieden beslaan. De website biedt momenteel bestanden aan voor verschillende landen in Azië en Europa. Binnenkort zal daar bovendien een groot databestand voor China aan toegevoegd worden. Naast deze bestanden is er op de site een *applet* beschikbaar waarmee de waarde van de euro/gulden op verschillende momenten in de periode 1450-2005 vergeleken kan worden. Verder zijn er links naar andere webpagina's en congrespapers te vinden. Databestanden worden volgens een vast format aangeboden. De data zelf staan in een Excel spreadsheet. Daarnaast wordt altijd een beschrijving van bronnen, munteenheden, maten en gewichten en publicaties opgenomen. Indien mogelijk wordt ook een digitale versie van het bijbehorende paper of artikel aangeboden. Omdat de voertaal op de website en in de bestanden grotendeels Engels is, is de website goed toegankelijk voor de wereldwijde onderzoeksgemeenschap. (Christiaan van Bochove)



Prijslijst van suikers en specerijen, 21 april 1667

Unieke opgravingstekeningen duurzaam digitaal gearchiveerd

In het kader van een nieuw project worden vijftig duizend opgravingstekeningen digitaal gearchiveerd, waarvan maar één exemplaar bestaat. DANS en de Rijksdienst voor Archeologie, Cultuurlandschap en Monumentenzorg (RACM) werken hiervoor samen.

Bij een archeologische opgraving worden de vondsten en hun exacte ruimtelijke ligging nauwkeurig gedocumenteerd. Uit de ligging kunnen bijvoorbeeld ouderdom, gebruiksduur en uitgevoerde activiteiten op een nederzetting worden afgeleid. Bij een opgraving wordt dan ook veel tijd geïnvesteerd in het documenteren van de ruimtelijke context van grondsporen en vondsten. Deze wordt vastgelegd op grote A0-veldtekeningen, die op schaal alle details in de horizontale vlakken en verticale profielen laten zien. Deze tekeningen zijn unieke documenten, die aan wettelijke normen moeten voldoen.

DANS nam samen met de Universiteit Leiden en de RACM medio 2004 het initiatief tot een elektronisch depot voor de Nederlandse archeologie. De nauwe samenwerking tussen DANS en de RACM heeft geleid tot een baanbrekend nieuw project.

Het digitaliseren van die veldtekeningen is zowel vanwege het behoud als de beschikbaarheid een noodzakelijk proces geworden. De gewenste kwaliteit (hoge resolutie, ware kleuren en opslag als ongecomprimeerde TIFF-bestanden) en de fysieke omvang (veelal A0-formaat) maken de digitale bestanden zeer groot. Tot voor kort was digitale archivering bij deze grootte van de tekeningen, aantallen en kwaliteitseisen vrijwel onmogelijk.

Dankzij een nieuwe samenwerking tussen DANS, RACM en het zogenaamde BIG GRID project kunnen de Nederlandse archeologen de veldtekeningen nu wel in 24-bits kleuren scannen op een resolutie van 300 dpi en als ongecomprimeerde bestanden opslaan. Van deze hoge kwaliteit *moederscans* zijn bij de RACM kleine, gecomprimeerde afgeleiden beschikbaar, die door de eindgebruikers snel op het beeldscherm kunnen worden beoordeeld en gebruikt voor het archeologisch onderzoek. Mocht er in de toekomst een hogere kwaliteit, een ander bestandsformaat of een reservekopie noodzakelijk zijn, dan biedt het BIG GRID, via het archief van DANS, uitkomst. (Milco Wansleeben)

