e data & research

Jaargang 6 nummer 1 / juni 2011

Kwartaalblad over data en onderzoek in de alfa- en gammawetenschappen

Mogelijk gemaakt door: Data Archiving and Networked Services, CentERdata, Centraal Bureau voor de Statistiek, CLARIN-NL, Huygens ING, Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica

WILLIAM HOOGTEYLING

KNAW: Open Access ook voor data

De Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) kiest voor Open Access en vraagt van haar instituten om dat ook te doen. Dat heeft het bestuur van de Akademie dit voorjaar besloten. In een brief aan de instituten wordt gevraagd 'het beleid rondom publicaties zodanig vorm te geven dat het past bij de doelstellingen' van de KNAW, en dit nog dit jaar vast te leggen in een notitie. Daarbij kunnen ze gebruik maken van een door de moederorganisatie aangereikt 'handvat'. Om de digitale duurzaamheid door de hele organisatie heen van de grond te krijgen is een budget gereserveerd van acht ton over een periode van vier jaar te beginnen met 2011.

Opvallend is dat de KNAW expliciet spreekt over open toegang tot publicaties èn data. Vaak wordt dat onderscheid in discussies over open access niet gemaakt en blijken de gedachten van discussianten en beleidsmakers vooral uit te gaan naar publicaties zoals tijdschriftartikelen.

Voor publicaties 'volgt de KNAW het voorstel dat de (...) Nederlandse universiteiten vorig jaar hebben aangenomen', aldus de beleidsnotitie waarop het bestuursbesluit is gebaseerd. Voor onderzoeksdata is het beleid dat data (...) duurzaam opgeslagen en voor iedereen toegankelijk worden, tenzij er bijvoorbeeld wettelijke bepalingen (privacy, nationale veiligheid), contracten of andere dwingende argumenten zijn die dat belemmeren.' Uitzonderingen dienen te worden verantwoord en elk onderzoekplan moet een korte dataparagraaf bevatten die aangeeft 'hoe het hergebruik van de onderzoekdata bevorderd kan worden'.

Het budget voor Open Access wordt over vier gelijke porties per jaar verdeeld, waarbij steeds 25 duizend euro voor scholing en voorlichting wordt begroot en een zelfde bedrag voor het uitbreiden van de reeds bestaande repository voor publicaties van de KNAW. Stimuleringsfondsen voor open access publicaties en voor permanente toegankelijkheid van data krijgen ieder steeds 75 duizend euro. (MdG)

Geld voor e-projecten uit NWO Middelgroot

Verschillende onderzoekvoorstellen met een sterke e-dimensie zijn in de prijzen gevallen bij de subsidieronde van NWO die in maart plaatsvond onder de titel Investeringen NWO-middelgroot. Die ronde, speciaal bedoeld voor infrastructurele voorzieningen, ging in totaal over 12,2 miljoen euro.

Zo werd een aanvraag gehonoreerd van een groep onder leiding van Christine Arnold (zie ook p.7) voor een databank over politieke responsiviteit. De bedoeling is om een geïntegreerde en openbare databank te maken met gegevens over voorkeuren van Europese electoraten, politieke partijen en regeringen, verkiezingsuitslagen en wetgeving. Met die databank moet onderzoek worden gedaan naar de mate waarin het publiek reageert op overheidsbeleid en de mate waarin de overheid zijn politiek uitdraagt door de tijd en door beleidsgebieden en bestuurlijke niveaus heen.

Ook kwam er subsidie voor de Leidse psycholoog Bernhard Hommel, voor het opzetten van een *virtual identity*-laboratorium om het samenspel te onderzoeken tussen waarneming en actie als basis voor menselijke cognitie. In het lab worden *virtual-reality*- en *brain-imaging*-technieken gecombineerd met de online-registratie van handelingen. Volgens Hommel gaat het om een 'wereldwijd baanbrekend en uniek' platform voor onderzoekers uit diverse disciplines. Op het gebied van de sociale wetenschappen kwam er financiële steun voor een project van de Utrechtse sociologe Anne-Rigt Poortman om een databank aan te leggen met gegevens over de manier waarop ouders hun scheidjng juridisch regelen. Ook dit leidt tot een 'unieke dataverzameling'.

Sonttolregisters te raadplegen

WIKIPEDIA

Willem van de Velde (1611-1693): Slag op het Sont, 8 nov 1658; v.l.n.r. de Carolus, de Eendraght, de Mercurius, de Rotterdam, de Pelikan, de *Morgen-stjarna*, de Halve Maen, de Krona en de Brederode (detail)

Dit voorjaar is het eerste deel van de database Sonttolregisters-online online beschikbaar gekomen, met de gegevens over meer dan 125 duizend doorvaarten uit de laatste dertien jaar van de achttiende eeuw. De gegevens zijn te raadplegen op www.soundtoll.nl.

De koning van Denemarken hief van ca. 1400 tot 1857 tol op schepen die door de Sont voeren, de zeestraat tussen de Oostzee en de Noordzee. De registers zijn bewaard vanaf 1497. De helft van de schepen in het register kwam uit Nederland. In de zeventiende eeuw ging het vooral om Hollandse schepen, in de loop van de achttiende eeuw namen schepen uit Friesland die rol over. De tolboeken bevatten gegevens over 1,8 miljoen doorvaarten en vormen daardoor een belangrijke bron voor onderzoek..

De data worden in een database ingevoerd op initiatief van de Universiteit van Groningen en het Fries historisch en letterkundig centrum Tresoar. Het project duurt tot 2013.

Ook tweets in SoNaR-corpus

Ruim vierhonderd Nederlandse twitteraars hebben tot nu toe gehoor gegeven aan een oproep van taalkundigen om hun tweets beschikbaar te stellen aan de wetenschap. Die oproep werd afgelopen februari gedaan voor SoNaR, een groot corpus geschreven Nederlands.

'We hebben ook gezocht naar de accounts van bekende Nederlanders zoals politici, waarbij we er vanuit gaan dat we vrij gebruik mogen maken van de informatie die zij via Twitter verspreiden', vertelt Henk van den Heuvel, coördinator van de verzameling nieuwe media van SoNaR. 'Verder verzamelen onze Vlaamse collega's ook nog twitteraccounts in Vlaanderen.'

Naast tweets worden ook chats uit chatboxen, posts van internetfora, blogs en e-mails in het corpus opgenomen. 'Uiteindelijk zullen bijna alle tekstcategorieën voorkomen in SoNaR', legt Van den Heuvel uit. 'Zo hebben we boeken en brochures, bijvoorbeeld patiëntenfolders en voorlichtingsmateriaal. Maar ook rapporten en jaarverslagen en zelfs werkstukken en scripties van scholieren en studenten.'

SoNaR zal in december dit jaar afgerond zijn en bevat dan ruim 500 miljoen woorden geschreven Nederlands. 'De Centrale voor Taal- en Spraaktechnologie (TST-Centrale) van de Nederlandse Taalunie gaat het corpus vervolgens beheren', aldus Van den Heuvel. 'Onderzoekers en ontwikkelaars kunnen daar aankloppen om de data te gebruiken voor hun project.'

Wie deze gebruikers precies zullen zijn weten de samenstellers niet; daarom proberen ze het corpus zo compleet mogelijk te maken. De verzamelde teksten moeten een realistisch beeld geven van de manier waarop mensen in het Nederlands schrijven. Dit gaat lang niet altijd volgens de officiële regels voor grammatica en spelling. Verder worden de teksten voorzien van allerhande metadata zodat onderscheid naar bijvoorbeeld leeftijd, geslacht of woonplaats mogelijk is.

Het SoNaR-corpus wordt gefinancierd door het Nederlands-Vlaamse STEVIN-programma voor spraak- en taaltechnologie. Aan het project werken onderzoekers mee van de universiteiten van Nijmegen, Tilburg, Enschede, Utrecht , Gent en Leuven. Projectleider is Nelleke Oostdijk van de Radboud Universiteit Nijmegen. (ER)

http://lands.let.ru.nl/projects/SoNaR/

AGENDA

5-10 juni
University of Alabama – USA
Digital Preservation Management Workshop
Fostering critical thinking in a technological realm and providing the means for exercising practical and responsible stewardship of digital assets in an age of technological uncertainty, the workshop sessions are geared towards making a digital preservation program doable for any organization. Intended for managers who are responsible for digital preservation programs in libraries, archives, and other cultural institutions.
www.icpsr.umich.edu/dpm/workshops/fiveday.html

8 juni
Den Haag, Ministerie van OCW
Expertmeeting data management en de nationale Roadmap voor grootschalige onderzoeksfaciliteiten
Expertmeeting van het ministerie van OCW en DANS. Toegang verschaffen tot databases is een doelstelling bij meer dan de helft van de 25 voorgestelde grootschalige faciliteiten op de Nationale Roadmap. Daarnaast richten veel voorstellen zich op data-integratie en interoperabiliteit, standaarden, metadata en datakwaliteit. Ook het snelle transport van grote hoeveelheden data over netwerken en *high performance computing* keren regelmatig terug als vereisten. De meeting gaat na of er elementen in de datavoorziening zijn die efficiënt gedeeld of gezamenlijk aangepakt kunnen worden. De bevindingen worden aangeboden aan de commissie die aanstaand najaar de Roadmap gaat herzien. www.dans.knaw.nl

13-17 juni
University of Ottawa – Canada
Joint Conference on Digital Libraries – Bringing Together Scholars, Scholarship and Research Data
Major international forum, encompassing the many meanings of the term 'digital libraries', including new forms of information institutions and organizations; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, distributing, and accessing digital content; theoretical models of information media, including document genres and electronic publishing; and theory and practice of use of managed content in science and education. www.jcdl2011.org/

14 en 15 juni
Den Haag, DANS
PersID seminar on object identification
Various persistent object identifier solutions will be compared with regard to services, and future cooperation and convergence will be explored. This seminar will take place in The Hague and will be hosted by PersID, SURFfoundation and DANS. Attendees will be experts, representatives of the participating PID systems, national libraries, data archives and other stakeholders. www.persid.org/news.html

28-30 juni
Köln, GESIS
Workshop 'Digital preservation management: implementing short-term strategies for long-term problems'
See 5-10 June, University of Alabama.
www.gesis.org/en/events/gesis-workshops/digital-preservation-management/

4-6 juli
Leeuwarden, Fryske Akademy
ICISO 2011 – Computational Humanities
The 13th International Conference on Informatics and Semiotics in Organisations, ICISO 2011, previously named International Conference on Organisational Semiotics (ICOS), is devoted to the latest research and applications in the field. The main theme is: 'Problems and Possibilities'. The aim is to provide a focal forum for participants from various domains of information management and information systems, computational science, semiotics, humanities, and business and engineering, being researchers, practitioners, or business and industrial professionals and academics.
www.orgsem.org/2011/

1-3 Augustus
University of Oxford, Oxford
IATU Conference 2011 – Measuring and mapping activities.
The Centre for Time Use Research invites all people interested in research into people's daily activities to join us at the 2011 Conference.
http://iatur2011.timeuse.org/

'Vernieuw de wetenschap'

'Innovating Scholarship' was het motto van een brainstorm-meeting op 29 maart bij het Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS).

