e-data & research

Jaargang 4 nummer 3 / december 2009

Kwartaalblad over data en onderzoek in de alfa- en gammawetenschappen

Mogelijk gemaakt door: Data Archiving and Networked Services, Centraal Bureau voor de Statistiek, Huygens Instituut, Internationaal Instituut voor Sociale Geschiedenis, Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica

Pagina 3 • Soundbites online • Eind september gaat de website met 'sprekende kaart' van het Soundbites-project online. Het Meertens Instituut maakt met de kaart zo'n duizend uur geluidsopnames van gesprekken in dialect beschikbaar.

Pagina 4 • Social Media • Social Media op het internet worden steeds meer zelf een aantrekkelijk onderzoekterrein.

WIEBE KIESTRA

Pagina 9 • Peter Doorn • DANS-directeur Peter Doorn noemt zichzelf een omnivoor en een handelaar in tweede-hands data. 'Ik vind alles interessant, als er maar data ter beschikking komen'.

Pagina 10 • Kohnstamm Instituut • De rubriek Focus besteedt deze aflevering aandacht aan het Kohnstamm Instituut dat dit najaar werd verzelfstandigd.

Pagina 12 • Kwaliteitsstempel • Een kwaliteitsstempel voor survey onderzoek is zo makkelijk nog niet, reageert Ineke Stoop op een suggestie in de vorige e-data.

EN VERDER

Brieven Van Gogh compleet digitaal

Negenhonderdtwee brieven van en aan Vincent van Gogh zijn sinds 6 oktober voor iedereen te bestuderen op de website http://vangoghletters.org. Het is voor het eerst dat de complete verzameling brieven van de schilder op deze manier is samengebracht èn volledig digitaal beschikbaar gemaakt.

VAN GOGH MUSEUM

Aquarel die van Gogh bijsloot bij brief 271: Bloeiende boomgaard met paartje: lente.

De website werd tegelijk met een zesdelige boekuitgave gelanceerd bij de opening door koningin Beatrix van een tentoonstelling over de brieven in het Van Gogh Museum in Amsterdam. Dat museum is samen met het Huygens Instituut verantwoordelijk voor het megaproject, waaraan vijftien jaar is gewerkt.

De brieven worden aangeboden in hoge resolutie facscimile maar ook als tekstbestanden in Nederlands en Engels, in verschillende vormen en steeds voorzien van uitvoerige annotaties en verantwoording. Ze zijn volledig doorzoekbaar en bevatten links naar alle schilderijen en kunstenaars waarnaar de brievenschrijvende schilder verwijst. Daardoor zijn de gebruiksmogelijkheden en informatie zo compleet dat wetenschappelijk medewerker Hans Luijten van het museum tegenover *e-data&research* de bewering aandurft: 'Ik zou niet weten wat wij op dit moment nog meer aan kunnen bieden'. Huygensdirecteur Henk Wals ziet zulke mogelijkheden wel, maar dan in de sfeer van toegevoegde faciliteiten. 'Je zou er nog analyses van Van Goghs taalgebruik aan toe kunnen voegen of informatie over de plek waar hij zich op zeker moment bevond. We staan nog maar aan het begin met digitale edities'. Als wetenschappelijke editie ziet echter ook Wals de brievensite als de ultieme Van Gogh-uitgave: 'Dit is de beste digitale editie die er op dit moment in de wereld bestaat'.

Soortgelijke kwalificaties waren er op de presentatie voor de boekeditie, die eveneens alle brieven en een overvloed aan gerelateerd beeld bevat (*'The art publishing event of the decade,'* aldus de Britse Telegraph). Het project was volgens de bezorgers halverwege de jaren negentig nog bedoeld voor een vijfdelige en later zelfs voor een twaalfdelige boekuitgave: de ultieme en complete printuitgave. Na de milleniumwisseling rees wel het besef dat de digitale vorm betere mogelijkheden zou bieden voor een complete annotatie èn voor onderzoekers over de hele wereld om gebruik te maken van het materiaal. In 2004 viel het besluit om het streven naar volledigheid in de annotaties te richten op de webeditie terwijl, aldus Wals, 'van de boekenserie een aantrekkelijke publikatie voor een breder publiek zou worden gemaakt. De webeditie is de eigenlijke editie.' Dat die laatste versie in de publiciteit minder aandacht kreeg dan de fraaie, door Wim Crouwel vormgegeven boekenserie in drie talen vindt Wals niet verwonderlijk. 'De betrokken uitgevers hebben voor deze uitgave grote financiële risico's genomen waarvoor we ze trouwens zeer erkentelijk zijn, en voor de verspreiding ervan is publiciteit belangrijk.'

Projectleider Peter Boot van het digitale deel van de editie ziet geen dreigende concurrentie tussen de gratis webeditie en de niet goedkope boekuitgave. 'We verwachten dat het boek goed verkocht zal worden. Het zijn twee verschillende soorten beleving die elkaar niet hoeven te bijten'. De site had op zijn eerste dag zo'n vijfduizend bezoekers en dat aantal liep in de loop van oktober uit tot vijftig duizend. Boot: 'We zijn benieuwd waar het zich stabiliseert als de publiciteitsgolf wat is weggeëbd'.

Extra Jubileumpagina's

Dit jaar is het vijfenveertig jaar geleden dat in Nederland het eerste digitale wetenschappelijk archief werd opgericht: de Steinmetz Stichting voor de sociale wetenschappen. Twintig jaar geleden werd het Nederlands Historisch Data Archief in het leven geroepen, vijftien jaar geleden het Wetenschappelijk Satatistisch Agentschap en vijf jaar geleden het e-Depot voor de Nederlandse Archeologie.

Reden voor data-instituut DANS waar de jubilerende instellingen nu deel van zijn, om het symposium 'Door Data Gedreven' te organiseren op 2 december in Den Haag. En voor *e-data&research* om in dit nummer vier extra pagina's op te nemen. Op het symposium en in de bijlage ligt het accent op de toekomst. De Steinmetz Stichting was een eenzame verkenner in het voorportaal van het digitale tijdperk. Inmiddels heeft dat tijdperk de alfa- en gammawetenschappen volledig in zijn greep. Welke nieuwe mogelijkheden dienen zich aan? Welke innovatieve kansen liggen er, en welke creatieve oplossingen zijn al bedacht?

Lees het op pagina 4 t/m 8 van deze *e-data&research*.

Veel activiteiten in Open Access Week

In de internationale Open Access Week van 19 tot en met 23 oktober heeft de Nederlandse academische wereld van zich laten horen. Er waren veel verschillende activiteiten om de aandacht op Open Access te vestigen. Op verschillende universiteiten en hogescholen werden lezingen, symposia en workshops gehouden, bedoeld om onderzoekers te wijzen op de voordelen van open toegang tot onderzoeksresultaten en uit te leggen hoe dat kan. Zo organiseerde de universiteit Utrecht een goed bezochte lezing over citaties en zichtbaarheid en ontving Tim Berners Lee, de grondlegger van het World Wide Web, aan de Vrije Universiteit een eredoctoraat. Ook sportief kreeg Open Access aandacht: in Groningen deden de Open Access Runners van de universiteitsbibliotheek mee aan de vier mijl van die stad. In de Leidse actie *Eerste Hulp bij Open Access* hielp een team van *OA dokters* dat de faculteiten bezocht met het digitaal beschikbaar stellen van publicaties. Daardoor konden bijna driehonderd wetenschappelijke publicaties worden toegevoegd aan het Leids Repositorium. Voor de mensen die niet bij deze activiteiten aanwezig konden zijn organiseerde DANS (Data Archiving and Networked Services) een online forum: de Open Data Speakers Corner. Op deze 'virtuele zeepkist' werd vanuit verschillende invalshoeken over de vrije toegang tot digitale gegevens gedebatteerd, met key-note bijdragen van bekende namen uit onderzoek, uitgeverij en bibliotheekwereld.Hier werd duidelijk dat wetenschappelijke data mee moeten tellen als onderzoeksoutput, net als bij publicaties nu het geval is, althans volgens de meerderheid van stemmers op de poll van maandag.

De universitaire ict-organisatie SURF heeft in deze week de nationale website over Open Access www.openaccess.nl gelanceerd. Die geeft de Nederlandse situatie op het gebied van Open Access weer en biedt praktische informatie over Open Acces publiceren in de verschillende vakgebieden. Daarnaast zijn er nieuwsberichten en korte films te vinden. Een zeer goed bezocht onderdeel van de site is *experts speak* waarin onder andere KNAW-voorzitter Robbert Dijkgraaf zijn visie geeft op de voordelen van Open Access. (Marjolein van den Dries)

AGENDA

2 december
Den Haag, De Glazen Zaal
Jubileum Symposium 'Door Data Gedreven'
Georganiseerd door DANS ter gelegenheid van verschillende jubilea van archiefinstituten. Onderzoekers uit zeer uiteenlopende disciplines binnen de alfa- en gammawetenschappen aan het woord over nieuwe vragen, methodes en antwoorden die zij met behulp van innovatief datagebruik hebben verkend. Zie ook elders in dit nummer. www.dans.knaw.nl/nl/dans_symposia/2009_4/

3 en 4 december
Bonn, International Data Service Center of the Institute for the Study of Labor
1st Annual European DDI Users Group Meeting: DDI – The Basis of Managing the Data Life Cycle
Bringing together DDI users and professionals from all over Europe. Anyone interested in developing, applying, or using DDI is invited to attend and present. www.iza.org/

2-4 december
London, Millennium Gloucester Hotel & Conference Centre Kensington
5th International Digital Curation Conference – IDCC
Moving to Multi-Scale Science: Managing Complexity and Diversity
www.dcc.ac.uk/events/dcc-2009

2-4 december
Paris , La Sorbonne
Berlin7 Open Access Conference
International follow-up conference to the 'Berlin Declaration'. The sessions will focus on various aspects of open access especially when reaching diverse communities. www.berlin7.org/

9, 10 december
Rotterdam, De Doelen
Digital Strategies for Heritage – DISH
A new bi-annual international conference on digital heritage and the opportunities it offers to cultural organisations. Triggered by changes in society, heritage organisations face many challenges and need to make strategic decisions about their services. DISH2009 aims at sharing knowledge about and experiences with digital strategies.
www.dish2009.nl/

9-11 december
Oxford, Kassam Stadium Conference Centre
5th IEEE International Conference on e-Science
Designed to bring together leading international and interdisciplinary research communities, developers, and users of e-Science applications and enabling IT technologies. The conference serves as a forum to present the results of the latest research and product/tool developments and to highlight related activities from around the world.
www.escience2009.org/

14 december
Den Haag, Koninklijke Bibliotheek
PLANETS – Toolkit voor ons digitaal geheugen
Bijeenkomst van de NCDD, de Koninklijke Bibliotheek en het Nationaal Archief voor iedereen die te maken heeft met het beheer van digitale informatie, die weleens gehoord heeft van PLANETS maar niet weet wat het precies is, of die nu wel eens in het Nederlands uitgelegd wil krijgen wat PLANETS doet. www.ncdd.nl/toolkit.php

17-20 februari
Frankfurt, Frankfurter Goethe-Museum und Universität Frankfurt
Medienwandel – Medienwechsel in der Editionswissenschaft
Dreizehnte internationale Tagung der Arbeitsgemeinschaft für Germanistische Edition. Die Tagung ist sowohl international als auch interdisziplinär ausgerichtet.
www.ag-edition.org/html/aktuell.html

19, 20 februari
Londen – University College – Institute of Archaeology
CAA UK 2010 meeting
Annual conference on Computer Applications and Quantitative Methods in Archaeology
http://unstan.arch.ucl.ac.uk/caauk/

Digitale duurzaamheid: wie, wat en hoe?