Studiedag beeldcollecties

Op vrijdag 8 december is er in Gent een studiedag rond digitale beeldcollecties. Organisatoren zijn de Vereniging voor Geschiedenis en Informatica (VGI), het Vlaams Centrum voor Volkscultuur (VCV), de Koninklijke Academie voor Nederlandse Taal en Letterkunde (KANTL) en het Centrum voor Teksteditie en Bronnenstudie (CTB). De dag bestaat uit twee delen. De eerste twee lezingen zijn technisch en handelen over de opvolging na het digitaliseren en de toegankelijkheid van digitale beeldcollecties. Daarna volgen drie getuigenissen uit het veld waarbij beelden op het internet worden geplaatst vanuit een verschillend perspectief, namelijk: Beeldbank Waasland, Musea Oost-Vlaanderen in Evolutie (MovE) en www.erf-goed.be. www.vgi-online.org

IASSIST Call for papers

Het thema van de 33ste jaarlijkse conferentie van de International Association for Social Science Information Service and Technology (IASSIST) is 'Building global knowledge communities with open data'. IASSIST vraagt om voorstellen voor papers, sessies, panel discussies, poster demonstraties en workshops op het brede gebied van dit onderwerp. De deadline van de *call for papers* is 16 januari 2007. Informatie: http://agiig.wordpress.com/2006/10/16/iassist-2007-call-for-papers/ of www.iassistdata.org/

NESSTAR beschikbaar voor onderzoekers

Datainstituut DANS wil het gebruik van NESSTAR uitbreiden en onderzoekers de mogelijkheid geven om zelf hun surveydata met NESSTAR op het web te publiceren. NESSTAR is speciaal bedoeld voor het publiceren van variabelen. Ook is het mogelijk om online analyses uit te voeren en data te downloaden. Geaggregeerde data kunnen met NESSTAR gepubliceerd worden in voorgedefinieerde tabellen, ondersteund door grafieken en kaarten. DANS heeft een NESSTAR server waarmee een aantal belangrijke surveys worden ontsloten. Nieuwe projecten die NESSTAR gaan gebruiken zijn o.a.: SCP-sport, COMPSOC, MESS, OSA-arbeidsaanbod en -vraagpanel. Informatie: Marion Wittenberg (DANS). http://nesstar.steinmetz-archief.nl/webview/index.jsp.

EASY on-line

Op 1 december wordt het Electronic Archive System (EASY) van DANS opengesteld. Onderzoekers kunnen dan eenvoudig data aanleveren en ophalen. DANS gebruikt de maand december om de intern ontwikkelde software te *finetunen* en het eigen archief over te brengen naar EASY. Momenteel is DANS in staat om meer dan 15 Terabyte aan data op eigen systemen op te slaan. Wanneer in voorjaar 2007 de aansluiting op het GRID in Nederland tot stand is gekomen, is er alleen nog een virtuele limiet. Begin februari 2007 zal EASY officieel in gebruik worden genomen. Volgens adjunct-directeur Harmsen van DANS gaat het om *state of the art* software, geschreven in JAVA en gebaseerd op het Eclipse framework van IBM. Het systeem maakt gebruik van internationale archiveringsstandaarden en kan data voor onbeperkte tijd veilig opslaan en beschikbaar stellen.

‘Onderzoekers moeten beste methoden en technieken toepassen’

‘De moeite die gestoken is in het verzamelen en digitaliseren van deze bronnen schept een morele verplichting bij de onderzoekers die er gebruik van maken, om bij hun analyse de beste methoden en technieken toe te passen’. Dat zegt Richard Zijdeman, een van de onderzoekers die op een speciaal symposium analyses presenteerden op basis van de digitale volkstellingen.

Op vrijdag 29 september organiseerden DANS en het CBS het symposium, getiteld *Uitgeteld & Ingevoerd: Analyse van de Nederlandse volkstellingen 1795-2001*. Aanleiding was het afronden van de digitalisering van de Nederlandse volkstellingen in het kader van het project *Life Courses in Context*.