Organisatoren waren naast het NIAS de nieuwe e-Humanitiesgroep van de KNAW en NWO Geesteswetenschappen. Rond de zestig geesteswetenschappers verzamelden zich in Wassenaar om na te denken over de mogelijkheden om de geesteswetenschappen met een computationele aanpak te vernieuwen. Verantwoordelijk voor de dag waren Sally Wyatt, hoofd van de KNAW e-Humanitiesgroep, en Jaap van den Herik, onder andere voorzitter van de executive board van het cultureel-erfgoedprogramma CATCH.

Behalve een inhoudelijke brainstorm was ook de vorming van een 'community' doel van de dag. Er was bovendien een beleidsmatig doel: onderzoek naar mogelijkheden en draagvlak voor een nationaal programma voor de verdere ontwikkeling van e-Humanities. Zo'n programma zou op zijn beurt als doel hebben om in de komende tien jaar honderd aio-plaatsen op het terrein van de e-Humanities te financieren.

Belangrijke rol

Het ochtendprogramma werd geopend door Rick Johnson onder de uitdagende titel *A Modest Proposal for World Domination by the Netherlands in e-Humanities*. Johnson rapporteerde over het onderzoek waarbij hij automatisch draden telt in doeken van schilderijen, en uit de patronen in het doek kan afleiden welke schilderijen uit hetzelfde doek zijn gesneden: typisch voorbeeld van het soort onderzoek dat voorheen te arbeidsintensief was om uitvoerbaar te zijn. Nederland heeft, zo betoogde Johnson, de infrastructuur om een belangrijke rol te spelen in de e-Humanities, in elk geval op het terrein van de kunstgeschiedenis: centrale onderzoeksinstituten, een traditie van samenwerking, en prachtige kunstwerken. Vooruitgang hoeft niet te komen van veel geld, maar kan ook bevorderd worden door grote uitdagende dataverzamelingen te creëren, waar de informatici vanzelf op afkomen.

Complexiteit

In het vervolg gaven zes wetenschappers aan hoe ze digitale hulpmiddelen in hun zeer diverse onderzoek gebruikten: van het verschuivend gebruik van voornaamwoorden in de Bijbel tot elektronische ondersteuning van onderhandelaars. In workshops werd, ook aan de hand van het onderzoek van andere deelnemers, over de onderwerpen doorgepraat. In die discussies kwam van alles aan de orde: het nut van interdisciplinair werken (dwingt tot expliciet maken van begrippen), de vereiste hoeveelheid technische kennis bij de geesteswetenschapper, de complexiteit van geesteswetenschappelijke vragen voor de computerwetenschap (soms uiterst simpel, soms nog ver buiten de huidige mogelijkheden).

Eén van de laatste sprekers was Demetrius Waarsenburg van het ministerie van Onderwijs, Cultuur en Wetenschap, die moest toegeven dat in deze tijden de mogelijkheden voor dure initiatieven beperkt zijn. Hij schetste de contouren van het beoogde nationaal programma voor e-Humanities, waarvoor de resultaten van de dag als invoer konden dienen. (PB)

www.nias.nl/Pages/NIA/29/458.bGFuZz1FTkc.html


MICHIEL VAN DUSSELDORP


MICHIEL VAN DUSSELDORP

Boven: Rick Johnson zet zijn *modest proposal for world domination* uiteen als luisteraars op de eerste rij deeldagvoorzitter Jaap van den Herik, mede-organisator Alice Dijkstra en NIAS-directeur Aafke Hulk. Onder: Reflectie tijdens de lunch in de NIAS-tuin

Vrije documenten door open standaarden

Open standaarden en open documentformaten zouden ook in de erfgoedwereld het uitgangspunt moeten zijn, niet alleen om de duurzaamheid te waarborgen maar ook om hergebruik te stimuleren en te bevorderen. Dat was de slotconclusie van een boeiende middag met als thema 'Open Erfgoed' op woensdag 30 maart bij de Koninklijke Bibliotheek (KB) in Den Haag.

Deze derde Nederlandse editie van *Document Freedom Day*, een wereldwijd initiatief om het gebruik van open documentformaten en open standaarden te stimuleren, werd georganiseerd door Document Freedom Day, Digitaal Erfgoed Nederland (DEN) en de Vereniging voor Geschiedenis en Informatica (VGI).

Een veertigtal bezoekers hoorde de zes sprekers vanuit verschillende invalshoeken licht werpen op de toepassing van open standaarden en open formaten in de erfgoedsector. De KB krijgt als nationale bibliotheek in toenemende mate *born-digital* publicaties aangeleverd, in verschillende formaten, soms open maar ook gesloten. Bovendien worden er op grote schaal kranten, tijdschriften en boeken gedigitaliseerd, zoveel mogelijk met open standaarden en formaten. De overheid probeert middels het actieplan Nederland Open in Verbinding instellingen aan te sporen open standaarden en *open source* te gebruiken. Europeana moedigt erfgoedinstellingen aan hun digitale collecties met behulp van open standaarden zoals OAI-PMH en Dublin Core beschikbaar te stellen, zodat ze eenvoudig 'geharvest' kunnen worden en aan andere collecties kunnen worden verbonden. De Digitale Bibliografie Nederlandse Geschiedenis stelt de gedigitaliseerde publicaties zoveel mogelijk in open formaten beschikbaar, vooral ook om het de onderzoeker mogelijk te maken zelf met het materiaal aan de slag te gaan. (Edwin Klijn)

www.dbng.nl/ www.linuxmag.nl/nieuws/716-document-freedom-day-wordt-volwassen.html

Ethisch verantwoord surveyonderzoek

Onderzoekers hebben een belangrijke verantwoordelijkheid in het beschermen van de rechten en het welzijn van mensen die vrijwillig meedoen aan hun onderzoek. Maar hoe vul je die verantwoordelijkheid in? Dat was een belangrijke vraag op de lezingenmiddag over Ethiek in Survey-onderzoek het Nederlandstalig Platform voor Surveyonderzoek (NPSO) op 15 februari organiseerde.

Ieder van de vier sprekers belichtte de huidige afspraken over bescherming van respondenten vanuit de eigen invalshoek. Wim van Slooten gaf een overzicht van de afspraken die de marktonderzoeksorganisaties in Nederland hebben gemaakt. Jelke Bethlehem sprak over internationale regels, vooral voor officiële statistieken. Een andere invalshoek bracht Agnes van der Heide in met haar verhaal over de ethische aspecten bij sociaal-medisch onderzoek: welke plichten heeft een onderzoeker daarbij tegenover de deelnemers, en hoe zit het met medisch-ethische toetsing? De presentatie van Ronald Leenes over databescherming werd gevolgd door veel vragen en discussie uit de zaal: moderne onderzoeksmethoden leiden tot steeds grotere hoeveelheden data en paradata over respondenten en er worden ook steeds vaker bestanden aan elkaar of aan registerdata gekoppeld. Daarbij kunnen gegevens herleid worden tot individuen en komt de privacy van de respondenten mogelijk in het geding. En die herleidbaarheid beperkt zich niet alleen tot naam, adres en woonplaats, maar beslaat bijvoorbeeld ook geregistreerde IP-adressen. Privacy heeft dus ook een technologische kant.

De discussie, geleid door de dagvoorzitter Gerty Lensvelt-Mulders, ging op de lezingenmiddag daarnaast natuurlijk ook over juridische aspecten, waarin op dit moment een groot grijs gebied bestaat. Er werden veel vragen gesteld over hoe ver de morele verantwoordelijkheid van de onderzoeker gaat.

Op de website van het NPSO kunnen nog vragen worden gesteld en kan er verder worden gediscussieerd over het thema. (Eric Balster)

www.npso.net/node/36

e-Humanities Group van start bij Meertens

Sinds maart van dit jaar is de nieuwe e-Humanities Group van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) gevestigd in het gebouw aan de Joan Muyskenweg in Amsterdam dat ook onderdak biedt aan het Meertens Instituut.

Volgens programmaleider Sally Wyatt is dat een belangrijke stap omdat het samenbrengen en samen laten werken van onderzoekers een van de belangrijke functies is van de groep. Juist omdat de onderzoekers die zich bij de groep aansluiten, merendeels in dienst blijven bij hun eigen instituut of universiteit, spelen ontmoeting en nabijheid een belangrijke rol, licht Wyatt toe. 'Er is op allerlei plaatsen al heel veel gaande op het gebied van eHumanities. Wat wij daaraan toe kunnen voegen is een plek die synergie bevordert door goede omstandigheden en allerlei activiteiten zoals bijeenkomsten, symposia en workshops. Aan de Muyskenweg hebben we prettige werkruimtes met grote kamers die zich lenen voor samenwerking'. Het is de bedoeling om voor dat doel in elk geval wekelijkse bijeenkomsten te organiseren.

Er werken nu vijf mensen bij de eHumanties Group, die de KNAW oprichtte nadat vorig jaar werd besloten de Virtual Knowledge Studio (VKS) op te heffen. Dat kunnen er snel meer worden als de lopende competitie om financiële steun van de Akademie voor projecten in de sfeer van de computationele geesteswetenschappen resultaten gaat opleveren. Het ziet ernaar uit dat dat in de komende weken gaat gebeuren, denkt Wyatt, die zelf ook afkomstig is van de VKS en daarnaast in Maastricht een leerstoel bezet in de *Digital Cultures in Development*. Er is enige vertraging opgetreden want de *Call for Proposals* voor deze competitie ging al in het voorjaar van 2010 uit (zie *e-data* 5-1). Wyatt: 'Er is een aantal voorstellen ingediend, allemaal met een flinke omvang van zo rond de zeven ton. Elk voorstel wordt minstens door twee instituten en een universiteit gedragen. De KNAW kiest er vijf uit. Als die projecten worden toegewezen komen de mensen die ze gaan uitvoeren voor een groot deel van hun tijd bij ons werken. Dan groeit vanzelf het niveau van de activiteiten maar ook het synergetisch effect.'