Feedback verzamelen op het eerste interimrapport van het PARSE.Insight project. Dat was het doel van de PARSE Insight Workshop op 21 en 22 september in Darmstadt. De Europese Commissie hecht veel belang aan een Europese e-onderzoeksinfrastructuur, waarbij digitale duurzaamheid een belangrijke rol speelt. PARSE.Insight is een twee jaar durend project, mede betaald door de Europese Unie, gericht op digitale duurzaamheid binnen de wetenschap. In contrast met de disciplinegerichte projecten DARIAH, CESSDA en CLARIN, richt PARSE. Insight zich op het ontwikkelen van een infrastructuur voor blijvende toegang tot digitale wetenschappelijke informatie *in het algemeen*.

Het interimrapport bevat de resultaten van een grootschalige enquête onder de verschillende stakeholders van het digitaal preserveren van onderzoeksdata (zie pagina 10). Het stond centraal in de diverse presentaties, en het werd besproken in door de PARSE-projectleider David Giaretta volgens steeds andere criteria ingedeelde groepen uit de vijftig deelnemers.

Een greep uit de belangrijkste issues: wie heeft toegang tot welke onderzoeksdata (alleen voor de eigen onderzoeksdiscipline? Wat mag wel en wat niet in het kader van de Europese privacy- en copyrightwetgeving?); data repositories (organiseren per discipline of institutioneel, instituut of universiteit?); wat is de rol van uitgevers bij digitale duurzaamheid (uitgevers staan niet bepaald te springen om zelf faciliteiten daarvoor te ontwikkelen); de uitwerking van het concept *enhanced publications* (tijdschriftartikelen met onderliggende data); kostenmodellen; wat gaan Google en Microsoft doen op dit terrein?

Ook werd duidelijk hoe belangrijk coördinatie van alle Europese en landelijke projecten hierover is (ESFRI, voor Nederland: NCDD en/of SURF?). (Heiko Tjalsma)

www.parse-insight.eu

Hoe verder met MIXED?

Van 10 tot 12 september vond er in Scheveningen een *consultation workshop* plaats van het MIXED project van DANS. MIXED is erop gericht om data in ontoegankelijke bestandsformaten bruikbaar te maken door ze te om te zetten naar helder gedefinieerde XML formaten. Het project heeft resultaten opgeleverd voor de formaten van dBase, DataPerfect, MS Access en MS Excel. Ook is er een elders ontwikkelde *converter* voor het sociale statistiekprogramma SPSS in opgenomen.

Voor de workshop waren experts uit binnen- en buitenland uitgenodigd met deskundigheid op de terreinen van digital preservation, data archivering, research infrastructuur en open source software ontwikkeling. De grote vraag aan de deelnemers was: hoe kunnen we de resultaten van MIXED een optimale plaats geven in het geheel van inspanningen om onderzoeksdata herbruikbaar te houden?

De workshop leverde een aantal duidelijke adviezen op: 1. Presenteer het XML conversieresultaat niet zozeer als een bijdrage om de data te behouden, maar als een bijdrage om ze te hergebruiken; 2. Werk samen met het Europese PLANETS project, dat een testbed voor preservatie gereedschappen aan het ontwikkelen is (een omgeving waar gebruikers experimenteel kunnen vaststellen hoe tools met zelf-ingebrachte data omgaan); 3. Vind concrete gevallen waar een onderzoeker baat heeft bij conversies naar nieuwe formaten. Zo kan er een sneeuwbaleffect ontstaan.

Inmiddels is vanuit MIXED, in het verlengde van het tweede advies, op een workshop van PLANETS ontwikkelaars in Wenen begonnen om MIXED op hun testbed te zetten.

RENÉ VAN HORIK

Deelnemers demonstreren dat zij met meerdere petten op aan de MIXED workshop deelnemen

Verder zijn er projectvoorstellen in de maak om teksten vanuit allerlei formaten om te zetten naar TEI (Text Encoding Initiative), als hulpmiddel voor corpusbouwers. Dit kan een mooie uitwerking van het derde advies worden. (Dirk Roorda)

http://mixed.dans.knaw.nl/node/434
www.planets-project.eu/

Beheer en gebruik van historische bronnen verzorgingsstaat

Beheerders en gebruikers van papieren en electronisch bronnenmateriaal, in het bijzonder over de Nederlandse verzorgingsstaat, lijken te leven in 'twee werelden die soms niet op elkaar lijken aan te sluiten'. Dat constateerden althans de organisatoren van een expert-meeting op vrijdag 30 oktober bij het Kenniscentrum Historie Zorgverzekeraars op de VU. Een debat, of althans nader contact, tussen beide groepen leek daarom nuttig. Het kenniscentrum nam de organisatie op zich samen met het onderzoeksbureau Ecade en het Instituut voor Nederlandse Geschiedenis.

Er ontstond inderdaad een brede en nuttige discussie met nogal wat praktijkvoorbeelden uit de geschiedschrijving van de Nederlandse verzorgingsstaat. Voor beheerders van archiefbronnen, wel of niet uitgegeven, is het vaak onduidelijk wie hun gebruikers zijn en hoe de toegangen tot de bronnen worden gebruikt en gewaardeerd. In het digitale internet-tijdperk is het vaak nog lastiger dan in de traditionele papieren situatie om (groepen van) gebruikers te identificeren: weblogs zeggen niet alles. Of de juiste prioriteiten worden gelegd bij de keuzes voor het digitaliseren van papieren archiefmateriaal en het maken van toegangen werd vanuit het perspectief van de gebruiker, de onderzoeker, betwijfeld. De digitale dienstverlening van de Nederlandse archieven is de laatste tijd weliswaar sterk toegenomen, maar de overgrote meerderheid van de archiefbronnen is nog steeds alleen van papier. De op zichzelf terechte focus van de Nederlandse archieven op *digitale* ontsluiting gaf aanleiding tot zorgen over het behoud van deze papieren archieven en de toegang daartoe op de klassieke studiezaal.

Voor wat betreft de historische bronnen in Nederland staan we aan het begin van een waarschijnlijk langdurige overgang van papier naar digitaal. Een dag als deze draagt ertoe bij dat beheerders én gebruikers hierover op positieve wijze met elkaar in gesprek komen. (Heiko Tjalsma)

Europese bibliotheken willen digitaliseren, maar kunnen vaak niet

De gezamenlijke Europese wetenschappelijke bibliotheken hebben enorme collecties waarvan nog in 2008 nog maar twee à drie procent digitaal beschikbaar was, zo becijferde de Europese Commissie. Daarom kwam de Commissie zelf in actie met het *i2010 Digital Libraries* initiatief en organiseerde LIBER, de Europese organisatie van wetenschappelijke bibliotheken, van 19-21 oktober in Den Haag zijn tweede digitaliseringscongres.

Gastheer Bas Savenije, directeur van de Koninklijke Bibliotheek, maakte geen geheim van zijn missie om zoveel mogelijk informatie zo open mogelijk ter beschikking van de wetenschap te stellen toen hij in een T-shirt van de net geopende Open Access Week het podium besteeg. Ook de aanwezige bibliotheken hebben zo langzamerhand de schroom overwonnen om hun collecties prijs te geven aan het internet; ze doen bijvoorbeeld in groten getale mee aan het *Europeana* project, dat de metadata van vele bibliotheken bij elkaar brengt en als Europese collectie presenteert.

En toch valt het aanbod nog tegen. Vooral onderzoekers in de alfawetenschappen, die meer dan de gamma- en betawetenschappen afhankelijk zijn van gedigitaliseerde papieren bronnen, moeten nog vaak de reis naar een fysieke bibliotheek maken. Het congres besprak de voornaamste obstakels voor meer digitale toegang: gebrek aan financiering en auteursrechtelijke belemmeringen. Voor beide zal in Brussel nog veel lobbywerk gedaan moeten worden. Tenzij alle bibliotheken in zee gaan met Google, maar daar voelen slechts weinigen voor. Google staat voor groot en veel, maar lage kwaliteit en geen garanties voor continuïteit. Het is de vraag of de wetenschap daarmee gediend is. (IA)

www.libereurope.eu/
www.europeana.eu

Steun voor interim-rapport Digitaal Geheugen

Op 18 september vond in de Koninklijke Bibliotheek een conferentie plaats om de resultaten te bespreken van de Nationale Verkenning Digitale Duurzaamheid. Die verkenning bestond uit een zes maanden durend onderzoek naar de duurzaamheid van digitale informatie in Nederland, zowel in de wetenschap als in de sectoren overheid/archieven en cultureel erfgoed.

De conferentie was georganiseerd door de Nationale Coalitie Digitale Duurzaamheid (NCDD), die ook het initiatief had genomen tot de verkenning, waarvan het interim-rapport op 1 juli verscheen onder de titel *Toekomst voor ons digitaal geheugen*. In de NCDD werken de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), Data Archiving and Networked Services (DANS), de Koninklijke Bibliotheek (KB), SURF en het 3TU Datacentrum samen met andere publieke organisaties om te onderzoeken hoe de kwetsbaarheid van digitale informatie effectief kan worden aangepakt. Doel van het congres was te toetsen of alle betrokkenen bij het onderzoek zich herkenden in de conclusies van het projectteam. Dat bleek het geval.