Zijdeman studeerde sociologie aan de Universiteit van Utrecht en is momenteel aio aan het Interuniversitair Centrum voor Sociaal-Wetenschappelijke Theorievorming en Methodenontwikkeling (ICS). Zijn onderzoek richt zich op veranderingen in de sociale mobiliteit tijdens de industrialisatie. Daarbij kan hij voor het eerst gebruik maken van historische data op individueel en gemeenteniveau die een gedetailleerde analyse van regionale verschillen mogelijk maken.

FOTO P.E.A. DÜRR

Socioloog Zijdeman: Geen strijd nodig tussen kwalitatieve en kwantitatieve stromingen

Op het symposium kwam sociale mobiliteit in Zijdemans presentatie aan de orde aan de hand van de beroepsoverdacht van vader op zoon. Hij toetste verschillende theorieën door middel van een multiniveau analyse die gebruikt maakt van zowel *Life Courses* als contextuele data, onder andere afkomstig uit de gedigitaliseerde volkstellingen. De socioloog is blij dat deze data nu digitaal beschikbaar komen. Naast de gegevens afkomstig uit de Historische Steekproef Nederland (HSN) en de gedigitaliseerde Volkstellingen gebruikt hij nog verschillende andere bronnen zoals jaarverslagen van de PTT, verslagen betreffende het onderwijs en van de Dienst voor het Stoomwezen.

Richard Zijdeman vindt niet dat het verschil tussen de kwalitatieve en kwantitatieve wetenschappelijke stromingen tot strijd hoeft te leiden. Verwijzend naar het interview met prof. Engelen in de vorige aflevering van *e-data&research* benadrukt hij dat kwalitatief onderzoek goed geschikt is om mechanismen in kaart te brengen en op basis daarvan theorieën te vormen. Kwantitatief onderzoek is juist weer geschikt om op grote schaal en op een veel hoger niveau die theorieën of mechanismen te onderzoeken. Dat daarbij de menselijke kant van de zaak naar de achtergrond verdwijnt, vindt Zijdeman logisch. (Luuk Schreven)

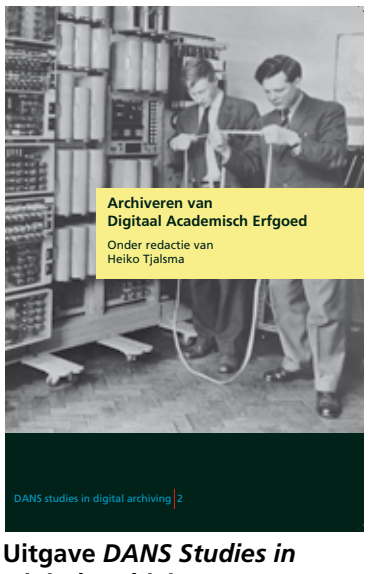
http://www.volkstellingen.nl/nl/nieuwsarchief/symposium_2006/

Archiveren van Digitaal Academisch Erfgoed

Verschillende organisaties bieden richtlijnen voor digitale archivering binnen instellingen: in Nederland bijvoorbeeld het Bureau Digitale Duurzaamheid van het Nationaal Archief, in Engeland de *Digital Preservation Coalition*. Deze richtlijnen zijn echter opgesteld om lopende zaken te archiveren. Voor retroarchivering van digitale informatie is nu door DANS de zogenaamde ADA-methode ontwikkeld, die wordt beschreven in het tweede deel van de reeks *DANS Studies in Digital Archiving*.

De ADA-methode is tot stand gekomen uit een samenwerking tussen het Meertens Instituut en het voormalige Nederlands Historisch Data Archief (NHDA, vorig jaar opgegaan in DANS). Het NHDA was wel beleidsmatig met digitale duurzaamheid bezig geweest, maar niet met de praktische uitwerking daarvan. Het Meertens Instituut bood de gelegenheid voor die uitwerking. Volgens auteur Tjalsma bevond zich bij het Meertens Instituut ‘letterlijk een kast met oude bestanden. Van daaruit ontstond op een gegeven moment de vraag om die te inventariseren en archiveren.’ Met een subsidie van SURF kon het ADA-project worden gerealiseerd.