Het is een 'ingewikkelde constructie', geeft Wyatt toe, en ook een die nieuw is voor de KNAW: een groep die zelf weinig mensen in dienst heeft maar als centrum voor samenwerking en inspiratie dient voor onderzoekers die elders op de loonlijst staan. 'Maar als je innoveert neem je altijd risico'.

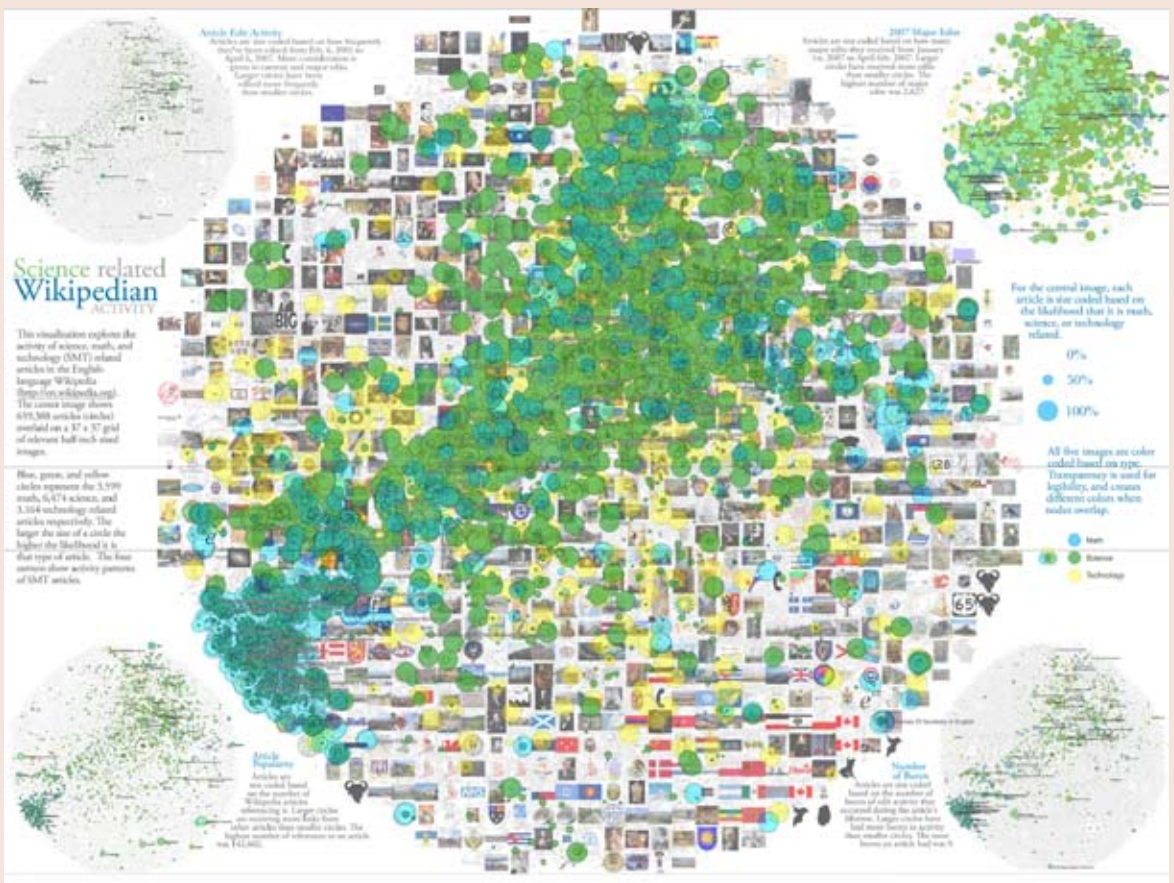
MICHIEL VAN DUSSELDORP

Programmaleider Sally Wyatt

De eHumanities Group moet ook een soort erfopvolger worden van het project Alfalab dat binnenkort afloopt, aldus Wyatt. 'Er zijn al een paar projecten in het kader van Alfalab afgerond. Andere komen in de komende maanden zo ver.Op 29 september komt er een grote bijeenkomst om te laten zien wat er is bereikt. Eén van de problemen met digitalisering in de geesteswetenschappen is dat er een aantal jaren wordt geïnvesteerd in mooie diensten of gereedschappen, en dat die daarna in verval raken als de projecten zijn afgesloten. De e-Humanities Group moet dat voorkomen, vooral door in te spelen op de behoeften van de geesteswetenschappers zelf.' (MdG)

http://ehumanities.nl/

Wetenschappelijke activiteit op Wikipedia



Börner, Katy. *Atlas of Science: Visualizing What We Know*. (2010). The MIT Press. Pg 45.

Wikipedia staat bekend als de vrije encyclopedie met een hoog betrouwbaarheidsgehalte. Dat er ook veel wetenschappelijke activiteit is op Wikipedia is veel minder bekend. In het boek *Atlas of Science* van Katy Börner staat deze afbeelding die de (Engelstalige) wetenschappelijke activiteit binnen Wikipedia in beeld brengt. Getoond worden 659.388 artikelen op een achtergrond van gekoppelde relevante afbeeldingen. Elke cirkel is een artikel. De kleuren blauw, groen en geel representeren respectievelijk de wiskundige (3.599 stuks), de wetenschappelijke (6.474 stuks) en technologische artikelen (3.164 stuks). Hoe groter de cirkel, des te groter de kans dat het om een wetenschappelijk artikel uit de betreffende categorie gaat. Bij een heel kleine kans wordt geen kleur toegekend en valt de cirkel bijna weg. Zie ook de rubriek Gelezen (p.8).

http://scimaps.org/atlas/part2.html (nummer 44-45 'Timeline 2006-2007' de twaalfde afbeelding)

Oude teksten bruikbaar voor nieuw onderzoek

Begin dit jaar zijn twee CLARIN-NL projecten afgerond die het mogelijk maken om historische Nederlandse teksten te gebruiken als bron voor onderzoek. Voorheen leidde dit vaak tot problemen door de vele spellingsvariaties die er in de teksten voorkomen.

TICCLops (*Text-Induced Corpus Clean-Up Online Processing System*), een project geleid door Martin Reynaert (Tilburg University), biedt een online dienst die deze spellingsvariatie automatisch corrigeert nadat de teksten zijn ingescand en herkend met behulp van OCR. Het project Adelheid, onder leiding van Hans van Halsteren (Radboud Universiteit), biedt de mogelijkheid om historische teksten automatisch te voorzien van onder meer lemma's en zinsontleding. Het richt zich op Nederlandse teksten uit de veertiende eeuw. Daarbij wordt gebruik gemaakt van hetzelfde systeem als TICCLops, dat zelf niet is gebonden aan een specifieke taal of periode. De diensten zijn beschikbaar via de CLARIN-centra.

Adelheid en TICClops zijn twee van de elf projecten die in 2010 zijn gehonoreerd binnen de eerste oproep van CLARIN-NL. Al deze projecten zijn nu afgerond of in de afrondende fase. Na de tweede oproep, in de zomer van vorig jaar, werden negen projecten gehonoreerd die op dit moment in volle gang zijn. In de loop van dit jaar zal nog een derde (open) oproep volgen. De gehonoreerde projecten dragen bij aan de ontwikkeling van de CLARIN-infrastructuur, een Europese digitale onderzoeksomgeving voor geesteswetenschappers. (ER)

Sociale wetenschap vraagt om andere manier van webarchiveren

Onderzoekers in de sociale wetenschappen vragen de webarchieven om zich niet alleen te concentreren op de inhoud van websites, maar ook informatie te gaan verzamelen over interacties op internet.

Dat bleek tijdens een conferentie van het International Internet Preservation Consortium (IIPC) op 9 mei bij de Koninklijke Bibliotheek in Den Haag. De IIPC is een internationaal samenwerkingsverband van circa veertig instellingen die websites archiveren; in de praktijk zijn dat vooral nationale bibliotheken.

Voor de jaarlijkse conferentie had de IIPC een aantal onderzoekers uitgenodigd om te discussiëren over het gebruik van de gearchiveerde websites. Paul Girard van SciencesPo, Eric Meyer van het Oxford Internet Institute, en Anne Helmond en Esther Weltevrede van de Universiteit van Amsterdam lieten zien dat voor hen juist de dynamiek van het interactieve internet onderwerp van onderzoek is. Maar daarover verzamelen de webarchieven nog maar nauwelijks gegevens. De techniek is de grootste barrière. Gewone html/http-bestanden zijn goed te archiveren, maar het interactieve web zit vol met technische functionaliteiten die webarchieven nog niet aan kunnen.

Een andere vraag van onderzoekers is om letterlijk alles wat er met een bepaald webarchief gebeurt nauwkeurig te documenteren, zodat de wetenschappers precies weten op welke data ze hun onderzoek baseren. Daarbij gaat het niet alleen om de keuzes die bij selectie zijn gemaakt, maar bijvoorbeeld ook om preserveringsacties, om gegevens over mislukte *crawls* en grote systeemstoringen. (IA)

http://tinyurl.com/3brxrac

Digitale editie Anne Frank

Het Huygens ING en de Anne Frank Stichting gaan, m.m.v. het NIOD, nieuw onderzoek doen naar het dagboek en de andere geschriften van Anne Frank. Het onderzoek richt zich op de ontwikkeling van Anne Frank als schrijfster en op de gebeurtenissen die zij beschrijft. Het zal resulteren in een Nederlands- en Engelstalige webeditie, waarin het mogelijk wordt om het schrijfproces van het dagboek op de voet te volgen. Annotaties en illustraties zullen de teksten ondersteunen. (PB)

Arbeidsvraag- en aanbodpanels

NATIONALE BEELDBANK

Begin maart zijn de *waves* van 2006 en 2008 van het Arbeidsaanbodpanel beschikbaar gekomen. Ook zijn onlangs de data van het Arbeidsvraagpanel 1991-2008 door het Sociaal en Cultureel Planbureau opengesteld voor onderzoekers.

De enquête van het Arbeidsaanbodpanel wordt sinds 1986 elke twee jaar afgenomen bij 4.500 personen in meer dan tweeduizend huishoudens. Er zijn inmiddels dertien metingen voor onderzoekers beschikbaar. Het Arbeidsvraagpanel verzamelt data over de vraag naar arbeid door organisaties zoals industriële bedrijven, handelsondernemingen onderwijsinstellingen, overheidsinstanties, zorg- en welzijnsinstellingen, etc. Beide bestanden kunnen direct en kosteloos worden gedownload uit DANS EASY. https://easy.dans.knaw.nl

Webweegschaal verbetert datakwaliteit in surveyonderzoek

ERIC BALSTER

Sinds augustus vorig jaar zijn er duizend internetweegschalen in gebruik genomen bij deelnemers aan het LISS panel, waarin huishoudens via webvragenlijsten meewerken aan allerlei wetenschappelijk onderzoek. Ze zijn bedoeld om op een objectieve manier gewicht en vetpercentage van respondenten te meten. Een van de doelen is om de huidige meetmethode van zelfrapportage te valideren. Maar er is veel meer mogelijk, aldus onderzoekers Peter Kooreman en Annette Scherpenzeel.