Nog lang niet overal wordt digitale informatie duurzaam beheerd, aldus de rapportage, die verschillende oorzaken aanwijst. Zo is er te weinig bewustzijn van het probleem en te weinig kennis en informatie over duurzaamheid, is de informatiehuishouding zelf niet op orde en is de financiering vaak projectmatig, zodat er te weinig continuïteit is. Ook zijn vaak verantwoordelijkheden niet goed benoemd: het particuliere belang van de dataproducent strookt lang niet altijd met het algemeen belang. Een digitaal depot is bovendien onbetaalbaar voor kleine organisaties en er zijn nog te weinig gereedschappen en diensten beschikbaar. Ten slotte zijn er nog onvoldoende selectiemethoden die zijn aangepast aan het digitale tijdperk

JACQUELINE VAN DER KORT

Sprekers op de conferentie, met de klok mee: René Smit en Inge Angevaare (beiden NCDD), dagvoorzitter Yola de Lusenet en Bas Savenije (KB).

In parallelsessies werkten de drie sectoren wetenschap, overheid/archieven en cultureel erfgoed aan oplossingen voor al deze knelpunten, maar daarvoor was de tijd van anderhalf uur wel erg kort. René Smit, voorzitter van de NCDD en van het bestuur van de Vrije Universiteit, benadrukte in zijn slotwoord de eigen verantwoordelijkheid van alle betrokken organisaties. De NCDD, zo zei hij, zal het probleem niet voor de aanwezigen oplossen. Wel mag men van de coalitie verwachten dat zij bestuurlijke impulsen organiseert om duurzaam gedrag te bevorderen. Smit steunde ook de veel gehoorde roep tijdens het congres dat het hoog tijd is om vooral pragmatisch aan de gang te gaan. (IA)

www.ncdd.nl/documents/NCDDcongres20090918verslag.pdf

Beschermheiligen digitaal

Op 1 november zijn het Meertens Instituut en DANS (Data Archiving and Networked Services) begonnen aan het project *Beschermheiligen in Nederland*. Gefinancierd vanuit het DANS-programma Kleine Dataprojecten, zal dat project in het najaar van 2010 een digitale atlas van beschermheiligen opleveren.

De databank in wording is gebaseerd op de *Volkskundige Atlas van Nederland en Vlaanderen*, een transnationaal project uit de jaren zestig van de vorige eeuw. Daarin werd een aantal verschijnselen op het terrein van de Nederlands-Vlaamse volkscultuur in hun verspreiding beschreven. Die verschijnselen, waaronder een onderdeel van de toenmalige heiligenverering in beide landen, zijn vastgelegd in verspreidingskaarten. Onder andere J.J. Voskuil heeft er aan het begin van zijn aanstelling bij het Meertens Instituut aan gewerkt en de kaarten van commentaar voorzien. Het geheel blijkt een schat aan informatie te bevatten.In de atlas, die nu alleen nog in papieren vorm is te raadplegen, is de focus gericht op een aantal beschermheiligen van mens en dier, die konden worden vereerd en aangeroepen om bij ziekte en andere problemen hulp te bieden. De beschermheiligen zijn een belangrijk onderdeel van het brede veld van heiligencultussenen daarmee van de religieuze cultuur van het dagelijks leven. De digitale kaart van beschermheiligen die het Meertens Instituut en DANS tot stand willen brengen zal onder andere gecombineerd worden met de Databank en kaart Bedevaart en Bedevaartplaatsen in Nederland. Voor het eerst zal met de voorgestelde digitale kaart systematisch onderzoek naar heiligenverering in Nederland kunnen plaatsvinden. (Douwe Zeldenrust en Peter Jan Margry)

PETER JAN MARGRY

Christoffel als beschermheilige van het verkeer

www.meertens.knaw.nl/bedevaart/
www.dans.knaw.nl/nl/projectenpagina/kdp/

Digitaal Wetenschapshistorisch Centrum in het Engels

In november is de Engelstalige versie van het Digitaal Wetenschapshistorisch Centrum (DWC) opengesteld voor onderzoekers. Het DWC, waarvan de Nederlandstalige versie sinds mei te raadplegen is, bevat bronnen en gereedschappen voor de wetenschapshistorisch geïnteresseerde onderzoeker in de breedste zin. Het omvat niet alleen de geschiedenis van de natuurwetenschappen, maar ook die van andere academische disciplines en van academische instellingen.

Enkele bronnen die nu deel uitmaken van het DWC bestonden al en zijn er nu in opgenomen, andere zijn speciaal voor de site opgezet. Het DWC bevat onder meer een lopende bibliografie van de Nederlandse wetenschapsgeschiedenis, een agenda- en nieuwssectie, een uitgebreide verzameling relevante weblinks en een mailinglijst.

Het DWC zal in de komende jaren verder worden ontwikkeld en uitgebreid, onder meer met een 'biografisch apparaat' dat onderzoekers in staat moet stellen om cohortanalyses van wetenschappelijke netwerken te maken. Ook zal de Digital Library met publicaties van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), die nu al via het DWC bereikbaar is, verder worden uitgebreid en geïntegreerd in de website.

Het DWC is opgezet en wordt onderhouden door medewerkers van het Huygens Instituut van de KNAW. (Ilja Nieuwland)

www.dwc.knaw.nl

Een foto als een bel

Cyclorama's zijn foto's die je als een bel helemaal omhullen, samengesteld uit een tweetal afbeeldingen genomen met een fisheye-lens op één punt. Ze worden gemaakt vanuit een auto die maximaal tachtig kilometer per uur rijdt en bieden een betere kwaliteit en een grotere nauwkeurigheid dan de beelden van bijvoorbeeld Google's StreetView. Het bedrijf CycloMedia maakt zulke beelden in opdracht van overheden, voor de WOZ-bepaling of een beoordeling van de bestrating – aldus het boek *Geo-innovatie in Nederland*, dat op 3 september verscheen als resultaat van het vijfjarige project *Ruimte voor Geo-informatie*. Uit dat boek komt ook deze schematische voorstelling. In oktober werd het project afgesloten met een eindrapportage aan het ministerie van VROM en innovatie-agent SenterNovem. Het boek is verkrijgbaar in print en verschillende digitale vormen.

www.rgi.nl www.cyclomedia.nl

Nationaal Georegister

Onlangs werd het Nationaal Georegister gelanceerd. Het brengt informatie over bestaande geo-informatie bijeen in één geïntegreerd portaal. Wie bepaalde gegevens zoekt, kan deze snel filteren met behulp van de gepubliceerde metadata. In veel gevallen is de geo-informatie via het register direct te raadplegen en te downloaden. Het Nationaal Georegister richt zich op de professionele gebruiker. Dat is bijvoorbeeld de specialist op zoek naar datasets, services of andere geo-informatie-elementen. Maar het kan ook een Geo-ict specialist zijn, die een website of toepassing ontwikkelt. Het Nationaal Georegister wordt mogelijk gemaakt door de aanbieders van geo-informatie in Nederland. Dat zijn zowel alle Nederlandse overheden als verschillende onderzoeksinstellingen en bedrijven.
www.nationaalgeoregister.nl

COOL in EASY

De eerste resultaten van het scholierenonderzoek COOL 5-18 zijn beschikbaar gemaakt via het dataarchief EASY van DANS. COOL (Cohortonderzoek Onderwijsloopbanen) is een longitudinaal project rond leerlingen van vijf tot achttien jaar. Van meer dan 70.000 leerlingen wordt de schoolloopbaan gevolgd door het primair en voortgezet onderwijs en het mbo. In drie rondes met vragenlijsten en toetsen worden gegevens verzameld om de ontwikkeling van kinderen tijdens hun schoolloopbanen en hun prestaties te kunnen beschrijven en verklaren.
COOL 15-18 is de opvolger van het PRIMA-cohortonderzoek in het primair onderwijs. Het wordt uitgevoerd in opdracht van NWO en het Ministerie van Onderwijs, Cultuur en Wetenschap.
www.cool5-18.nl
http://easy.dans.knaw.nl

Nederlands Europabeleid

Het Instituut voor Nederlandse Geschiedenis publiceerde in september een eerste serie digitale bronnen over de totstandkoming van het Nederlandse beleid voor de Europese integratie. Het project, tevens een pilot voor een mogelijke pan-Europese bronnenuitgave, wordt gefinancierd door het Ministerie van Buitenlandse Zaken en loopt tot 2014. Het zal dan ongeveer tienduizend documenten bevatten. De documenten zijn totnogtoe vooral afkomstig uit het archief van Buitenlandse Zaken en het Nationaal Archief. Ze zijn beschikbaar als pdf (ge-OCR-de afbeeldingen) en ontsloten op onder meer datum, persoon, trefwoord, documentsoort en samenvatting. www.inghist.nl/Onderzoek/Projecten/Europaeenwording

Hugo de Groot online

Het Huygens Instituut heeft onlangs een basisversie gepresenteerd van de digitale brieven van Hugo de Groot. De meer dan 7500 brieven zijn volledig doorzoekbaar en toegankelijk op jaar, ontvanger en verzender. De site biedt een gedigitaliseerde versie van de zeventiendelige boekuitgave die is gemaakt in de jaren 1928-2001. In de komende jaren zullen de mogelijkheden van de site geleidelijk worden uitgebreid. De correspondentie is een belangrijke bron voor studie van vele aspecten van de zeventiende eeuw.
http://grotius.huygens.knaw.nl/

Bronnenonderzoek vanaf de bank

RONALD VAN DER BIE

Stel: je bent student economische geschiedenis, je moet alleen nog je scriptie maken, je hebt een hockeyblessure en je zit op de bank met een dikke knie. Is het dan mogelijk om thuis, achter de computer, je scriptieonderzoek zo te organiseren dat je voor het arbeidsintensieve bronnenonderzoek de deur niet uit hoeft?

De moderne student komt een heel eind. Sinds enkele maanden geleden het Instituut voor Nederlandse Geschiedenis de nijverheidsenquêtes van 1816, 1819, 1843 en 1848 online beschikbaar heeft gemaakt, is broninformatie over vrijwel alle negentiende-eeuwse bedrijfstellingen digitaal beschikbaar. Behalve de nijverheidsenquêtes waren er al de gedigitaliseerde versies van de Nijverheidstatistiek van Struve en Bekaar (1888–1890) en de volks- en beroepstellingen van de KNAW en het CBS (1795–1971). De enquêtes zijn ge-ocr-d, de statistiek is als Excelfile te downloaden. De volks- en beroepstellingen zijn zowel het een als het ander. Alleen de Staat van de Nederlandse fabrieken (1859) en de Statistiek van de fabrieks- en ambachtsnijverheid in Nederland (1874) wachten nog op bewerking.