Zowel het NHDA als het Meertens Instituut heeft concrete resultaten met het project behaald. Voor het Meertens Instituut zijn niet alleen belangrijke databestanden behouden gebleven die anders verloren waren gegaan, maar ze zijn ook gedocumenteerd en daardoor beter bruikbaar. Het NHDA heeft geleerd hoe je digitale retroarchivering aan moet pakken. Er is een standaard ADA-aanpak in zeven fases gedefinieerd die in de publicatie wordt beschreven, en die nu door DANS wordt aangeboden. DANS heeft de ADA-aanpak gebruikt in het project ‘e-depot Nederlandse Archeologie’ en onderzoekt met welke andere kandidaten een ADA-project kan worden opgestart. (Luuk Schreven)



Uitgave *DANS Studies in Digital Archiving 2*

Hollanders wel of niet op drift?

Migratie en demografische ontwikkelingen in Holland in de 19de en 20ste eeuw; dat was het onderwerp van een symposium op 6 oktober in Den Haag. Voor een groot aantal van de gepresenteerde onderzoeken was de Historische Steekproef Nederland (HSN) de belangrijkste bron.

Het gepresenteerde onderzoek riep vele vragen uit het publiek op en daarmee ontstond een levendig symposium. De vragenstellers putten regelmatig uit hun persoonlijke ervaring of kennis om beweringen van de onderzoekers bij te vallen of juist ter discussie te stellen. Zo werd de aanname dat de regel van gelijke erfdeling in Kennemerland ook voor meisjes opging door een participant stellig tegengesproken. Ook de gedachte dat de Hollanders op drift waren zoals de titel van het symposium suggereerde, werd door een aantal sprekers betwijfeld en zelfs ontkend. In zijn studie over verwantschap en migratie van Akersloters zag Jan Kok nog geen drie procent buiten Noord-Holland terechtkomen en minder dan twee procent in het verre buitenland.

Onderzoekers en reacties van het publiek lieten zien dat de levensloopbenadering niet hoeft te ontaarden in kille cijfermatige exercities. Naast statistische analyses van het materiaal werden ook talloze individuele levenslopen geschetst die tot de verbeelding spreken. De kracht van dit onderzoek zit in generalisaties op basis van individueel materiaal. Zo konden Ulbe Bosma en Kees Mandemakers aantonen dat de ‘Oost-Indië-gangers’ geen proletarische of agrarische achtergrond hadden maar dat het eerder om stedelijke, geschoolde arbeiders en zelfstandigen ging. Henk Laloli zag in de levenslopen van losse arbeiders en havenarbeiders van Amsterdam enig perspectief, omdat een deel er in slaagde zich te ontworstelen aan de krotten van de oude stad en te verhuizen naar de 19e of 20ste eeuwse wijken.

Alle onderzoekers bleken complexe databases te hebben aangelegd, waarmee de digitale historicus realiteit geworden is. (Henk Laloli)

http://www.iisg.nl/~hsn/news/hollandersopdrift.php

DONOR-project afgerond, toekomst onduidelijk

‘In de toekomst zal kennisdeling steeds belangrijker worden. Deze ontwikkeling moet ondersteund worden met ICT-tools. Er ligt nu een prototype klaar om met die kennisdeling aan de slag te gaan. Het zou ontzettend jammer zijn en kapitaalvernietiging betekenen als niemand dit oppakt’. Dat zegt Lisette Bros, projectleider van DONOR (Data Onderwijskundig Nederland Online Research). Dit project uit het SURF-programma Digital Academic REpositories (DARE), is uitgevoerd bij de Radboud Universiteit en maakt onderwijsdatabestanden die in onderzoek zijn of worden geproduceerd, online toegankelijk.