Gewicht is een belangrijke gezondheidsindicator. In surveyonderzoek wordt vaak aan respondenten zelf gevraagd wat hun gewicht is. Ook in het LISS panel (Langlopende Internet Studies voor de Sociale wetenschappen) was dit tot nog toe het geval. Er kleven echter de nodige nadelen aan deze zelfrapportage. 'Binnen ons internetpanel hebben we de ruimte die beschikbaar is voor innovatie aangegrepen om hier verbetering in aan te brengen', zegt Annette Scherpenzeel. En haar collega Peter Kooreman is blij: 'Door deze *state of the art technology* kunnen gewicht en vetpercentage veel preciezer en veel frequenter gemeten worden.'

WIEBE KIESTRA

Anette Scherpenzeel en Peter Kooreman: ...compleet beeld van alle factoren...

Hoe het werkt

De weegschaal die wordt gebruikt meet het gewicht en de weerstand in het lichaam, waarmee het vetpercentage bepaald kan worden. Zodra een deelnemer op de weegschaal staat wordt deze informatie via een radiosignaal doorgegeven aan de ontvanger die aangesloten is op het modem of de router van het deelnemende huishouden. De ontvanger stuurt via het internet de informatie naar de database van het LISS-panel. Daar worden de gegevens gekoppeld aan de juiste persoon in het huishouden en worden ook de Body Mass Index (BMI), het vetpercentage en het spiermassapercentage berekend. De respondent zelf kan deze gegevens meteen terugzien op zijn of haar persoonlijke LISS inlogpagina.

Waarom deze oplossing

Het vragen naar gewicht in een vragenlijst of interview brengt een aantal problemen mee. Het zou nogal belastend zijn om respondenten elke maand, week of zelfs elke dag naar hun gewicht te vragen, terwijl juist de variaties in gewicht door de tijd heen relevant zijn voor onderzoekers. Op de weegschaal hoeven mensen maar enkele seconden te gaan staan, wat eenvoudig een dagelijkse of wekelijkse routine kan worden. De computer hoeft er niet eens voor aan te staan. Daarnaast kunnen veel meetfouten optreden, bijvoorbeeld doordat de weegschalen die mensen zelf hebben allemaal verschillend zijn afgesteld, mensen niet op dezelfde manier afronden, fouten maken bij het aflezen en overtypen, of een waarde doorgeven die dichter bij hun streefgewicht ligt dan hun werkelijke gewicht. En dat terwijl correct gemeten gewicht en vetpercentage en over langere tijd gevolgd heel goede gezondheidsindicatoren zijn.

Het nut van de data

Het unieke van de studie is dat de gegevens over gewicht en vetpercentage gecombineerd kunnen worden met allerlei andere gegevens die bekend zijn over de LISS panel respondenten. De panelleden vullen elke maand online vragenlijsten in over allerlei onderwerpen, zoals hun opleiding, werk, inkomen, levensstijl, eet-en drinkgewoonten, persoonlijkheid, hobbies en sport, etc. Zo ontstaat een compleet beeld van alle factoren die op lange termijn samen kunnen hangen met gewicht, gewichtsbeheersing en gezondheid. Kooreman: 'Dit onderzoek is van groot maatschappelijk belang, omdat het ons meer inzicht kan geven in het gedrag van mensen met bijvoorbeeld overgewicht.'

De weegschaal en verder

Het project heeft vooralsnog een doorlooptijd van een jaar maar het kan, als blijkt dat heel veel onderzoekers de data willen gebruiken, verlengd worden en ook eventueel uitgebreid naar een groter aantal huishoudens. De data zullen na de zomer op de website van LISS (www.lissdata.nl) gratis ter beschikking gesteld worden aan wetenschappelijk onderzoekers. 'Hoe meer onderzoekers deze data gebruiken, hoe waardevoller het experiment is', geeft Annette Scherpenzeel aan. De onderzoekers denken ook aan uitbreiding van de studie met bloeddrukmeters die ook gegevens via internet kunnen versturen want, aldus Scherpenzeel: 'datakwaliteit staat voorop'.

CBS zoekt dialoog met *social media*

RONALD VAN DER BIE

Al bijna 120 jaar publiceert het Centraal Bureau voor de Statistiek de uitkomsten van zijn onderzoek. Een arbeidsstatistiek was in 1894 de eerste eigen publicatie, toen nog van de Centrale Commissie voor de Statistiek, de voorloopster van het CBS. Papierwerk was dat lange tijd, maar sinds 1996 ontsluit het CBS zijn data ook elektronisch, met de *online* database StatLine die gratis toegankelijk is via de website (www.cbs.nl/statline) en sinds 2010 ook via een iPhone app. Jaarlijks brengt het bureau ruim drieduizend publicaties zoals persberichten, conjunctuurberichten, boeken, elektronische publicaties, StatLinetabellen en (interactieve) visualisaties, naar buiten.

Sinds september 2009 verspreidt het CBS haar output ook via verschillende *social media*. Daarmee hoopt het bureau het bereik van zijn cijfers verder te vergroten en de zichtbaarheid van de organisatie te verbeteren, in het bijzonder onder jongeren die het meest actief zijn op de sociale netwerken. Door daar actief te zijn, zoekt het CBS de dialoog: het bureau wil informatie delen en open staan voor reacties.

De activiteiten met de *social media* passen in de online strategie van het bureau, dat op dit moment output verspreidt via de volgende kanalen:

- Twitter (@statistiekcbs en @statisticscbs): op twitter.com/statistiekcbs en twitter.com/statisticscbs (Engels) verschijnen de berichten uit de RSS-*feeds* van cbs.nl. In het weekend en op dagen dat er geen publicaties op de homepage van cbs.nl verschijnen wordt een statistisch weetje getwitterd. Vooral deze weetjes worden door followers geretweet en verder verspreid. De CBS-Infoservice houdt in de gaten of er vragen binnenkomen via Twitter en zorgt voor beantwoording. Het CBS heeft inmiddels ruim 2100 volgers op zijn Nederlandstalige Twitter-account en ruim 260 op de Engelstalige.
- YouTube (youtube.com/statistiekcbs) het bureau heeft twaalf korte filmpjes geplaatst met uitleg over statistische begrippen, statistieken en het gebruik van StatLine. Deze filmpjes zijn ook met Engelse ondertiteling geplaatst en tussen de tweehonderd en vierduizend keer bekeken.

Sinds januari dit jaar staan op de homepage van het CBS icoontjes die linken naar de Twitter- en YouTube-accounts. In maart zijn onder alle artikelen op de site zogenoemde *sharebuttons* geplaatst. Hiermee kunnen bezoekers van cbs.nl artikelen, persberichten en visualisaties met één klik onder de aandacht brengen van hun contacten op Facebook, LinkedIn, Twitter en Hyves.

Zelf onderzoek doen vanaf de eigen werkplek bij het Centrum voor Beleidsstatistiek op microdata via zogeheten *remote access* of *on site* faciliteiten, dat kan natuurlijk nog steeds.

INTERVIEW

Sociaal-geograaf en wetenschapsbestuurder Pieter Hooimeijer:

‘Registerdata toegankelijker maken, dat is typisch iets voor het ministerie van EL&I’

MARTIJN DE GROOT

Zijn fascinatie voor onderzoek bloeide destijds bijna tegelijk op met die voor grote dataverzamelingen. Nu is sociaal-geograaf Pieter Hooimeijer vooral pleitbezorger van een systeemaanpak: ‘Veel maatschappelijke problemen ontstaan niet omdat mensen foute dingen doen, maar omdat de verbindingen in het systeem niet goed functioneren.’

‘Op het gebied van registerdata liggen we in Nederland behoorlijk achter. In de Verenigde Staten is het al heel gewoon dat alle data die op kosten van de gemeenschap bij elkaar zijn gebracht, ook beschikbaar moeten zijn voor onderzoek.’ Als er iemand is die aan een pleidooi voor openheid van registerdata gezag kan toevoegen is het de Utrechtse hoogleraar Sociale Geografie en Demografie Pieter Hooimeijer. Hij koos al in de jaren zeventig voor de grote dataverzamelingen – in een periode waarin zijn medestudenten het nog heel gewoon vonden om zich op de uitwerking van een honderdtal enquêtes te storten. Daarna zette hij zich in voor de beschikbaarheid van die datasets, onder meer in het bestuur van het Wetenschappelijk Statistisch Agentschap dat in de jaren negentig de data van het Centraal Bureau voor de Statistiek (CBS) voor de wetenschap ging losweken. En hij speelde een belangrijke rol bij de oprichting van datainstituut DANS, waar hij nu nog de wetenschappelijke adviesraad leidt. Hooimeijer somt desgevraagd in hoog tempo een aantal argumenten voor zijn standpunt op dat mogelijke twijfelaars al bij voorbaat de moed in de schoenen doet zinken.

WILLIAM HOOGTEYLING

Hooimeijer: ‘Je wil alle variabelen in je analyse betrekken die er een rol in kunnen spelen’

Geen steekproef

Ten eerste is een systeemaanpak van maatschappelijke problemen – waarvan Hooimeijer groot voorstander is – niet mogelijk zonder registerdata. ‘Bestudeer je bijvoorbeeld een stedelijk systeem, dan wil je daarin alle variabelen betrekken die er een rol in kunnen spelen – dat is immers het wezen van een systeemaanpak. Dan gaat het dus niet alleen over de mensen maar ook over bijvoorbeeld woningen, de infrastructuur, kantoorlocaties, water, transport van goederen en personen. Daar zijn allemaal heel goed geregistreerde data over.’ Een steekproef nemen uit zulke data levert niet minder werk op, zoals het geval is bij enquêtes, maar eerder meer. Bovendien, aldus Hooimeijer, voldoet een steekproef niet:

‘Je bent geïnteresseerd in de terugkoppelingen tussen de verschillende onderdelen waaruit zo’n systeem bestaat: de ondergrond, de infrastructuur, het vastgoed, maar ook het gebruik dat van die ruimtelijke dimensies wordt gemaakt. Als je dan met steekproefgegevens werkt loop je steeds de aansluiting tussen die onderdelen mis.’