De bronbestanden bieden belangrijke informatie. Een economische geschiedenis van Soest of Aalten is op basis van de beroepstellingen van 1899 goed te schrijven. Voorbeelden van dergelijk onderzoek zijn er genoeg. Economisch-historicus Jacques van Gerwen van het Nederlands Economisch Historisch Archief (NEHA) heeft voor zijn onderzoek *Ondernemers in Nederland,* dat hij vorig jaar samen met Ferry de Goey heeft gepubliceerd, de digitale volks- en beroepstellingen intensief gebruikt. Hij noemt de goede digitale toegankelijkheid van deze bronnen 'een geweldige stimulans voor het onderzoek'. De bronnen verdienen volgens Van Gerwen dan ook een breed gebruik, niet alleen voor economisch- en sociaalhistorisch, maar ook voor lokaal en zelfs genealogisch onderzoek. De bundel *Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795–2001* (DANS en CBS, Den Haag 2007) geeft een aantal mooie voorbeelden van onderzoek dat met de bron kan worden gedaan.

IISG/HANS DRIEMAN

Jacques van Gerwen: 'Geweldige stimulans'

Nederlandse situatie riant

Van Gerwen roemt de Nederlandse situatie: 'Mijn indruk is dat Nederland zeker wat betreft de digitalisering van ruim tweehonderd jaar volkstellingen, met vrije toegang voor iedereen, een voorloper is. Buitenlandse data die wij hebben gebruikt voor de internationale vergelijking, vooral na 1945, zijn bij mijn weten vooral bewerkte data. Wellicht zijn er wel gedigitaliseerde oorspronkelijke tellingen, maar ik ken ze niet.' Al is de Nederlandse situatie dan riant, Van Gerwen kan gemakkelijk nog enkele tellingen noemen die het verdienen te worden gedigitaliseerd: de twee genoemde 19de-eeuwse tellingen, de bedrijfstellingen (vanaf 1930), de landbouwverslagen (negentiende en twintigste eeuw) en de productiestatistieken van het CBS (vanaf 1917).

De gepresenteerde datasets geven wel wat problemen bij het gebruik. Van Gerwen werkt het liefst met het originele materiaal omdat je op die manier een beter zicht hebt op de rijen en kolommen van de vaak zeer omvangrijke tabellen. 'Verder is de kwaliteit van de prints vaak slecht en het is niet altijd goed te zien of er een 3, een 6 of een 8 staat.'

Volledig zijn de nijverheidstatistieken ook niet en voor een totaalbeeld moet er nog wel wat onderzoek gedaan worden. Economisch-historicus Michael Jansen gebruikte de gegevens uit de nijverheidsenquêtes van 1819 in zijn onderzoek naar de industriële ontwikkeling van Nederland (1800–1850) slechts voor aanvullende productieschattingen van de wat kleinere bedrijfstakken. Het leeuwendeel van zijn exercitie kwam uit meer robuuste accijns-, in- en uitvoer- en landbouwstatistieken. De nijverheidsstatistiek van de ingenieurs H.W.E. Struve en A.A. Bekaar is uiteindelijk het meest geschikt voor waar de statistiek voor was bedoeld: een indruk geven van de arbeidsomstandigheden, de veiligheidssituatie en de verspreiding van socialezekerheidsregelingen in het midden- en grootbedrijf. De ruim drieduizend bedrijven die de beide heren hebben onderzocht en waarvan het Internationaal Instituut voor Sociale Geschiedenis de gegevens in nette overzichten per provincie presenteert, vormen nog geen tien procent van alle bedrijven. Voor aanvullend archiefbezoek moet de gedupeerde student dus zeker nog achter een paar stevige krukken aan.

Internationale bestanden

De genoemde bronbestanden vormen maar een fractie van wat er inmiddels aan online-onderzoeksmateriaal beschikbaar is. Een uitstekend overzicht biedt de *Virtual library economic and business history*, te vinden op de website van het NEHA. Dat instituut biedt er een schier eindeloze reeks datafiles aan met, bijvoorbeeld, gegevens van historische lonen en prijzen op alle continenten van de wereld. Allemaal zeer goed gedocumenteerd en goed ingeleid, met links naar talloze internationale databestanden, zoals the European State Finance Database (met bijvoorbeeld de boekhoudingen van het Vaticaan in de 16de en 17de eeuw) en de History Data Service. De Nederlandse reeks met lonen en prijzen gaat terug tot 1450, de tabellen zijn voor enkele perioden per provincie uitgesplitst.

www.inghist.nl/Onderzoek/Projecten/Nijverheid1816-1848
www.neha.nl/struve
www.cbs.nl
www.neha.nl/w3vl/index.html

Vragen stellen of alleen luisteren

Met Social Media Monitoring naar een nieuw onderzoekstijdperk?

PETER KANNE

De groeiende rol van de social media op internet maakt dat ze een aantrekkelijk terrein voor onderzoek worden. Hoe verhoudt zich zulk onderzoek tot het traditionele opinieonderzoek?

In opinieonderzoek besteden we veel aandacht en zorg aan de formulering van vragen, de opbouw van de vragenlijst, de volgorde, het wel of niet geven van achtergrondinformatie of noemen van de afzender. Wanneer bijvoorbeeld een standpunt aan Geert Wilders wordt toegeschreven geeft dat andere uitkomsten dan bij de meeste andere politici, weten we uit experimenten. Vragen moeten valide zijn; we willen meten wat we willen meten. De steekproef moet representatief zijn voor de populatie, en de respons zo hoog mogelijk. Alleen dan kunnen we ervan uitgaan dat de onderzoeksuitkomsten een redelijk beeld geven van de meningen van de populatie.

Social Media Monitoring

Nee, dan de Social Media Monitoring, dat is heel andere koek met geheel andere wetten. Met Social Media Monitoring bedoelen we het monitoren van de communicatie op het internet: newssites, blogs, discussiefora, twitter. Dat social media niet langer te negeren zijn als informatiebron, daar twijfelen weinigen aan. Vraag het ex-minister Ella Vogelaar die door *Geenstijl* min of meer eigenhandig aan de kant werd gezet. Vraag het Barack Obama, die mede dankzij de inzet van internet vele harten won en meer dan honderd miljoen dollar ophaalde.

Nieuws is geen lopend vuurtje meer, maar een uitslaande brand. Nieuws op internet is *headlines* en ANP-berichten. Geen plek voor achtergronden, nuances en duiding. Collectieve wijsheid (*citizen journalism*), maar ook collectieve waan ('Bassie wordt pappa!'). Niets van waarde gaat nog verloren (uitzending gemist), maar niets dat waardeloos is evenmin (de parkeergaragezoen van Yolanthe en Wesley). Iedereen heeft wereldwijd en direct toegang tot alle denkbare bronnen en iedereen geeft overal zijn mening over.

Veel van wat we willen weten van burgers is dus gratis van het net te halen. En dat zonder ruis van de opdrachtgever, sturing van de enquêteur of beroepsdeformatie van de onderzoeker. *Gefundenes Fressen!* Veel opdrachtgevers willen geen antwoord op een vraag die we hun klanten kunnen stellen, ze willen weten wat de klanten *zelf* zeggen.

Met Social Media Monitoring, zoals dat bij NIPO gebeurt met het systeem TNS WebLedge, rubriceren we *user generated content*. We volgen trends op het internet: over welke onderwerpen en spelers praat men? In verband met welke onderwerpen en met welke lading? Zo is er bijvoorbeeld veel verschil in *ranking* volgens de officiele websites en newssites (de aanbodkant) in vergelijking met blogs en discussionboards (de web 2.0 kant).

Allemaalpolitiek.nl

Een volgende stap – na het monitoren van Social Media– is het *zelf* creeren van web 2.0 communities. Met onze website www.allemaalpolitiek.nl organiseren we bijvoorbeeld een politieke ontmoetingsplaats die verrassende 'kiezersinformatie' oplevert. Het is een open online politieke arena. Burgers worden uitgenodigd openhartige en diepgaande discussies te voeren over alle denkbare politieke onderwerpen en kwesties. Deelnemers aan allemaalpolitiek.nl kunnen niet alleen met elkaar in discussie gaan, maar ze kunnen ook hun eigen politieke webblog starten, elkaars bijdragen belonen met virtuele cadeaus, video's en foto's delen en nog veel meer. Een NIPO-moderator voedt de discussie op aangeven van een (mogelijke) opdrachtgever.

Voor ons als politieke onderzoekers (maar vooral voor onze klanten) is het heel interessant te zien hoe burgers met elkaar discussiëren en reageren op de input. De informatie die we verzamelen is niet zo spontaan en 'puur' als bij het uitsluitend monitoren van social media, maar is spontaan, eerlijk en creatief genoeg om te kunnen concurreren met kwalitatief onderzoek.

Bij dit alles zijn representativiteit, validiteit en betrouwbaarheid begrippen die in eerste instantie minder relevant zijn. Met social media monitoring signaleren we vroegtijdig trends, nieuwe issues, 'reputatie-vandalisme', reacties op slogans, nieuwe ideeën, etc. Dat deze inzichten daarna weer kwantitatief getoetst moeten worden ligt voor de hand.

Het is dus: luisteren én vragen stellen! En dan weer luisteren...

De wetenschap wordt anders

Het world wide web fungeert als wetenschapsversneller, zegt de Amsterdamse hoogleraar Kennisrepresentatie en Redeneren Frank van Harmelen. Daarmee verandert er veel. Maar dat is nog niets vergeleken bij wat ons te wachten staat.

MARTIJN DE GROOT

‘Natuurlijk heeft het Web ons wetenschappelijke leven veranderd. Als wetenschappers bloggen en hyperlinken we dat het een lieve lust is,’ aldus de informaticus die, heel toepasselijk, op 20 oktober de Diësrede van de Vrije Universiteit mocht houden in aanwezigheid van eredoctor en web-uitvinder Tim Berners-Lee. En hij vervolgde: ‘Maar dat is allemaal oude koek. Ik beweer dat het Web in de nabije toekomst een veel diepgaander invloed op de wetenschap zal hebben dan we op het moment om ons heen zien.’