‘Uit wetenschappelijk oogpunt moet elk onderzoek repliceerbaar zijn’, aldus Bros, ‘Daarom hebben we in het kader van het DONOR-project faciliteiten ontwikkeld waarmee onderzoekers hun data kunnen documenteren en archiveren in een zogenaamd e-depot.’ De data blijven hierdoor beschikbaar voor andere onderzoekers.

Om de beschrijving van onderzoeksdata te automatiseren, is er een DONOR-tool in Java ontwikkeld. ‘Doel van de tool is om bestanden in e-depots te kunnen deponeren met uitgebreidere metadata dan de gebruikelijke bibliografische kenmerken’, zegt ICT-specialist Jaap van der Linden. De DONOR-tool is gebaseerd op het open source-systeem DSpace, een systeem dat onder andere wordt gebruikt om publicaties in repositories te zetten. Van der Linden: ‘Aan DSpace is een extra niveau toegevoegd waardoor onderzoekers ook afzonderlijke variabelen kunnen beschrijven.’

De toekomst van de tool is nog onduidelijk. De bestandsinventarisatie die daarnaast binnen het project werd uitgevoerd, is in DARENET opgenomen en diverse databestanden zijn via DANS beschikbaar. (Luuk Schreven)

www.donoronline.nl en www.darenet.nl

COLOFON

e-data&research is het kwartaalblad over data en onderzoek in de alfa- en gammawetenschappen, verschijnend onder auspiciën van DANS, het Huygensinstituut, het Internationaal Instituut voor Sociale Geschiedenis en de Vereniging voor Geschiedenis en Informatica.
e-data&research is ook een voortzetting van Historia&Informatica en van Data News – Steinmetz Archive.
Toezending kosteloos aan relaties van de stakeholders en op verzoek aan studenten in de alfa- en gammarichtingen.
Oplage: 6500.

e-data&research is online te raadplegen op www.edata.nl

Uitgever: Edita-KNAW, Postbus 19121, 1000 GC, Amsterdam

Redactieadres: Postbus 93067, 2509 AB Den Haag; Anna van Saksenlaan 51, 2593 HW Den Haag; **T** (070)3494450 **F** (070)3494451 **E** edata@dans.knaw.nl

Redactie: Peter Boot, Martijn de Groot (hoofd/eindredacteur a.i.), Marien van der Heijden, Jetske van der Schaaf, Luuk Schreven.

Aan dit nummer werkten mee: Christiaan van Bochove, Tom Brink, Peter Doorn, Tom van Dijk, Bert van Elsacker, Henk Harmsen, Lex Heerma van Voss, Peter Kegel, Henk Laloli, Heleen van Luijn, Laurents Sesink, Milco Wansleeben.

Redactiesecretariaat: Lucas Pasteuning, Jetske van der Schaaf

Pre-press, productie en vormgeving: Edita-KNAW
Druk: PlantijnCasparie, Almere
ISSN: 1872-0374

Open toegang tot data in China?

In de laatste week van oktober kwamen in Peking zo'n zeshonderd academici uit alle werelddelen bijeen op de zogenaamde CODATA-conferentie over vrije toegang tot wetenschappelijke informatie. Ik was daar één van en ik was vooral nieuwsgierig naar de beschikbaarheid van data en andere informatie achter the Great Chinese Firewall.

Ik kan me voorstellen dat China alle belang heeft bij een zo open mogelijke toegang tot wetenschappelijke data. Van westerse onderzoekers want de toegang tot Chinese data voor het Westen wordt natuurlijk in de eerste plaats beperkt door de taalbarrière. Maar als het gaat om maatschappelijke of politieke informatie is de reputatie van China bepaald niet onbesproken.