Daar komt nog bij: ‘Als je met registerdata werkt hoef je alleen de ontbrekende gegevens nog maar te verzamelen en kun je dubbele bevraging vermijden. En als het om schaarse gevallen gaat, bijvoorbeeld in het onderzoek naar bijwerkingen van medicijnen, dan heb je registerdata nodig omdat je anders niet aan voldoende cases komt.’ Registerdata zijn tenslotte ook nog nuttig als het erom gaat de representativiteit te beoordelen van data die je wèl steekproefsgewijs hebt verzameld.

‘Ik heb collega’s naar de Verenigde Staten zien gaan omdat daar de gegevens wel beschikbaar waren die ze hier niet konden krijgen’

Een duidelijke zaak. Daar moet beweging in komen, vindt de sociaal-geograaf die onder meer als voorzitter van de Sociaalwetenschappelijke Raad van de KNAW ook graag zelf zijn bijdrage zal leveren. ‘Er is nu politieke actie nodig, en dan is de systeemverantwoordelijke aan de beurt: dit is typisch iets voor het ministerie van Economische Zaken, Landbouw en Innovatie! Als registerdata gemakkelijker toegankelijk worden, zal dat het onderzoek en de kennis over Nederland te versterken. Ik heb collega’s naar de Verenigde Staten zien gaan omdat daar de gegevens wel beschikbaar waren die ze hier niet konden krijgen.’ Aan de andere kant zien we nu al, aldus Hooimeijer, dat ‘de betere beschikbaarheid van het Sociaal-Statistisch Bestand van het CBS leidt tot vragen, die daarvoor niet te onderzoeken waren’.

Vaak wordt de groeiende beschikbaarheid van elektronische data aangevoerd als aanleiding voor het gebruik van grote bestanden en de roep om beschikbaarheid daarvan. Hooimeijer had die aanleiding niet nodig. Hij kreeg in de eindfase van zijn studie toegang tot de gegevens van het toenmalige Woningbehoefte Onderzoek. ‘Dat was een kwestie van enorme mazzel want het CBS was in die tijd niet toeschietelijk met het beschikbaar stellen van data.

Markovketens

In dit stadium was de jonge Amsterdammer al gecharmeerd geraakt van onderzoek en van de methoden en technieken waarmee dat onderzoek kon worden gedaan. ‘De Vrije Universiteit, waar ik studeerde, had een zekere reputatie opgebouwd op het gebied van kwantitatief onderzoek en dat was precies wat mij aansprak. Ze hadden daar ook de beschikking over computers en cursussen om die te programmeren. Databasebeheer, FORTRAN, daar had je allemaal toegang toe.’ De mogelijkheid om met de informatie uit het Woningbehoefte Onderzoek te werken wekte bij Hooimeijer de fascinatie voor grote bestanden, die hem daarna niet meer los zou laten. Hij kon zich storten op de rol van *Markovketens* en verhuisketenmodellen dankzij het ‘voortreffelijke materiaal’ dat uit het latere WoONonderzoek voortkwam en deed dat ook in de jaren die volgden, met onder meer in 1988 als resultaat een dissertatie ‘in een nieuw vak: huishoudensdemografie, dat gaat over de gehele tijd *tussen* geboorte en sterven. De huwelijksdata waren in de jaren zeventig van de vorige eeuw immers wat in de versukkeling geraakt’.

Een van de mooie dingen van de WOoN-data is dat ze ook de dynamiek van het woongedrag laten zien. ‘Belangrijke vragen werden ook in retrospectief gesteld, bijvoorbeeld wanneer mensen zelfstandig waren gaan wonen, of dat ze in de voorafgaande periode waren gescheiden.’ En natuurlijk werden de meeste vragen elke vier jaar herhaald zodat je heel goed de beweging kon analyseren. ‘Daar kon je echt wat mee doen,’ aldus Hooimeijer. ‘Statistische verbanden, waar je natuurlijk in de sociale wetenschappen altijd naar zoekt, zeggen nog niet veel over hoe iets zich in de tijd ontwikkelt. Maar daarvoor kan je bij deze data wel terecht’. Er was nu ook voldoende materiaal om zogenaamde transitiemodellen te maken: dynamische simulaties van huishoudens en hun woongedrag, waarin plaats was voor zowel volgtijdelijke als ruimtelijke relaties. ‘Als twee mensen gaan samenwonen betekent dat ook dat er een gaat verhuizen, en dat er een woning vrij komt. Als je dat aan elkaar kan koppelen, betekent het dat je demografische dynamiek aan de dynamiek van de woningmarkt kan relateren, en weer terug. Dat kon allemaal met deze gegevens!’

Pieter Hooimeijer (1955) studeerde stadsgeografie en planologie aan de VU en promoveerde in 1988 in Utrecht op het proefschrift *Vergrijzing, individualisering en woningmarkt*. In 1990 werd hij hoogleraar demografie aan de UvA en in 1995 hoogleraar sociale geografie en demografie aan de Universiteit Utrecht. Hij is ook wetenschappelijk directeur van de Netherlands Graduate School of Urban and Regional Research en voorzitter van de Sociaal-Wetenschappelijke Raad van de KNAW.

Redundantie organiseren

Het WOoN-bestand is nog in volle glorie beschikbaar, en natuurlijk uitgebreid en geactualiseerd. Maar intussen heeft Hooimeijer zijn aandacht op andere dingen gericht. ‘Systeemanalyse is een tijd behoorlijk ondergeschoven geweest, waarschijnlijk omdat daarvoor – in de jaren zeventig en tachtig van de twintigste eeuw – de pretenties veel te hoog waren opgeschroefd. Maar het is een zeer bruikbare aanpak om maatschappelijke problemen mee te benaderen.

‘Er ligt een grote uitdaging voor de sociale wetenschappen om met voorstellen te komen’’

De bankencrisis is een mooi voorbeeld. Als er iets mis gaat roept iedereen dat er meer toezicht moet komen. Veel problemen komen echter niet voort uit toezicht of gebrek daaraan, maar uit het feit dat verbindingen binnen het systeem niet goed functioneren. President Obama zei het mooi na die mislukte aanslag op dat vliegtuig in Detroit: ‘De veiligheidssystemen werkten goed, maar de informatie werd onvoldoende uitgewisseld.’ Verkeerde systeemsluitingen, noemen wij dat. Met systeemanalyse kan je daar heel goed naar zoeken omdat het wezen daarvan juist is dat de grenzen van het systeem niet worden gesloten. Je mag er alles bij halen’. Het belang voor de maatschappij is enorm, benadrukt Hooimeijer, juist in een tijd waarin strenger toezicht steeds meer als panacee wordt gezien terwijl we om ons heen kunnen waarnemen dat het problemen niet voorkomt. ‘Er ligt een grote uitdaging voor de sociale wetenschappen om met voorstellen te komen waardoor je kunt laten zien hoe het misgaat, maar ook waar je *resilience*, je veerkracht is te vinden. En hoe je *redundantie* kunt organiseren – krachten die het overnemen als het niet goed gaat’.

Een advies over deze nieuwe rol van de sociale wetenschappen, *Kwetsbaarheid en veerkracht van maatschappelijke systemen* is in voorbereiding bij de Sociaal-Wetenschappelijke Raad van de KNAW.

Focus

CLARIN-NL – Common Language Resources and Technology Infrastructure in Nederland

INGE ANGEVAARE

Bestaande digitale bestanden en de *tools* om ze te gebruiken zo goed zichtbaar en toegankelijk maken dat vele onderzoekers in de geesteswetenschappen er nieuw onderzoek mee kunnen doen. Dat is de missie van het project CLARIN-NL. *e-data&research* sprak met bestuursleden Jan Odijk (Utrecht) en Arjan van Hessen (Twente) op het Programmabureau van CLARIN-NL aan de Trans in hartje Utrecht.

In de alfawetenschappen behoorden taal- en spraakwetenschappers bij de eerste groepen die de grote mogelijkheden van het digitale tijdperk ontdekten. Ze begonnen al snel met het samenstellen van digitale tekstcorpora en het ontwikkelen van gereedschappen om die te kunnen analyseren. Maar omdat ieder dat op zijn eigen manier deed, kon men elkaars data niet goed gebruiken; de benodigde interoperabiliteit ontbrak. Jan Odijk: 'Zelfs een simpel overzicht van wat er allemaal in Nederland beschikbaar is, ontbreekt.' CLARIN-NL wil dat veranderen door, zoals het officieel heet, een *e-science* infrastructuur te bouwen voor talige data en tools in de geesteswetenschappen. 'Wij richten ons daarbij op alle onderzoekers uit de geestes- en sociale wetenschappen die met talig materiaal werken, niet alleen op de taal- en spraakwetenschappen,' benadrukt Jan Odijk. 'Ook veel onderzoekers in de geestes- en sociale wetenschappen kunnen gebruik maken van de corpora en de tools om die te exploreren.' Als ze tenminste goed vindbaar zijn, goed gestructureerd zijn, en duurzaam beheerd worden. Het is de kern van het CLARIN-project om alles zo te structureren dat tools en data van verschillende herkomst goed in combinatie gebruikt kunnen worden.

Metadata spelen hierbij een cruciale rol. CLARIN-NL werkt niet met één vast metadataschema, maar met een flexibel, modulair systeem. Onderzoekers kunnen daarin zo nodig zelf componenten aanmaken als zij die nodig hebben. Daarnaast moeten de data voldoen aan algemeen geaccepteerde en binnen CLARIN ondersteunde technische standaarden.

INGE ANGEVAARE

Jan Odijk (links) en Arjan van Hessen: 'Het lukt ons steeds beter om ook minder technisch georiënteerde onderzoekgroepen bewust te maken van de mogelijkheden'

Duurzaamheid niet vanzelfsprekend

Duurzaam databeheer is nog lang niet overal ingeburgerd. Daarom schrijven de CLARIN-regels voor dat onderzoeksgegevens na afloop van projecten moeten worden ondergebracht bij één van de vijf CLARIN-centra: het Max Planck Instituut voor Psycholinguïstiek, het Meertens Instituut, het Instituut voor Nederlandse Lexicologie, het Huygens Instituut en DANS. CLARIN-NL hoopt dat dit netwerk uitgebreid zal worden en dat op korte termijn ook beheerders van enorme digitale corpora als de Koninklijke Bibliotheek, het Nationaal Archief en het Nederlands Instituut voor Beeld en Geluid data zullen gaan leveren die passen in het CLARIN-systeem.