Twee gevolgen van de beschikbaarheid van het web zullen daarvoor zorgen, licht Van Harmelen bij navraag toe: ten eerste de manier van publiceren en de mogelijkheid om data uit te wisselen, en ten tweede de mogelijkheid om het web zèlf als bron van data te gebruiken. ‘Tot nu toe wordt er gepubliceerd in wetenschappelijke artikelen, die welbeschouwd dienen als een soort ‘staatsbegrafenis voor wetenschappelijke resultaten’. Je maakt hypotheses, tabellen en plots, en vervolgens kan niemand er meer iets mee doen. Je kan de data niet checken want je kan er niet bij. Maar je kan ze ook niet gebruiken voor andere, nieuwe doeleinden. Dat wordt allemaal compleet anders. We kunnen nu de data publiceren zodat iedereen erover kan beschikken. En niet een paar jaar nadat de onderzoeker zijn of haar werk voorlopig heeft afgesloten, maar meteen.’

Als een golf

Daardoor ontstaan oneindig veel nieuwe mogelijkheden, betoogt Van Harmelen, en niet alleen om voort te borduren op het werk van een ander maar vooral ook om data te combineren en daardoor volledig nieuwe onderzoekvragen te beantwoorden. In sommige exacte disciplines zoals de sterrenkunde, de deeltjesfysica of de levenswetenschappen, is deze manier van publiceren al heel gewoon. In andere, zoals de scheikunde en de sociale wetenschappen, nog niet. ‘Maar deze beweging trekt als een golf over de wetenschap’. Wel is er meer nodig dan alleen de mogelijkheid om uit te wisselen, benadrukt de informaticus. ‘Je moet het op zo’n manier doen dat de gegevens voor anderen ook bruikbaar zijn. Het Semantisch Web heeft daarvoor een mooie standaard opgeleverd: het Resource Description Framework of RDF. Met dat RDF kan je van elke dataset de objecten, de variabelen en de relaties daartussen beschrijven. Dat maakt het makkelijker om ze te hergebruiken en vooral om dat door computers te laten doen.’

Het web als observatorium

In de sociale wetenschappen zijn al voorbeelden van projecten die met deze aanpak veelbelovende resultaten bereiken, aldus Van Harmelen. Communicatiewetenschappers laten bijvoorbeeld inhoudsanalyses van berichten in de media uitvoeren door hun computer – in plaats van een legertje studenten of huisvrouwen – en slaan de data op in RDF zodat ze ze gemakkelijk kunnen verbinden met data uit vele andere bronnen. Ander voorbeeld: het aan de VU verbonden Netwerk Instituut bestudeert het wetenschapsproces en heeft daarvoor nu niet alleen citatieanalyses beschikbaar – ‘de dataset daarvan is niet zo groot en loopt ook nog eens vijf jaar achter,’ – maar het hele internet. ‘Wetenschappers doen nog veel meer dan publiceren en citeren. Ze zitten op conferenties, ze bloggen, en dat zie je allemaal op internet dus je kan het ook meten.’ Een soortgelijke aanpak van het web als ‘observatorium’ toont de studie van organisatiewetenschappers aan de VU naar kennisnetwerken tussen bedrijven. ‘Vroeger stuurde je een vragenlijst op en dan mocht je blij zijn als je 30% terugkreeg. Die netwerken zijn tegenwoordig steeds meer op het net te achterhalen. Kijk maar naar LinkedIn.’

Het gaat hard en de perspectieven zijn stralend, vindt Van Harmelen. Maar niet alleen stralend. Het kan ook te hard gaan, citeert hij een bekende uitdrukking: ‘What to do when succes is becoming a problem?’

IRMA BULKENS

Frank van Harmelen

Lezing: Het Semantic Web als wetenschapsversneller

Sprekers op het symposium ‘Door Data Gedreven’

Deze en de volgende pagina’s van e-data&research zijn gewijd aan het symposium ‘Door Data Gedreven’, volgende week in Den Haag. Een terugblik op het eerste data-archief van ons land treft u op pagina 7. Net als het symposium gaat deze bijlage verder vooral over de toekomst van het alfa- en gammaonderzoek. Welke ongedachte mogelijkheden dienen zich aan door het creatief gebruik van reeds verzamelde data? Welke kansen doen zich voor? Lees hier de vooruitblik.

Plaats: De Glazen Zaal, Prinsessegracht 26, Den Haag
Tijd: 2 december 2009, 12.00 – 18.00 uur

Begin november was het symposium zo goed als volgeboekt. Nog beschikbare of vrijkomende plaatsen worden uitgegeven via www.dans.knaw.nl/nl/dans_symposia/, waar u ook het volledige programma treft.

Digitaal onderzoek kan nog veel beter

Wetenschappers, journalisten en onderzoekers maken weliswaar volop gebruik van digitale bronnen, maar ze laten zich daarbij onnodig remmen door de relatief primitieve zoekmogelijkheden op internet. Zonder veel moeite kun je veel geavanceerder en systematischer zoeken, waardoor de kwaliteit van je onderzoek enorm kan toenemen. Dat zegt Ewoud Sanders volgende week in zijn bijdrage aan het symposium ‘Door Data Gedreven.’

Aansluitend bij de Bert van Selm-lezing, die hij eerder dit jaar hield, wil Sanders laten zien hoe je gebruikmakend van openbare bronnen in relatief korte tijd op je eigen pc of laptop een digitale materiaalverzameling kunt aanleggen van honderden miljoenen woorden. Een flexibele verzameling die voor ieder onderzoek aan te passen is en die zeer geavanceerd kan worden doorzocht. Zo kunnen de resultaten chronologisch worden geordend, omgekeerd chronologisch, op relevantie en thematisch. Ook kan, gebruikmakend van indexeringssoftware, met één zoekopdracht razendsnel worden gezocht op allerlei spellingvarianten van een woord of naam, en op talloze woordcombinaties.

Sanders laat een en ander zien aan de hand van een voorbeeld: leven en werk van een jong gestorven dichter die actief was tussen 1880 en 1925. Hij liet enkele dichtbundels, een stapel brieven en een dagboek na. Kunnen wij in grote lijnen de leefwereld van de dichter digitaal reconstrueren – de boeken, tijdschriften en kranten die hij las? En kan ons dit helpen om zijn brieven te annoteren en moeilijk leesbare passages in zijn dagboek te ontcijferen? Ja dat kan. Doordat wij op onze eigen pc zoveel bronnen uit de tijd van de jong gestorven dichter met jokertekens letter voor letter kunnen doorzoeken, blijkt het zelfs mogelijk om woorden in zijn brieven die door inkt- of wijnvlekken grotendeels onleesbaar zijn geworden, te ontcijferen.

Sanders bouwde de afgelopen drie jaar een digitale taalbibliotheek van ruim 4,5 miljoen pagina’s

Ewoud Sanders

en ruim twee miljard woorden, volgens hem momenteel de grootste digitale bibliotheek voor het Nederlandse taalgebied. ‘Geen mens kan zoveel lezen als een computer,’ zegt hij. ‘Mijn computer leest, gebruikmakend van optische tekenherkenning (OCR) ruim vierhonderd boeken per maand. En het geheugen van mijn pc is veel betrouwbaarder dan het mijne: iedere naam, ieder woord, ja zelfs ieder woorddeel is supersnel terug te vinden, tikfouten uitgezonderd. Je eigen digitale materiaalverzameling aanleggen, je ontworstelen aan de huidige beperkingen van internet, heeft mij enorm veel gebracht. Ik kan hierdoor gerichter en beter gestructureerd zoeken dan ooit tevoren – wat vrijwel wekelijks ontdekkingen oplevert die hiervoor onmogelijk waren geweest. En als je een materiaalverzameling aanlegt op basis van openbare bronnen zijn de kosten zo gering dat dit binnen ieders handbereik ligt.’

Lezing: In het hoofd van een vergeten dichter. Een digitale reconstructie.

Sociaal Statistisch Bestand: een voorbeeld uit de praktijk

Het Sociaal Statistisch Bestand (SSB) van het Centraal Bureau voor de Statistiek (CBS) is ontwikkeld om meer themaoverstijgende, longitudinale en gedetailleerde informatie samen te kunnen stellen, aldus projectleider Johan van Rooijen van het SSB.

Inmiddels omvat het meer dan veertig registers met informatie over uiteenlopende terreinen zoals banen, uitkeringen, zelfstandigen, processen-verbaal, opleidingen, woningen en demografie. Gegevens uit verschillende bronnen worden geïntegreerd om ze consistent te krijgen, en daarna op een gestructureerde manier beheerd en ontsloten. Standaardisatie en documentatie zijn daarbij belangrijk. De gegevens uit het SSB worden door veel gebruikers binnen en buiten het CBS gebruikt. Van Rooijen zal die toepassingen illustreren aan de hand van een onderzoek naar de gevolgen van bedrijfseconomisch ontslag. Voor dat onderzoek werden werknemers die in 2003 op die manier hun baan verloren twee jaar gevolgd. Zo kon worden vastgesteld in welke mate zij weer toetraden tot de arbeidsmarkt. De meerderheid blijkt daar toe in staat maar een niet te verwaarlozen deel ondervindt langdurige negatieve gevolgen van het ontslag.

Lezing: Het Sociaal Statistisch Bestand: een veelzijdige bron voor onderzoek

Een digitale bibliotheek van jaarringen in hout

Al zeventig jaar zijn onderzoekers in Europa bezig met de studie van de groeipatronen van hout uit het verleden. Het gaat daarbij om boomsoorten zoals eik, es, beuk, iep, grove den, zilverspar en fijnspar. Ze onderzoeken cultureel erfgoed zoals scheepswrakken, schilderijen, gebouwen en archeologische vondsten (huisplattegronden, waterputten, grafkisten). Daarnaast richten ze zich op in de bodem geconserveerde resten van oude bosvegetaties.De kennis die dat oplevert is van groot belang om de herkomst van (verhandeld) hout te bepalen.

Een belangrijke onderzoeksvraag van deze dendrochronologen is steeds: hoe oud is een stuk hout precies? Met dendrochronologie kun je namelijk vaststellen in welk jaar elke groeiring in hout is gevormd, dus ook in welk kalenderjaar een boom doodging. Dat kan door een ongedateerd jaarringpatroon te vergelijken met absoluut gedateerde dendrochronologische kalenders. Die datering zegt iets over het moment waarop het onderzochte object is gemaakt. Zo is het hout van een Romeinse rivierkade in Leidsche Rijn omgehakt in het voorjaar van 100 n.Chr. De bouw van deze kade zal niet veel later hebben plaatsgevonden.