Zo blijkt uit onderzoek van Zittrain en Edelman in 2002 dat zo'n twintig duizend websites niet vanuit China bereikbaar waren maar wel vanuit de

VS: vooral websites met nieuws, politiek, religie, gezondheid, handel en entertainment waaronder sex. De sites van Amnesty International, het Centre for Anti-Communism, de Taiwan Student Club in Austria en de Voice of America hoorden daarbij. De Chinese overheid voert een politiek van blokkades, die echter vooral effectief zijn tegen gebruikers die ze niet proberen te omzeilen. De blokkeringssystemen worden ook steeds verfijnder al vergen ze veel onderhoud, waaronder handmatige fijnregeling. Aan menskracht is in China echter geen gebrek, zo bleek mij iedere ochtend als drie Chinezen op één beeldscherm controleerden of het ontbijt wel bij mijn kamerprijs was inbegrepen.

In een landenstudie van het *OpenNet Initiative* (ONI) staat dat China's Internet filtering regime is the most sophisticated effort of its kind in the world. Er zijn tal van organisaties en duizenden medewerkers op verschillende niveaus bij betrokken. Het filterregime schijnt heel fijnmazig te zijn, over de tijd steeds anders te worden ingesteld en zowel plaats te vinden op het niveau van het backbone van het Chinese netwerk als op dat van internet service providers. Er wordt ook op zoektermen gefilterd en Chinese internetcafé's zijn wettelijk verplicht om het zoekgedrag van hun klanten zestig dagen te bewaren.

Maar in hoeverre is webinformatie nu eigenlijk effectief af te schermen? Van spamfilters weten we dat ze gebrekkig werken. En hetzelfde geldt voor de filters die angstige ouders installeren om hun kinderen weg te houden van smoezelige websites: kinderen weten sneller hoe ze die kunnen omzeilen dan dat hun ouders ze kunnen installeren.

Ik kreeg nu de kans de effectiviteit zelf uit te proberen vanuit mijn hotel in Peking (alle kamers voorzien van Internet, voor omgerekend vijf euro per dag). Eerst maar eens de Nederlandse Google. In een handomdraai was allerlei vunzigheid, politiek of anderszins, op het scherm te toveren. Waarschijnlijk vindt de Chinese overheid het gedachtegoed van Geert Wilders helemaal niet gevaarlijk voor het voortbestaan van het Rijk van het Midden. Maar ook 'Taiwanese independence', Falun Gong en Dalai Lama leverden talloze hits op, al weet je bij Google natuurlijk nooit wat je *niet* vindt. Mijn experiment was niet wetenschappelijk in de zin dat elders op de aardbol iemand gelijktijdig dezelfde zoektermen intypte.

En de Chinese Google? Dat werkt immers schaamteloos mee aan de Chinese censuur? Bij de zoekmachine voor afbeeldingen valt het verschil duidelijk te zien. De Chinese Google levert slechts 173 plaatjes op voor het trefwoord 'Falun Gong', vooral van aanhangers van de sekte die moordaanslagen hebben gepleegd; de Nederlandse Google levert 45.200 afbeeldingen.

Bij het zoeken naar tekst verandert het beeld. Als je geen Chinees kunt lezen is het natuurlijk moeilijk na te gaan wat je krijgt, maar het intypen van Engelse termen geeft telkens vergelijkbare aantallen hits in de Chinese en de Nederlandse Googles. Zou het verfijnde systeem tijdens de CODATA-conferentie even op een laag pitje zijn gezet, tenminste in het hotel waar de conferentie werd gehouden? Of kon men zelfs per kamer op een knopje drukken: in 1707 zit een westerling, die hoeft niet te worden geblokkeerd? *Reporters sans frontières* meldden op 6 juni dat Google.com sinds 31 mei werd geblokkeerd in verband met de herdenking van het bloedbad op het Tiananmenplein op 4 juni, zeventien jaar geleden. Maar een paar dagen na de herdenkingsdag was deze blokkade opgeheven, blijkbaar omdat de urgentie van de censurering toen alweer was weggeëbd.