Maar ook de studenten en onderzoekers zelf moeten bewust worden gemaakt van de mogelijkheden. Van Hessen: 'Vooral de opleiders spelen hier een sleutelrol. Door cursussen te ontwikkelen en gastcolleges te geven, maken we opleiders en studenten bewust van de mogelijkheden en leren we ze te werken met grote digitale bestanden. We streven ernaar de technieken onderdeel te laten worden van het standaardcurriculum in de geesteswetenschappen, zowel in de bachelor- als in de masterfase.'

Onderzoekers niet altijd informatici

CLARIN-NL wil ook onderzoekers helpen bij het omzetten van hun gegevens naar CLARIN-standaarden. Van Hessen: 'Je mag niet van elke onderzoeker verwachten dat hij technisch onderlegd is'. De hulpschermen naar allerlei functies binnen de CLARIN-infrastructuur zoals zoek- en browsefuncties of het exploreren of bewerken van data, moeten daarom zeer intuïtief en gebruiksvriendelijk zijn.

Het totale budget is negen miljoen euro. Odijk: 'Daar kunnen we heel wat mee bereiken, vooral op het vlak van bewustwording. We ontwikkelen in elk geval overtuigende *showcases* om aan onderzoekers te laten zien wat er allemaal mogelijk is, zodat het voor wetenschappers vanzelfsprekend wordt om de CLARIN-regels te volgen omdat je anders de boot mist.'

www.clarin.nl;
www.isocat.org

Interedition: interoperabiliteit voor duurzaamheid

Op een recente bijeenkomst van Interedition op de Ludwig Maximilians Universität in München telden samengeschoolde literair-historici en IT-onderzoekers in nauwelijks vijf minuten niet minder dan achttien projecten om software te ontwikkelen voor de transcriptie en annotatie van gedigitaliseerde teksten. In de geesteswetenschappen, waar budget en capaciteit voor IT-ontwikkeling steevast beperkt zijn, is het opmerkelijk dat digitale instrumenten met hetzelfde doel in veelvoud ontwikkeld worden. Er schuilt ook een aanzienlijk risico in, omdat het onderhoud dat nodig is om de ontwikkelingen bij te benen vaak niet te financieren is. De toegankelijkheid en het behoud van de instrumenten en gerelateerde data komen daardoor snel onder druk te staan.

Sleutel tot het ontwikkelen van duurzamere digitale gereedschappen is interoperabiliteit, aldus de werkgroep 'Strategic IT Recommendations' van het Interedition-project. De gedachte is dat *tools* die elkaars gegevens en processen kunnen gebruiken - dus interoperabel zijn - meer gedefinieerde en gedeelde werkprocessen vereisen. Die kunnen op hun beurt leiden tot efficiëntere spreiding van de verantwoordelijkheden voor het bouwen en onderhouden van de software voor zulke processen. Interoperabiliteit wordt in het Interedition-model breed opgevat. Het is niet alleen een technische eigenschap die zorgt dat programma's met elkaar kunnen praten, maar heeft ook een sociaal aspect: zorg dragen dat betrokken specialisten kennis kunnen uitwisselen en kunnen samenwerken. Tenslotte betekent interoperabiliteit op methodologisch vlak de identificatie van congruente werkprocessen.

Dat dit alles niet slechts theorie is laten de *proof of concept* producten van Interedition zien. CollateX is van deze tools met een actuele 1.0 release het verst gevorderd. Het is software die met literair-kritische precisie variatie in verwante teksten opspoort. Dit is bijvoorbeeld relevant bij de analyse van Darwins *Origin of Species*. Dat werk verscheen in achttien opeenvolgende edities en kende nogal wat mutaties, wat tot debat leidde of Darwin stelliger dan wel onzekerder werd in de loop van de tijd. De analyse van de variatie in extenso die nu mogelijk is, leidt tot de conclusie dat Darwin steeds overtuigder werd van zijn inzichten.

De ontwikkelingsgeschiedenis van CollateX toont de relevantie van Interedition's interoperabiliteitsmodel aan. Gezamenlijke bijeenkomsten voor tekstonderzoekers en ontwikkelaars mondden uit in de definitie van een methodisch werkproces voor digitale collatie van variante teksten, het zogenaamde Gothenburgmodel. Daarnaast werden *bootcamps* voor IT-ontwikkelaars georganiseerd. Deze boden gelegenheid om het methodische model in een coproductie tussen ontwikkelaars wereldwijd te implementeren. Daarmee legde Interedition de basis voor een Open Source development community in de geesteswetenschappen. Zo'n community kan op termijn leiden tot een beter geïntegreerde aanpak van software-ontwikkeling voor de geesteswetenschap. Op technisch niveau leidde het interoperabiliteitsdenken tot een model van microservices. In dit model wordt een groter werkproces zoals de tekstcollatie door CollateX opgedeeld in een aantal kleinere services die onafhankelijk op een gedistribueerde infrastructuur (de *cloud*) kunnen bestaan. De implementatie van CollateX als een set van dergelijke gedecentraliseerde webservices maakt het delen en onderhouden van technische resources praktischer. Mede daardoor maken nu meerdere Europese, Amerikaanse en Canadese projecten gebruik van de CollateX webservice, en dragen zij actief bij aan de ontwikkeling ervan. (Joris van Zundert)

www.interedition.eu
http://collatex.sourceforge.net/

CATCHPlus maakt doorstart

Na de verhuizing van het projectbureau CATCHPlus van het Instituut voor Beeld en Geluid naar het Meertens Instituut heeft het met een grotendeels nieuw team een goede doorstart gemaakt.

CATCHPlus is in het leven geroepen om de onderzoeksresultaten van CATCH (*Continuous Access To Cultural Heritage*) te verzilveren door bruikbare tools en diensten voor de hele Nederlandse erfgoedsector op te leveren. 'Het is het knooppunt tussen ICT en erfgoed,' volgens de nieuwe projectleider Patricia Alkhoven. 'Het gaat om een unieke samenwerking tussen grote erfgoedinstellingen, universiteiten en bedrijfsleven. Dat is ingewikkeld en het is dan ook een uitdaging om resultaten te realiseren. We hebben nu in elk geval de zekerheid dat het project in zijn geheel volbracht kan worden, want we kregen onlangs het bericht dat de aanvraag bij het ministerie van Economische Zaken, Landbouw en Innovatie (EL&I) voor het laatste gedeelte van de financiering gehonoreerd is.'

CATCHPlus werkt aan een centrale infrastructuur om informatie en software te delen, met daarbinnen een repository voor het gemeenschappelijk opslaan en beheren van annotatiedata. Het project wordt gefinancierd door het ministerie van Onderwijs, Cultuur en Wetenschap, de Nederlandse Organisatie voor Wetenschappelijk Onderzoek en het ministerie van EL&I. Het loopt tot medio 2012. (Douwe Zeldenrust)

www.catchplus.nl

Luistert de regering naar het volk?

Het opzetten van een database over politieke responsiviteit was één van de projecten die in maart van dit jaar een middelgrote investeringssubsidie van NWO kregen (zie ook pagina 1). *e-data&research* sprak kort met hoofdaanvraagster Christine Arnold, onderzoekster aan de Universiteit van Maastricht. Haar project heeft een begroting van ruim € 400.000. Eén van Arnolds specialismen is het wetgevingsproces van de Europese Unie (EU). Arnold: 'We willen onderzoeken in welke mate meningen van Europese burgers vertaald worden in wetgeving. We willen daarvoor drie soorten informatie bijeenbrengen: voorkeur van het electoraat, meningen van partijen en regeringen, en gegevens over wetgeving. Veel van de informatie bestaat al. Het probleem is dat de informatie vaak betrekking heeft op een enkel land of op een enkel beleidsterrein, en vaak op verschillende manieren gecodeerd is. Bovendien wordt onvoldoende rekening gehouden met de getrapte besluitvorming binnen de EU. Door de gegevens te harmoniseren ontstaat de mogelijkheid om statistisch betekenisvol onderzoek te doen over meerdere landen en over een langere periode.'

Zijn die bestaande bronnen zomaar rechtenvrij beschikbaar?

Arnold: 'In vrijwel alle gevallen, ja. Eurolex bijvoorbeeld, de bron voor wetgeving van de EU, moedigt hergebruik aan. De meeste datacollecties zijn voor onderzoek opgezet en juichen verdere verspreiding toe. Op grond van onze data-analyse en daaruit resulterende database kunnen andere onderzoekers vervolgens weer gegevens voor eigen analyses downloaden. Daarvoor maken we een eenvoudig te bedienen webinterface.'

Er worden gegevens van vijftien landen verzameld?

Arnold: 'Ja, dat is vooral een praktische keuze, mede gebaseerd op toegankelijkheid, die ons in staat stelt de dataverzameling binnen de drie jaar van het project te kunnen uitvoeren. Onderzoeksassistenten met verschillende taalkundige achtergrond zullen helpen nationale wetgeving te analyseren.'

Hoe worden de politieke standpunten uit de wetgeving gedestilleerd?

Arnold: 'We maken gebruik van Legislative XML, waarin de wetsteksten worden gecombineerd met metadata. In veel van de bronnen worden standpunten al geïndexeerd naar onderwerp. De metadata worden gecodeerd in *Eurovoc*, een thesaurus om de activiteiten van de EU te beschrijven. Op basis van handmatige codering wordt de computer getraind om andere (wets)teksten te analyseren. Computationele linguïstiek is daarvoor onmisbaar. Politicologen maken nog te weinig gebruik van de hulpmiddelen die informatici hebben ontwikkeld, en wij willen dat veranderen.' (PB)

SINDS KORT BESCHIKBAAR

Het overzicht toont databestanden die recent voor onderzoekers beschikbaar zijn gekomen bij CBS, CentERdata, Huygens ING en DANS. Een volledig overzicht van de CBS-bestanden is te vinden op www.cbs.nl/microdata. De LISS panel bestanden van CentERdata zijn kosteloos beschikbaar via het LISS Data Archief www.lissdata.nl/dataarchive. De bestanden van Huygens ING zijn beschikbaar via www.inghist.nl. De DANS databestanden komen van diverse andere onderzoeksinstellingen. Deze kunnen kosteloos worden gedownload vanuit DANS EASY: https://easy.dans.knaw.nl