Er zijn de afgelopen decennia in Europa grote dendrochronologische datasets opgebouwd. De groeipatronen van in Nederland en Noord Duitsland gedateerde bomen bestrijken de laatste acht millennia, waarbij elk jaar van dat enorme tijdsinterval wordt gedekt door waarnemingen. Datasets uit andere regio´s zijn nog omvangrijker. Deze collecties kunnen een belangrijke rol spelen bij interdisciplinair onderzoek naar bijvoorbeeld veranderingen in het landschap en de menselijke omgang hiermee, houthandel en economie, en klimaat.

Voorwaarde is wel dat de collecties gedigitaliseerd zijn en een structuur hebben die ze doorzoekbaar en vergelijkbaar maakt. Veel gegevens zijn echter alleen vastgelegd op papier. Hier is dus een inhaalslag nodig. Datzelfde geldt voor de veelheid aan digitale dataformats die sinds de jaren tachtig van de vorige eeuw is ontwikkeld. Veel formats zijn intussen onbruikbaar geworden voor uitwisseling. Ook worden aan dendrochronologisch dateringsonderzoek (meestal *contract research*) in het algemeen geen kwaliteitseisen gesteld voor de registratie, verduurzaming en ontsluiting van gegevens. Die bevinden zich daardoor vaak in de pc´s van individuele onderzoekers, onderhevig aan digitaal verval.


BERT VAN AS
Esther Jansma

Nederland heeft, met een subsidie van NWO, het voortouw genomen om deze situatie te verbeteren. Laboratoria in België, Duitsland, Frankrijk, Nederland en Polen werken daarbij samen met de Rijksdienst voor het Cultureel Erfgoed en de Universiteit Utrecht. Dertigduizend dendrochronologische meetreeksen van 6000 v.Chr tot nu en metadata worden geverifieerd, opgewerkt en gecombineerd in dit project met als titel *A digital Collaboratory for Cultural Dendrochronology (DCCD) in the Low Countries*. Het project, dat zich met name richt op de Lage Landen, heeft met hulp van Europese en Amerikaanse wetenschappers en ict-deskundigen een internationale standaard opgeleverd voor digitaal archiveren en uitwisselen: de *Tree-Ring Data Standard* TriDaS. Deze standaard blijkt internationaal aan te slaan. Nog onlangs heeft een belangrijk Amerikaans instituut voor dendrochronologisch onderzoek de eigen collectie naar TRiDaS omgezet, waardoor uitwisseling met het DCCD mogelijk wordt. Andere Amerikaanse instellingen zoeken naar fondsen om hetzelfde te doen. Het TriDaS-model vormt de basis van de dataopslag- en webapplicatie die nu wordt ontwikkeld door DANS, dat de gegevens opslaat volgens de normen van het *Data Seal of Approval*. De DCCD, die eind 2010 gereed zal zijn, biedt een viertalige gebruikersschil waardoor onderzoekers nieuwe gegevens kunnen toevoegen, oude projecten kunnen herbewerken en zoekacties uitvoeren. Daarmee wordt het een levend, beweeglijk en steeds actueel archief.

In Europa bestaat binnen het vakgebied sterke belangstelling om deze infrastructuur uit te breiden. Een belangrijke reden is dat het DCCD aan dataleveranciers de mogelijkheid biedt om eigen gegevens ten dele af te schermen. Dit is vooral relevant voor onderzoekers in de private sector, die hun gegevens gebruiken als referentie bij dateringsonderzoek en hun gegevens niet willen ontsluiten voor concurrerende bedrijven, maar wel geïnteresseerd zijn in wetenschappelijke samenwerking. Het is voor het eerst dat deze groep, die zich altijd heeft verzet tegen de ontsluiting van gegevens, bereid is mee te denken over de digitale koppeling van collecties.

Esther Jansma is bijzonder hoogleraar in Utrecht en onderzoeker bij de Rijksdienst voor Cultureel Erfgoed.

Lezing: Duizenden jaren houtgebruik: een dendrochronologisch repository voor de Lage Landen (6000 v.Chr. - heden)

www.dendrochronology.eu
www.ncdc.noaa.gov/paleo/tree-ring.html
www.tridas.org

Hoe vrouwvriendelijk is de Tweede Kamer?

MAARTEN MARX

Ruwe tekst als onderzoeksmateriaal voor tekstanalyse door sociale wetenschappers komt in steeds grotere hoeveelheden beschikbaar, bijvoorbeeld via internet. Maar om van tekst in allerlei formaten naar een mooie invoerfile voor het gangbare rekenpakket SPSS te komen is vaak nog een hele stap. Toch zijn de hulpmiddelen daarvoor beschikbaar.

Een recent artikel in *Science*, genaamd 'Computational Social Science', pleit voor een curriculum waarin studenten aan de alfa en gamma faculteiten *tools* leren gebruiken om enorme hoeveelheden tekst te kunnen verwerken. Aan de hand van een voorbeeld kunnen we zien wat die tools inhouden, en dat de omgang ermee helemaal niet zo moeilijk is. Computers zijn tegenwoordig zo krachtig en eenvoudig geworden dat ook echte gamma-infobeten kwantitatief onderzoek op basis van gigantische hoeveelheden tekst kunnen uitvoeren. De kennis die daarvoor nodig is kan in een vak 'Tekstanalyse voor de sociale wetenschappen' van bescheiden omvang aan elke student geleerd worden.


ANP
CDA-parlementariër Mirjam Sterk heeft in de Tweede Kamer de aandacht van drie bewinsdslieden

Laten we ons richten op de volgende onderzoeksvragen. Er zit nu een recordaantal van meer dan veertig procent vrouwen in de Tweede Kamer. Zijn dat nu de bekende Excuus-Truusen of zijn ze ook evenredig veel aan het woord? En vervolgens: verschilt dat per onderwerp, of tussen de partijen? Hoe zat het vroeger?

Alle data om deze vragen te beantwoorden zijn aanwezig. De Handelingen der Staten Generaal zijn als pdf-bestanden op het internet beschikbaar vanaf 1917. Ze bevatten exact wat iedereen in de Tweede Kamer gezegd heeft. Daarnaast bevat de website *parlement.com* voor iedereen die ooit in de Kamer heeft gezeten een uitgebreide biografie. We zouden dus gewoon kunnen gaan turven hoe vaak elk parlementslid aan een debat heeft meegedaan, hoe vaak hij of zij heeft geïnterrumpeerd, aan het woord is geweest en hoe lang dat lid op de spreekstoel heeft gestaan. Tijden staan weliswaar niet vermeld in de Handelingen, maar die kunnen we benaderen door het aantal gesproken woorden te tellen.

Tot zo ver lijkt het dus niet moeilijk. Toch is de uitvoering niet eenvoudig omdat de data niet in het juiste formaat beschikbaar zijn. Er zijn drie problemen. In de eerste plaats gaat het om heel veel data: 3560 biografiën en meer dan honderd miljoen woorden, gesproken in de Tweede Kamer vanaf 1995. In de tweede plaats is de koppeling van de twee datasets moeilijk omdat er niet consequent met dezelfde namen naar parlementsleden verwezen wordt. Dit probleem is des te erger met data van voor 1995, die zijn ingescand en nog foutjes bevatten door het gebruik van optische tekenherkenning OCR. En ten slotte bestaan de Handelingen uit tekstbestanden in pdf-format, met summiere metadata. Voor elk woord in de Handelingen weten we wel op welke dag het is uitgesproken en op welke bladzijde het staat, maar niet door wie en in welke hoedanigheid: als parlementslid of lid van de regering; als betoog, interruptie of antwoord op een interruptie.

Nadat de Handelingen in een machineleesbaar formaat zijn gebracht kunnen technieken voor tekstanalyse het probleem van de herkenbaarheid en hoedanigheid van de sprekers oplossen. Met *named entity recognition* and *reconciliation* kunnen we sprekers herkennen, hun naam normaliseren en zo de hindernissen voor het combineren van de twee datasets opruimen. Op dat moment kunnen we computers inschakelen om het probleem van de grote hoeveelheid data op te lossen. We hoeven dan niet met dure codeurs en steekproeven te werken, en kunnen de analyse op de gehele dataset uitvoeren. De Universiteit van Amsterdam maakt op dit moment in samenwerking met de Koninklijke Bibliotheek de Handelingen beschikbaar in een XML formaat waarmee de hiervoor gestelde onderzoeksvraag echt eenvoudig op te lossen is.

Ondertussen hebben we bij wijze van voorproefje al wat ruwe tellingen gedaan om de vrouwvriendelijkheid te meten. Als we de voorzitter niet meetellen is in de periode van 3 februari 2009 tot en met 8 oktober 2009 33% van de spreektijd in de Tweede Kamer door vrouwen gebruikt. Bijna 20% minder dan je zou verwachten op basis van het aantal vrouwelijke leden. Slechts 30% van de tijd staat er een vrouw op de spreekstoel. Dit kan natuurlijk nog van alles betekenen. Misschien zijn vrouwen wel minder langdradig en geven kortere antwoorden op interrupties dan mannen.

Er is enorm veel geinvesteerd in taal-technologie tools voor het Nederlands. Echter, voor 'gewone' alfa's en gamma's zijn die tools nog vaak erg moeilijk toepasbaar, zeker als ze in combinatie gebruikt moeten worden. Dit is zonde want er is een schat aan ruwe data vrij beschikbaar. Het hiervoor aangehaalde Science-artikel beschrijft het reële gevaar dat de industrie (Google, Amazon, etc) het vakgebied 'Computational Social Science' voorgoed voor de neus van de wetenschap wegkaapt. Laten we zorgen dat dat niet gebeurt.

Binnenkort kan iedereen het genoemde onderzoek zelf uitvoeren, want we plaatsen alle Handelingen in een uniform XML-formaat in het EASY archief van DANS. En elke dag wordt het aangevuld.

Maarten Marx is politicoloog en informatie-wetenschapper en doceert aan de Universiteit van Amsterdam.

Lezing: Tekstanalyse voor de Sociale Wetenschappen

Het Steinmetzarchief: geboren uit een hausse aan veldonderzoek

MARTIJN DE GROOT

Vijfenveertig jaar geleden werd binnen de alfa- en gammawetenschappen het eerste data-archief in Nederland opgericht. Sindsdien is er veel veranderd, maar ook veel hetzelfde gebleven, blijkt uit een terugblik.

De oprichters van het latere archief gaven hun schepping in 1964 met opzet de naam Steinmetz Stichting. Het gaat hier niet in de eerste plaats om oorspronkelijke documenten maar om afgeleid materiaal zoals ponskaarten, codeboeken en interviewprotocollen, redeneerden ze. Daarom vermeden ze de naam 'archief' in naam en doelstelling.