De CODATA-conferentie maakte duidelijk dat China vrije toegang tot wetenschappelijke data nastreeft. Zo ver is het nog niet voor alle soorten informatie. Het is mogelijk om de web-blokkades te omzeilen. Maar het is natuurlijk wel de vraag of de gemiddelde Chinese internetter dat aandurft. (Peter Doorn)

www.codataweb.org/06conf/
http://cyber.law.harvard.edu/filtering/china/
www.opennetinitiative.net/studies/china/
www.rsf.org/article.php3?id_article=17936

Column
Tom van Dijk

Brulapen-onderzoek


MARIE CÈCILE THIJS

'Sociaal-wetenschappelijk onderzoek met behulp van een computer gaat sneller, maar duurt langer'. Die stelling poneerde Rudy Andeweg bij het verschijnen van zijn proefschrift in 1982. Zijn these is lang waar geweest. Sinds een aantal jaar is zij evenwel genadeloos gefalsifieerd. Want online-onderzoek is een snelheidsduivel. Mooi natuurlijk: tijd is immers geld en geld is god.

Wat al héél snel gaat, moet alras natuurlijk nòg sneller. En wat steeds goedkoper is geworden, moet nòg goedkoper. Dat kan ook makkelijk, want rapporten worden niet in meer in het Nederlands geschreven, maar in Powerpoints – dat op korte termijn zal worden uitgeroepen tot de officiële onderzoekstaal. In dat idioom gelden twee kapitale stijlfiguren. Eén: hele zinnen zijn absoluut verboden. Twee: pictogrammen zijn verplicht. Zo brengt internet ons terug naar de Middeleeuwse plaatjesboeken.

Kort na het moment waarop het volk aan het fietsen kon raken, schreven dominees: 'de fiets doet kwaad'. Dat hadden ze goed gezien. Want de fiets bracht hun jongeren naar het Sodom en Gomorra van de urbaniteit. De fiets bracht het platteland evenwel ook naar de universiteit. Internet is ook een soort fiets. Iedereen kan er op stappen. Waar vroeger – voor veldwerk – nog de drempel lag van forse investeringen, is deze nagenoeg volledig weggevallen. Voor paar honderd euro is software te koop waarmee data zijn te produceren. En eenmaal geland in tabellen zien alle data er hetzelfde uit.

Een jaar of wat geleden liet een verzekeraar een 'duur' onderzoek naar een nieuw product doen. Dat product had iets te maken met de beurs. Zo'n 17% van de doelgroep bleek kooplustig. Het onderzoek was nog niet gehouden of de beurs kachelde in elkaar. Geen probleem: een slimme jongen bij onze verzekeraar heeft toen via software van 325 euro een aantal enquêtevragen op de website van het eigen bedrijf gezet. Spectaculair resultaat: een koopintentie van 50%. Na veel gesoebat is het dure onderzoek toch nog een keer herhaald. De koopintentie was geslonken tot 3%.

McKinsey&Company, Planet Internet, NRC Handelsblad, AD, MSN en FHV BBDO hebben onlangs voor de tweede keer in successie 21minuten.nl gelanceerd. In 2005 zeiden ze over zichzelf: 'het belangrijkste internetonderzoek naar de toekomst van Nederland'. En: 'wij zorgen er voor dat er naar u geluisterd wordt'. Dat klinkt nog eens bemoedigend: Fortuyn is weliswaar dood, maar dan hebben we altijd 21minuten nog. Onlangs is de tweede meting gepubliceerd in een keurig rapport, met hele zinnen, zonder Powerpoints. Via herwegingen is gecorrigeerd voor internetvertekening en 'zelfselectie onzuiverheid'. Met die laatste term hebben ze het dan over de vele miljoenen die niet op eigen initiatief 21 minuten wilden besteden aan het invullen van de vragenlijst. Dat is heel knap. Mensen die niet aan het onderzoek hebben meegedaan worden via een goocheltruc toch nog in het onderzoek gewogen. Vroeger kon dat nog niet. Wat er niet was, kon toen nog niet door weging tevoorschijn getoverd worden. Geïnteresseerd? Die handige brulaap van onze verzekeraar heeft het recept.