Centraal Bureau voor de Statistiek	Periode
Bijzondere Zorgkosten – Zorg met verblijf	2009
Bijzondere Zorgkosten – Zorg zonder verblijf	2008-2009
Conjunctuurtest industrie: Verwachtingsenquêtes	2007-2010
Consumentenprijzen voedingsmiddelen	2007-2010
Huurenquête	2010
Enquete ICT-gebruik bedrijven	2009
Statistiek financien ondernemingen	2000-2009
Pensioenaanspraken	2005-2008
Pensioendeelnemingen	2005-2008
CentERdata LISS Data Archive	
LISS Panel	
Health Prevention,(Panidi, K)	2010
Measuring the Desire for Children in Low Fertility Settings: Wave 1 and 2, (Bühler, C., Gauthier, A.H.,Goldstein, J.R. and S.C. Hin)	2010
Are Effective Emotion Regulation Strategies Associated with Financial Capability, (Overveld, M. van, Smidts, A., Atkinson, A., Peffer, G)	2010
Public Attitudes Towards and Knowledge of Conditional Sentences: Part 1 and 2, (Gelder, J.L. van)	2010 2010
LISS Core Study – *Health, Wave 4,* (CentERdata)	2010
Solidarity in Health Care, (Houwen, K. van der)	
Tilburg Consumer Outlook Monitor, December 2010, (Yabar, J.; Pieters, R.; Leenheer, J.)	2010
Allochtonen panel	2010
Solidarity in Health Care,(Houwen, K. van der)	2010
Multiculturalism, (Vijver, F. van de)	2010
Namegenerator (social networks), (Houwen, K. v.d.)	
Dutch Parliamentary Election Studies, part 1 and 2, (Schmeets, H.; Houwen, K. van der)	2010 2010
Trust, wave 1 and 2, (Schmeets, H.; Houwen, K. v.d.)	2010
Operant Motive Test, (Bender, M.,Chasiotis, A., Vijver, F. van de)	
Huygens ING	
Gedigitaliseerde Rijks Geschiedkundige Publicatiën:	
Bescheiden betreffende de buitenlandse politiek van Nederland	1848-1945
De Nederlanders in KeralaBronnen tot de kennis van het leven en de werken van D.V. Coornhert	1663-1701
De tol van Iersekeroord. Documenten en rekeningen	1321-1572
Kroniek van Peter van Os. Geschiedenis van 's-Hertogenbosch en Brabant van Adam tot 1523	
Gedenkschriften van Anton Reinhardt Falck:	
Memories van overgave van gouverneurs van Ambon in de 17e en 18e eeuw	
The dispatches of Thomas Plott and Thomas Chudleigh, English envoys at The Hague	1681-1685
Londense dagboeken van jhr. ir. O.C.A. van Lidth de Jeude	1940-1945
Baltische archivalia:	
Dutch entries in the pound-toll registers of Elbing	1585-1700
Inventarissen van de inboedels in de verblijven van de Oranjes en daarmee gelijk te stellen stukken	1567-1795
Via DANS EASY	
Historische Wetenschappen	
Javanen in Diaspora – Interviewcollectie – (KITLV – Stichting Comité Herdenking Javaanse Immigratie)	2010
Biestkens Bijbel – Nicoline van der Sijs – Stichting Vrijwilligersnetwerk Nederlandse Taal	1560
Archeologie	
Almere Onderzoek 17e eeuws scheepswrak (ADC ArcheoProjecten)	2008
Venlo Maasboulevard. Venlo aan de Maas: van vicus tot stad (ADC ArcheoProjecten)	2009
Een Romeinse nederzetting in Huissen. Een archeologische opgraving in het verlengde van de Hortensialaan te Huissen, plangebied Agropark II, Gemeente Lingewaard – Gelderland (Archeological Research and Consultancy – ARC)	2011
Wageningen, Rouwenhofstraat (Onderzoek en Adviesbureau BAAC)	2011
Sociale Wetenschappen	
Familie-enquête Nederlandse bevolking 2009 (G.Kraaykamp, S.Ruiter, M.Wolbers – Radboud Universiteit Nijmegen)	2009
Aanvullend Voorzieningen Gebruiksonderzoek – AVO (Sociaal en Cultureel Planbureau – SCP)	2007
European Social Survey – ESS – updates and tools (Norwegian Social Science Data Services)	2011
Inkomens Panel Onderzoek – IPO – verbeterde versies (CBS beveiligd microbestand)	2005 2006 2007
Geo-data	
TOP50NL – Digitaal objectgericht topografisch basisbestand Nederland schaal 1:50.000 (Kadaster)	2010
Grondsoortenkaart 2006 - Simplified Soil Map of the Netherlands, (Wageningen UR – Alterra)	2006

EUROPESE RAAD

De Europese Ministerraad (hier bestaande uit de regeringsleiders) is het wetgevend orgaan van de EU

www.fdcw.unimaas.nl/staff/arnold

'Page cannot be found' verleden tijd

Verwijzingen naar bronnen op internet in de vorm van url's (*uniform resource locators)* zijn minder stabiel dan veel gebruikers verwachten. Vaak levert een url (herkenbaar aan de vorm http://www.knaw.nl/artikel.html) na het intypen in de *browser* de foutmelding op: '*page cannot be found*'. De bron is dan in veel gevallen verplaatst naar een andere locatie op internet. Met behulp van ingewikkelde zoekacties via bijvoorbeeld Google lukt het soms alsnog om het document te traceren, maar succes is niet verzekerd. Een herkenbaar probleem voor veel internetgebruikers.

Er is een systeem dat ervoor zorgt dat documenten altijd traceerbaar blijven, ook al worden ze verplaatst. De basis wordt gevormd door PI's (*persistent identifiers*). Het idee is dat elke Nederlandse wetenschappelijke instelling PI's toe kent aan haar bronnen (zoals artikelen, boeken, datasets, websites). Een PI blijft gekoppeld aan de bron, waar deze ook naar toe wordt verplaatst. Een PI is echter geen URL. Er is daarom ook een mechanisme nodig dat de PI herleidt naar de actuele URL van dat document. Met een Engelse term wordt dat mechanisme een *resolver* genoemd.

In het project PersID-NL werken Data Archiving and Networked Services (DANS) en de Koninklijke Bibliotheek (KB) met een subsidie van SURFfoundation samen om een nationale infrastructuur voor PI's inclusief *resolver* te bouwen. Daarnaast wordt de aanzet gegeven om een *Registration Agency* tot stand te brengen dat de afspraken zal gaan beheren tussen nationale instellingen die samenwerken in deze infrastructuur. (Arjan Hogenaar)

http://wiki.surffoundation.nl/display/vp/1.2+Persistent+Identifier+Resolver

Eerste keurmerken voor databeheer toegekend

In december 2010 werd het Data Seal of Approval (DSA) operationeel, bedoeld om kwaliteit en integriteit van onderzoekdata in archieven te waarborgen. Sindsdien hebben zes belangrijke data-archieven het DSA verworven: twee in het Verenigd Koninkrijk, twee in Nederland, een in de Verenigde Staten en een in Frankrijk. Het is voor onderzoekers die hun data willen delen én diegenen die data van anderen willen gebruiken van essentieel belang om te weten dat de kwaliteit en integriteit van die data gewaarborgd is. Daarom ontwierp DANS een keurmerk, het zogenaamde Data Seal of Approval (DSA), dat zestien basiseisen formuleert waaraan goed databeheer moet voldoen.

Sinds 2009 wordt dit DSA beheerd en verder ontwikkeld door een internationaal bestuur, dat een online zelfevaluatieformulier samenstelde aan de hand waarvan de kwaliteit van het databeheer in kaart kan worden gebracht. De zelfevaluaties worden beoordeeld door een DSA-bestuurder alvorens het keurmerk wordt afgegeven.

De tot nu toe gecertificeerde instellingen zijn de Archaeology Data Service (ADS, UK), het DANS Electronic Archiving System (EASY, NL), het Interuniversity Consortium for Political and Social Research (ICPSR, USA), het Platform for Archiving CINES (PAC, FR), het Language Archive van het Max Planck Institute for Psycholinguistics (NL) en het UK Data Archive.

Het DSA heeft internationaal veel waardering geoogst omdat het een relatief licht instrument is dat bovendien als een goede eerste stap fungeert binnen uitgebreidere certificeringsmethoden zoals die momenteel in Europa worden ontwikkeld. Bovendien kan het worden toegepast in alle disciplines: alfa, bèta en gamma. Na de toekenning van het DSA blogde Stuart Jeffrey van de ADS: 'We are delighted that we have achieved a distinction that reflects so well on all the hard work that our curatorial team have put into ensuring that the ADS conforms to internationally recognised best practice in the area.' (IA)

http://datasealofapproval.org/
www.trusteddigitalrepository.eu

Digitale en analoge boeken: bondgenoten of vijanden?

Is digitalisering van boeken een last of een zegen? En welke rol dient Google Books hierin te spelen? Dat was in het kort het thema van de vierde Leerstoel Boek.be op 18 maart in Antwerpen. Een pittig onderwerp voor een publiek van zowel uitgevers als academici, die doorgaans lijnrecht tegenover elkaar staan wanneer het over digitalisering gaat. Gastspreker was Robert Darnton, hoogleraar bij Harvard University, directeur van de hieraan verbonden University Library en grondlegger van het Gutenberg e-Project.

Analoog en digitaal vullen elkaar aan

Darnton stelt dat het analoge boek en het e-boek vaak onterecht gepresenteerd worden als twee uitersten. Analoge boeken en e-boeken ondermijnen elkaars positie niet, maar vullen elkaar aan. In de geschiedenis zie je dat verschillende verschijningsvormen van het boek lange tijd naast elkaar bestonden. Ook na de uitvinding van de boekdrukkunst werden nog regelmatig schrijvers ingehuurd voor het vermenigvuldigen van boeken. Bij oplagen van minder dan honderd was het simpelweg goedkoper het werk over te laten schrijven dan het te laten drukken. Gezien de lange periode dat het boek al bestaat, en zeker nu het aantal gedrukte boeken nog ieder jaar stijgt, verwacht Darnton niet dat het analoge boek heel snel zal verdwijnen: 'Een boek blijft een boek.'

Er zijn wel nieuwe toepassingen denkbaar. Waarschijnlijk zal printing-on-demand een hoge vlucht nemen. Een voorbeeld hiervan bestaat al in de vorm van de Espresso Book Machine. Mensen zijn toch erg gehecht aan de verschijningsvorm van het boek, ondanks apparaten als de iPad en de Kindle e-reader. Uitgevers hebben het meeste kans op succes als zij zich flexibel opstellen en zich richten op uitbreiding van de markt in plaats van die te beperken door bindende clausules op te leggen aan auteurs.