'Als hun onderzoek was afgesloten dachten onderzoekers: dat is klaar en nu gaan we aan ons volgende project beginnen. Dan voelden ze geen aandrang om hun data nog eens uitvoerig te gaan documenteren om te worden gearchiveerd.' Wie deze woorden hoort zou ze gemakkelijk kunnen toeschrijven aan een medewerker van DANS, Data Archiving and Networked Services, dat tegenwoordig is belast met het duurzaam toegankelijk maken van onderzoekgegevens. Mis. Aan het woord is Harm 't Hart, emeritus hoogleraar Methoden en Technieken aan de Universiteit Utrecht, en hij beschrijft de weerstand die hij in de jaren zestig moest overwinnen om datasets van onderzoekers los te krijgen. 't Hart was namelijk in die tijd coördinator van de *Steinmetz Stichting voor het Opslaan en Toegankelijk maken van Bestaand Materiaal van Sociaal Onderzoek*. Die stichting, de voorloper van het latere Steinmetzarchief, was op 27 november 1964 opgericht vanuit de zogenaamde zevende faculteit van de Universiteit van Amsterdam om te voorkomen dat waardevolle onderzoekgegevens verloren zouden gaan.

Gebouwd op surveys

De jaren vijftig en begin zestig hadden een enorme opbloei laten zien van wetenschappen die zich voor een groot deel baseerden op surveys: sociologie, politicologie, psychologie. De na-oorlogse eensgezindheid om het fascisme voorgoed buiten de deur te houden en de invloed van de Verenigde Staten, waar de kwantitatief georiënteerde gammawetenschappen sterk in opkomst waren, gaven die ontwikkeling de wind in de zeilen. De sociologie van de jaren vijftig en zestig was bijna gebouwd op surveys, maar ook de politicologie en belangrijke stromingen in de psychologie maakten er gretig gebruik van, bijvoorbeeld voor het ontwikkelen van allerlei schalen om houdingen en talenten in beeld te brengen. In het kielzog van die ontwikkeling gingen er stemmen op om ervoor te zorgen dat al die gegevens niet alleen door de eerste onderzoeker gebruikt zouden kunnen worden, maar ook door degenen die na hem of haar kwamen. Na de Verenigde Staten waren in Europa Duitsland, Frankrijk, het Verenigd Koninkrijk en Denemarken al een eind op weg toen in Nederland de eerste stappen werden gezet.

Archief voor gehele wereld

Initiatiefnemers waren prof. H.M. Jolles vanuit de Sociaal-Wetenschappelijke Raad en drs. M. Brouwer van het aan de UvA verbonden Seminarium voor Massapsychologie, Propaganda en Openbare Mening. Laatstgenoemde was ook degene die de toen pas afgestudeerde Harm 't Hart vroeg om bij de nieuwe stichting te komen werken. 'Die internationale ontwikkeling was een belangrijke drijfveer voor Brouwer en Jolles om ook in Nederland naar een archief te streven', zegt 't Hart nu. Vanuit Amerika werd dat ook bepleit. Volgens een notitie van beide mannen uit 1964 streefde bijvoorbeeld het Roper Public Opinion Research Centre, genoemd naar opinieonderzoekbureau Elmo Roper, ernaar 'om het centrale archief te worden voor surveyonderzoek over de gehele wereld' en overwoog het 'de oprichting van een aantal regionale centra in bijvoorbeeld West-Europa, Oost-Azië en Zuid-Amerika'. Behalve die internationale druk vormde ook de behoefte om studenten aan proefmateriaal te helpen een belangrijk motief. Die moesten immers ook worden getraind in het gebruik van de moderne kwantitatieve methoden.

Met commercieel onderzoek

Opvallend aan de nieuwe initiatieven, zowel in Nederland als in het buitenland, was dat er werd samengewerkt tussen universitair en commercieel onderzoek. 'Dat was in die tijd ook wel een soort ideaal,' herinnert 't Hart zich, 'Het Roper Instituut, dat nog steeds bestaat en nu is verbonden aan de Universteit van Connecticut, was toen volledig commercieel net als opinieonderzoeker Gallup, die zijn materiaal ook ter beschikking stelde.' In Nederland hield dat streven niet lang stand. Opinieonderzoekers als het NIPO, de Nederlandse Stichting voor Statistiek en Intomart hadden wel belangstelling, maar verbonden zich toch niet aan de op te richten stichting. 'Ze waren ook bang dat hun opdrachtgevers bezwaar zouden maken,' aldus 't Hart. Anderen zouden immers kunnen meeprofiteren van de investeringen die zij hadden gedaan.

Tel- en sorteermachines

Hoewel de overeenkomsten met de huidige tijd zich hier en daar opdringen, is er ook een groot verschil tussen de situatie in de jaren zestig en die van nu. Op het moment van de oprichting van de Steinmetz Stichting speelde de computer namelijk geen rol van betekenis. 'Het enige wat vastlag dat waren de enquêtes,' herinnert 't Hart zich. 'Die werden omgezet in codeformulieren. In het marktonderzoek en het universitaire onderzoek werden daarvan ponskaarten gemaakt. We hadden de beschikking over tel- en sorteermachines, die die kaarten konden tellen. Dan had je bijvoorbeeld een enquête met tweeduizend respondenten. Dat waren tweeduizend ponskaarten. Die kon je dan op geslacht sorteren en dan had je twee stapeltjes.' In dat proces, dat kon oplopen tot twaalf kaartenbakjes, konden de afgenomen enquêtes steeds opnieuw gesorteerd en geteld worden – iedere keer voor nieuwe variabelen of voor een ander onderzoek. Opmerkelijk is dat de pleidooien voor een goedwerkend archief, waarin tegenwoordig de computer en het internet een belangrijke rol spelen, vijfenveertig jaar geleden zonder die twee revolutionaire hulpmiddelen eigenlijk op dezelfde argumenten berustten als nu.

Leuren

Maar de ontwikkeling ging in de tweede helft van de jaren zestig heel snel. Drie jaar na de oprichting van de stichting bezocht 't Hart bijvoorbeeld een conferentie in Los Angeles die gewijd was aan 'het op moderne wijze gebruiken van computers bij de analyse van complexe verzamelingen van gegevens van sociaal-wetenschappelijk onderzoek'. In zijn verslag over die conferentie noemde hij een ontwikkeling waarvan ook nu nog veel wordt verwacht: 'Het kwantitatieve sociaal-wetenschappelijk onderzoek gaat duidelijk in de richting van geïntegreerd gebruik van bestaand materiaal en van nieuw verzamelde gegevens in plaats van het gebruik van één bron dan wel van meerdere bronnen zonder dat integratie tot stand komt'. Dat is mogelijk geworden door het ontstaan van archieven, voegt hij eraan toe, 'en door de ontwikkeling van de computertechniek'. Toch was het geloof in de vooruitgang misschien groter dan door de realiteit werd gerechtvaardigd. Er is in die eerste jaren veel energie gestoken in het werven van gegevens, weet 't Hart: 'Het was een beetje leuren. Ik ben bij al die instituten langs geweest. Universiteiten, particuliere instellingen, overheidsinstellingen. Men vond de voordelen wel belangrijk *als ik het zei*, maar het viel toch niet mee om actieve medewerking te krijgen.'

Ook rond het gebruik van de opgeslagen gegevens ontstonden bepaald geen opstoppingen. Beduchtheid om andermans fouten mogelijk te reproduceren speelde daarbij een belangrijke rol. Een schrijf- of codeerfout van de oorspronkelijke onderzoeker of diens medewerkers zou bij hergebruik onbewust gerepliceerd kunnen worden. 't Hart: 'De grote rol van de computer en van internet hebben wel enorme vooruitgang gebracht. Destijds werden de data wel gebruikt maar alleen door een paar mensen, terwijl wij toch het idee hadden dat daar veel meer onderzoekers van zouden kunnen profiteren'. Die paar mensen werkten vooral bij de Universiteit van Amsterdam en dat leidde weer tot het beeld bij anderen, dat de Steinmetz Stichting eigenlijk in de eerste plaats een zaak voor die universiteit was.

Complete Volkstelling

Mede om die reden werd de stichting in 1971 ondergebracht bij de Koninklijke Nederlandse Akademie van Wetenschappen. Bij die gelegenheid werd de naam veranderd in Steinmetzarchief. 't Hart bleef er werk voor verrichten en had op dat moment een tweehonderdtal datasets bij zich. Vijf jaar later, in 1976, waren dat er 650 geworden, waaronder driehonderd NIPO-weekenquêtes, psychologisch testmateriaal, gemeentelijke statistische gegevens èn – de trots van Harm 't Hart destijds – de complete Volkstelling van 1960. Het archief had zich volgens een artikel van zijn medewerker C.P. Middendorp verzekerd van uitstekende computerfaciliteiten: 'Het beschikt over een terminal-aansluiting op de CDC-Cyber 73-28 computer van het Academisch rekencentrum Amsterdam. De terminal bestaat uit een Data-point 220 mini-computer en een Tally line-printer. Ponswerk wordt uitbesteed aan het Mathematisch centrum en aan het bureau Gamma-Routine'.

HERBERT WIGGERMAN

Harm 't Hart, met op de achtergrond in de steigers het pand Herengracht 457 waar hij lange tijd zijn werk deed als coördinator van de Steinmetz Stichting: 'De behoefte om studenten aan proefmateriaal te helpen was een belangrijke drijfveer achter de oprichting'.

Een nieuwe kijk op de historische handel met de West

MARTIJN DE GROOT

Lange tijd concentreerden economisch historici zich voor wat betreft de internationale handel van Nederland op 'de Oost'. Het belang van de Atlantische handel werd niet hoog aangeslagen. Dat had te maken met de misvatting dat Suriname economisch niets voorstelde, maar vooral ook met het ontbreken van overzichtelijke bronnen, zegt historicus Henk Den Heijer. Maar dat gaat nu veranderen.