Tom van Dijk is directeur Beleidsonderzoek bij Intomart GfK

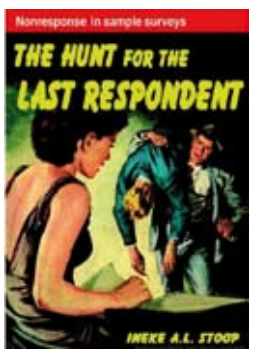
Gelezen

European Strategy Forum for Research Infrastructures (ESFRI): *European roadmap for research infrastructures – Report 2006*; ISBN 92 79 02694 1
ESFRI werd opgericht in 2002 om de Europese Commissie strategisch te adviseren over wetenschappelijke onderzoeksinfrastructuren. Aan deze eerste Europese 'roadmap' hebben zo'n duizend vooraanstaande onderzoeksexperts een bijdrage geleverd. De roadmap bevat 35 aanbevolen projecten. In de alfa- en gammawetenschappen zijn zes projecten benoemd: twee voor de alfawetenschappen, drie specifiek voor de gammawetenschappen en één dat beide gebieden bestrijkt. Online edition: http://cordis.europa.eu/esfri/home.html

Research Council United Kingdom, Economic and Social Data Service: *Guidance on data management*
Although written for Rural Economy and Land Use (RELU) research projects, this publication contains valuable information on data management in general. This guide replaces the *Guide to good practice: data management 2005*. Online edition: www.data-archive.ac.uk/news/publications.asp

K. Breedveld, A. van den Broek, J. de Haan, L. Harms, F. Huysmans en E. van Ingen: *De tijd als spiegel* (Den Haag, Sociaal en Cultureel Planbureau, 2006); ISBN 90 377 02 83x
Hoeveel uren werken we in Nederland? Wie zorgen voor het huishouden en de kinderen en hoeveel tijd besteden ze daaraan? Hoeveel vrije tijd hebben we, en wat doen we in die vrije uren? Kijken we televisie? Sporten we? En hoeveel tijd ruimen we in voor onze sociale contacten? In 2005 hielden 2200 Nederlanders gedurende een week in oktober hun activiteiten bij in een dagboekje. Sinds 1975 zijn om de vijf jaar vergelijkbare aantallen Nederlanders bereid geweest dat te doen. Dankzij hen is het mogelijk een beeld te schetsen van de tijdsbesteding van de Nederlandse bevolking over de periode 1975-2005. In De tijd als spiegel wordt de tijdsbesteding anno 2005 beschreven en vergeleken met die in 2000 en eerdere jaren. Online beschikbaar: www.scp.nl en zie ook: www.tijdbesteding.nl

Ineke A.L. Stoop: *The hunt for the last respondent, nonresponse in sample surveys* (Den Haag, Sociaal en Cultureel Planbureau, 2005); ISBN 90 377 02 155 (Ned. samenvatting)
Bij het analyseren van data of het publiceren van interessante uitkomsten wordt meestal weinig aandacht besteed aan de kwaliteit van de verzamelde gegevens. Geavanceerde analysemethoden kunnen nooit tekortkomingen in de dataverzameling goedmaken. Een hoge nonrespons kan grote gevolgen hebben voor de betrouwbaarheid van de resultaten. Daarom is investeren in de kwaliteit van de gegevens en kennis over de wijze waarop gegevens zijn verzameld van belang voor beleid en wetenschap. In Nederland is het steeds moeilijker om een hoge respons te krijgen. Dit



maakt onderzoek onder steekproeven uit de bevolking lastig en kostbaar. Een lage respons roept vragen op over de betrouwbaarheid van de uitkomsten. In dit proefschrift wordt een overzicht gegeven van de internationale literatuur over nonrespons in steekproefonderzoek. Online: www.scp.nl