Naar een nieuw model

Zorgwekkend vindt Darnton de stijging in de prijzen van abonnementen voor wetenschappelijke tijdschriften. Waar een abonnement in 1986 nog gemiddeld $154 kostte, was dat in 2009 $2000, terwijl hoogleraren vaak gratis input voor het tijdschrift leveren. Bij bibliotheken komt hierdoor het aanschaffen van ander materiaal dan tijdschriften onder druk te staan. Bij het vaststellen van de bijdragen aan deze tijdschriften valt de keuze vaak op gerenommeerde onderzoekers, waardoor jonge onderzoekers weinig kansen krijgen om te publiceren.

Alleen de onderzoekers en de universiteiten zelf kunnen aan deze situatie iets veranderen, aldus Darnton. Als zij er vaker voor kiezen om in een Open Access online tijdschrift te publiceren, wordt deze negatieve spiraal doorbroken. Dit vereist echter een radicale ommezwaai in het wetenschappelijk denken over publiceren. Bij Harvard is hiermee al een begin gemaakt: hoogleraren deponeren hun onderzoeksdata en publicaties in een Open Access repository, genaamd DASH – Digital Access to Scholarship at Harvard.

Google versus instellingen

Momenteel loopt er in de Verenigde Staten een rechtszaak tussen Google en uitgeverijen en auteurs over de rechten van door Google gedigitaliseerde werken, het *Google Books Settlement*. Google wil de rechten van

zogenaamde *orphan books*, boeken zonder aanwijsbare rechthebbende, verwerven wanneer het bedrijf het werk digitaliseert. Darnton hoopt dat dit geen doorgang zal vinden. Google streeft als commercieel bedrijf naar het maken van winst – niet naar het behoud en de verspreiding van cultureel erfgoed dat voor de universiteitsbibliotheken bovenaan staat. Hierom ook vindt Darnton het een slecht idee om Google digitalisering te laten coördineren en in eigen hand te laten houden.

Daarentegen juicht hij samenwerkingsverbanden tussen bibliotheken toe, zoals het project Europeana van o.a. het Nationaal Archief en Digitaal Erfgoed Nederland. Google kan binnen deze samenwerkingsverbanden wel uitvoerende taken op zich nemen.

Marieke Polhout

Naschrift: Een week na de lezing werd bekend dat het Google Books Settlement verworpen is door een Amerikaanse rechter. Google krijgt dus niet automatisch de rechten van door haar gedigitaliseerde boeken.

http://history.fas.harvard.edu/people/faculty/darnton.php
http://dash.harvard.edu/ www.europeana.eu/portal/
www.ondemandbooks.com/

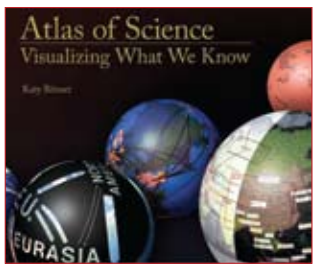
Lezen: Robert Darnton, *The Case For Books – Past, Present and Future*. New York, Public Affairs, 2009.

Gelezen

Marjan Grootveld, Jeff van Egmond, Brenda Sørensen (red.): *Data reviews, peer reviewed research data*. Den Haag, DANS, 2011. DANS studies in digital archiving 5; ISBN 978-94-90531-05-8
Het reviewen van data was een lang gekoesterde wens van DANS, die manifest werd na gesprekken met SURFfoundation over datakwaliteit. Zoals de kwaliteitsbewaking van wetenschappelijke publicaties plaats vindt door *peer review*, zo zouden ook datasets gereviewd kunnen worden. Reviews kunnen ook aanleiding geven tot verdere discussie over datasets, en zo bijdragen aan de vorming van communities van onderzoekers rond datacollecties. Natuurlijk zal het reviewen van data anders gaan dan dat van boeken of artikelen. Ook het online karakter zal de wijze van reviewen van datasets beïnvloeden. Dataset-reviews zullen meer lijken op gebruikers-reviews van producten zoals digitale camera's of van hotels. Eind 2010 is een verkennend onderzoek uitgevoerd onder een aantal afnemers van datasets uit DANS EASY, aan wie werd gevraagdom een gedownloade dataset te beoordelen. Dit rapport beschrijft de opzet van de pilot, de uitkomsten en aanbevelingen, waarvan de voornaamste is om een continue vorm van data reviewing door afnemers in te voeren.
www.dans.knaw.nl/content/publicaties

Koninklijke Nederlandse Akademie van Wetenschappen, *Universiteiten en onderzoeksinstellingen in Nederland 2011*. Den Haag, SDU Uitgevers, 2011; ISBN 978-90-12-57102-9
Zoekt u de weg in wetenschappelijk onderwijs of onderzoek? Dan vindt u met deze gids eenvoudig de juiste persoon of instelling met alle relevante contactgegevens met behulp van de uitgebreide persoons- en zaakregisters. De gegevens voor de gids zijn ontleend aan de Nederlandse Onderzoek Databank (NOD) en zijn eveneens terug te vinden in de wetenschapsportal NARCIS.
De nieuwe publicatie bestaat uit twee delen. Het eerste deel verstrekt per Nederlandse universiteit informatie over circa acht duizend hoogleraren, bijzonder hoogleraren en universitaire hoofddocenten. Het tweede deel bevat 497 beschrijvingen van onderzoeksinstellingen.
www.dans.knaw.nl/content/publicaties

Katy Börner: *Atlas of science; Visualizing what we know*. Cambridge-Ma USA, MIT Press, 2010; ISBN 978 0 262 01445 8
Cartographic maps have guided our explorations for centuries, allowing us to navigate the world. Science maps have the potential to guide our search for knowledge in the same way, helping us navigate, understand, and communicate the dynamic and changing structure of science and technology. Allowing us to visualize scientific results, science maps help us make sense of the avalanche of data generated by scientific research today. Atlas of Science, features more than thirty full-page science maps, fifty data charts, a timeline of science-mapping milestones, and 500 colour images; it serves as a sumptuous visual index to the evolution of modern science and as an introduction to 'the science of science' – charting the trajectory from scientific concept to published results. Not even the most brilliant minds can keep up with today's deluge of scientific results. Science maps show us the landscape of what we know. (See also page 3) http://mitpress.mit.edu/catalog



Column
Martijn de Groot

Datasabbatical

Twee druk bezette types proberen al maanden een eetafspraak te maken, maar het lukt niet omdat ze elkaar steeds niet kunnen bereiken. Als ze elkaar onverwacht tegen komen zonder agenda bij de hand, besluiten ze het er niet bij te laten zitten: 'Ik vraag mijn antwoordapparaat om jouw antwoordapparaat te bellen en dan lossen ze het samen maar op!'. Het is een oude grap, uit de tijd zonder mobiele telefoon, email, sms, ping, yammer of twitter. De antwoordapparaten hadden dus geen keus – ze moesten er met elkaar wel uitkomen. Wij weten natuurlijk dat dat niet is gebeurd. Zij wáren immers niet de afspraakmakende partijen; ze representeerden die alleen. Zonder tussenkomst van de echte mensen zou dat etentje er nooit komen.

Waarom schrijf ik nu, dertig jaar later, dat verhaal weer op? We leven in een tijd waarin je een eetafspraak simpelweg in de e-agenda van je disgenoot dumpt. Een kwartier van te voren sms je om de plek af te spreken, en op de afgesproken tijd en plaats laat je aan je netwerk weten waar je hebt ingecheckt, met coördinaten en al. Maar we leven ook in een tijd waarin steeds meer gegevens worden vastgelegd in de vorm van dataverzamelingen over van alles en nog wat, en daardoor moet ik steeds weer aan die antwoordapparaten denken. Die verzamelingen worden behouden, verzorgd, gecultiveerd en opengesteld. Ze worden bij elkaar gezet, gekoesterd en gekoppeld, en zo brengen ze nieuwe dataverzamelingen voort die op hun beurt weer met elkaar in contact kunnen worden gebracht terwijl wij er omheen dansen, juichend over de almaar groeiende onderzoekperspectieven.

Begrijp mij goed, ik dans van harte mee. Dat heb ik althans de laatste vijf jaar gedaan als hoofdredacteur van dit mooie datablad. Het lijkt me een voorrecht om onderzoeker te zijn in een tijd waarin zo veel empirisch materiaal klaar ligt om voor nieuw onderzoek te worden gebruikt. Maar toch.

Dataverzamelingen zijn niet de realiteit. Ze vormen een representatie daarvan. Van een déél daarvan. Misschien van een deel waarnaar we voor onze onderzoekvraag net niet op zoek waren, maar dat er wel behoorlijk op lijkt en voor niks beschikbaar is. En misschien zijn ze een vertekende afspiegeling, of zelfs een slechte afspiegeling. Als we te lang blijven juichen terwijl zulke verzamelingen zich vermenigvuldigen, kunnen we nog veel verder van de realiteit af komen te staan.

Ik zou er wel voor zijn dat hard core datagebruikers regelmatig – bijvoorbeeld eens in de vijf jaar – verplicht werden zich een tijdje van de data af te keren en zich onder te dompelen in de echte werkelijkheid. Een soort datasabbatical. Enquêteren in een supermarkt bijvoorbeeld, of oude – papieren – handschriften ontcijferen.

Zelf geef ik het goede voorbeeld. Ik heb dit datablad vijf jaar met plezier geleid, maar nu ga ik weer schrijven over alles wat er àchter die data schuil gaat. Dat wordt verfrissend. Voor u en mij.

Martijn de Groot is zelfstandig communicatieadviseur en tekstschrijver.

COLOFON

e-data&research is het kwartaalblad in Nederland over data en onderzoek in de alfa- en gammawetenschappen. Het verschijnt onder auspiciën van CentERdata, CLARIN-NL, DANS, het Huygens ING, het Centraal Bureau voor de Statistiek, de Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica. Toezending kosteloos aan relaties van de stakeholders en op verzoek aan studenten in de alfa- en gammarichtingen. Oplage: 9100.

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

Uitgever: Stichting Uitgeverij *e-data&research*, Postbus 93067, 2509 AB Den Haag

Redactieadres: Postbus 93067, 2509 AB Den Haag; **t** (070)3446484 **f** (070)3494451 **e** edata@dans.knaw.nl

Redactie: Inge Angevaare, Eric Balster, Ronald van der Bie, Peter Boot, Martijn de Groot (hoofd/eindredacteur), Thijs Hermsen, Erica Renckens, Jetske van der Schaaf

Aan dit nummer werkten mee:
Arjan Hoogenaar, Edwin Klijn, Marieke Polhout, Douwe Zeldenrust, Joris van Zundert.

Redactiesecretariaat:
Lucas Pasteuning, Jetske van der Schaaf

Vormgeving en opmaak:
Ellen Bouma, Alkmaar

Productie: Amsterdam University Press

Druk: Ten Brink, Meppel

ISSN: 1872-0374