'De focus was in de economische geschiedschrijving heel lang gericht op de Oost. Het belang van de handel in die richting werd veel groter ingeschat dan dat van de Atlantische handel,' zegt Den Heijer, die lid is van een onderzoekteam dat met hulp van NWO de *Dutch Atlantic Connections* in de zeventiende en achttiende eeuw in de schijnwerpers moet krijgen. 'En het valt niet te ontkennen: Nederlands Oost-Indië was de belangrijkste kolonie in termen van economisch profijt. Suriname werd gezien als een verliesverhaal, dat in de negentiende eeuw was weggezakt naar een onbelangrijk niveau. Maar die vergelijking gaat voor de achttiende eeuw mank. Bovendien was de handel met Suriname maar één onderdeel van het geheel van Nederlandse handelsactiviteiten met Afrika en Amerika. Nederlandse schepen kwamen in heel veel havens aan die kant van de wereld en bewogen zich ook daartussen. Ze onderhielden een netwerk van handelsrelaties dat een groot economisch en cultureel belang vertegenwoordigde.' Zonder politiek of economisch imperium hielden ze met z'n allen de machine van de Atlantische handel draaiende. Dat dat imperium tot nu toe nog niet de aandacht kreeg die het verdiende, blijkt vooral een kwestie van de beschikbare bronnen. Daarin gaat het project van Den Heijer en zijn collega's verandering brengen door het bijeenbrengen van de gegevens van vier knooppunten in het westelijke handelsnetwerk: Paramaribo, Curacao, Amsterdam/Rotterdam en het West-Afrikaanse Elmina.

Flarden samenvoegen

'De Verenigde Oostindische Compagnie had een monopolie en daarvan is een uitstekend archief bewaard gebleven. Dat bevat alle gegevens die je maar wil hebben en daar hebben zich dus vele instituten en onderzoekers in de loop der tijd op gericht. De West-Indische Compagnie daarentegen was een economische mislukking, waarvan ook nog eens het archief grotendeels door de papiermolen is gegaan. Maar de handel met de West had ook een heel ander karakter. Die werd vooral bedreven door kleine spelers, particuliere ondernemers. Om daar inzicht in te krijgen moet je dus zien dat je gegevens uit heel veel kleine bronnen tevoorschijn haalt, en vervolgens dat je ze met elkaar in verbinding brengt om inzicht in het gehele netwerk van relaties te krijgen. Dat zijn we nu aan het doen.'

Het kan gaan om notariële archieven, legt de Leidse historicus uit, of om archieven van zelfstandige ondernemers die op allerlei verschillende plaatsen opgedoken moeten worden. Ook een bron als de Middelburgse Commercie Compagnie, waarvan het Zeeuws Archief nog veel documenten heeft, levert belangrijke informatie op, bijvoorbeeld ladinglijsten en logboeken van vertrekkende schepen die aangeven met welke lading op welke havens wordt gevaren. 'Je moet op veel verschillende plaatsen zoeken om aan de goede informatie te komen. Veel stukjes zijn trouwens al boven water gehaald, bijvoorbeeld op het gebied van de slavenhandel door de Amerikaanse Nederlander Johannes Postma. Daar is ook een databank van die te raadplegen is op www.slavevoyages.com.' Maar het belangrijkste om de echte onderzoeksvragen te beantwoorden, is het elektronisch beschikbaar maken van al die losse flarden en het samenvoegen ervan tot één doorzoekbaar en bewerkbaar geheel. Deze zogenaamde relationele database is intussen al zo ver dat hij gedeeltelijk bij DANS is gedeponeerd en zo beschikbaar gemaakt voor andere onderzoekers. Het project, waarin bestaande en nieuwe databestanden worden samengevoegd, duurt nog tot 2013. 'De database van het Nederlands-Atlantische scheepvaartverkeer gaat zeker leiden tot veel nieuw onderzoek en nieuwe publicaties', aldus de Leidse historicus.

Gezicht op het West-Afrikaanse eiland Elmina, 17e eeuwse kaart uit de Atlas Blaeu van der Hem

Lezing: Schepen, mensen en goederen (Nederlands scheepvaartverkeer in het Atlantisch gebied, 1600 – 1800)

www.kitlv.nl/home/Projects?id=19
http://easy.dans.knaw.nl

Trekvogels in beeld met anderhalve Mb per seconde

MARTIJN DE GROOT

'Geen modellen zonder metingen, geen metingen zonder modellen.' Willem Bouten, hoogleraar Computational Geo-Ecology aan de Universiteit van Amsterdam, gelooft in modellen om de bewegingen van dieren op en boven het aardoppervlak te voorspellen, en hij gelooft in metingen om die modellen mee te bouwen. Heel veel metingen.

DIEDERIK VAN DER LAAN

Willem Bouten in zijn virtuele laboratorium: 'Wij gebruiken gegevens over de vogeltrek van allerlei radars in Europa'

Alles komt samen in het Virtual lab for eScience (VLe), waar Bouten waarnemingen uit de natuur vanuit een veelheid aan bronnen in Nederland en daarbuiten verzamelt. 'Wij maken hier gebruik van de nieuwste kennis en gereedschap uit de informaticawereld, en van het steeds maar groeiende vermogen om via glasvezelkabels enorme hoeveelheden data snel te verplaatsen.' Hij wil wel een voorbeeld noemen. 'Ik kan nu hier in mijn *virtuele laboratorium* gegevens over de vogeltrek gebruiken die door allerlei radars in Europa met anderhalve megabyte per seconde worden uitgespuugd. Dat kon een paar jaar geleden nog niet'.

Bouten begeeft zich met zijn modellering op een gebied dat eigenlijk precies tussen de twee oude tegenpolen in de ecologie in ligt, legt hij uit. 'Je hebt theoretisch ecologen, die werken met wiskundige vergelijkingen van het gedrag van dieren zonder ze van dichtbij te zien. En je hebt veldecologen, die waarnemingen doen in het veld met als voornaamste doel om de dieren te beschermen. Ik maak modellen om de bewegingen van dieren te voorspellen, waarvoor ik mijn laboratorium niet uit hoef. Maar ik gebruik daarbij van alle waarnemingen uit het veld die ik maar kan krijgen. Geen modellen zonder...'

Modellen kunnen gebruikt worden om te voorspellen. Bouten en zijn groep doen dat ook: het levert hen feedback op over de kwaliteit van hun modellen èn geld om ze verder te verbeteren. Bouten: 'Voor bouwend Nederland kijken we of en waar er natuur-compenserende maatregelen moeten komen. In het verleden zei men gewoon: daar zit deze of die zeldzame soort en dan moest een heel bouwproject stilgelegd worden. Om zulke beslissingen te onderbouwen is de Gegevensautoriteit Natuur in het leven geroepen, de GAN. Voor die GAN hebben we een informatiesysteem gemaakt waarmee we de verspreiding van soorten kunnen aangeven. Daarmee kunnen ze een onderbouwd oordeel vormen over de impact van een bouwproject voor zeldzame planten of dieren. De gegevens halen we uit een hele reeks bronnen. Je hebt de website waarnemingen.nl, maar er zijn in Nederland ook twintigduizend geregistreerde vrijwilligers die waarnemingen doorgeven aan vrijwilligersorganisaties als Vlinderstichting, Floron en Sovon. Maar ook Natuurmonumenten, Staatsbosbeheer en gemeenten en provincies verzamelen gegevens voor beheer en bescherming.'

In het virtueel laboratorium worden al die data samengebracht. Eerst gebeurde dat alleen voor wetenschappelijk onderzoek, maar al snel bleek dat het systeem zich ook goed leende voor uitspraken over de verspreiding van bepaalde soorten voor de bouwwereld.

Inzicht in vogeltrek stelt Bouten en zijn team in staat om tot maximaal twee dagen van te voren voorspellingen te doen over de intensiteit van vogeltrek, waar de Luchtmacht veel aan heeft bij het plannen van vliegoefeningen. Het *last minute* afgelasten van zo'n oefening kan al snel in de tonnen lopen aan zinloos gemaakte kosten, en het laten doorgaan terwijl er vluchten trekvogels passeren tot ernstige ongelukken. De ecologen maken daarbij gebruik van twee heel verschillende soorten metingen, die vooral in combinatie tot de gewenste resultaten leiden. Bouten: 'Vanuit mijn laboratorium kan ik de militaire radars van de Nederlandse en Belgische strijdkrachten raadplegen. We hebben ook de beschikking over meteorologische radars, die zijn ontwikkeld om windsnelheden te meten. Je ziet daarop tot miljoenen vogels door het beeld vliegen maar je weet niet wat voor soorten het zijn. Daarnaast hebben we zo'n vijftig vogels uitgerust met GPS-zendertjes, waarvan we dus de individuele bewegingen kunnen volgen. Die twee waarnemingen koppelen we hier aan elkaar en in combinatie met andere, bijvoorbeeld meteorologische gegevens, levert dat de voorspellingen van vogelstrek op dagelijks door de luchtmacht en door vogelaars gebruikt worden.

Lezing: e-Ecologie: combinatie van natuur en techniek

www.gegevensautoriteitnatuur.nl/pages/nieuws.aspx
https://ndff-ecogrid.nl/
www.ecogrid.nl/nl/info/ecogrid/
http://public.flysafe.sara.nl/bambas/

Peter Doorn, drijvende kracht in hergebruik van onderzoekgegevens:

‘Ik vind alles interessant, als er maar data ter beschikking komen’

HANS VAN MAANEN

Het Steinmetzarchief, het Nederlands Historisch Data Archief, het Wetenschappelijk Statistisch Agentschap, het Electronisch Depot voor de Nederlandse Archeologie: allemaal zijn ze inmiddels verenigd in DANS, Data Archiving and Networked Services. En één naam duikt in die archieven met een opmerkelijke regelmaat op: die van Peter Doorn. Wie is deze drijvende kracht in, zoals hij het zelf wat ironisch noemt, ‘de handel in tweedehands data’? Een interview.

‘Twee dingen vielen mij op toen ik met mijn proefschrift bezig was, en allerlei gegevens uit de Arbeidskrachtentelling wilde koppelen aan het Woningbehoefte-onderzoek.

Het eerste was dat het gebruik van die bestanden zo ontzettend duur was. Voor die Arbeidskrachtentelling moest een ton betaald worden – in guldens, we spreken over begin jaren tachtig, toen ook het profijtbeginsel volop speelde. Het tweede, veel belangrijker, was dat die bestanden eigenlijk amper te combineren waren. Er was geen koppeling of vergelijking van data mogelijk, omdat ze steeds hun eigen indelingen en definities hadden gekozen. Ik heb toen heel wat moeten puzzelen en programmeren om er toch uit te krijgen wat ik wilde – en ik had daar blijkbaar zo’n aardigheid in, dat het mijn vak is geworden. Ik had er in ieder geval, als sociaalgeograaf, nog voor mijn promotie meteen mijn baan bij geschiedenis in Leiden aan te danken: daar ging ik het vak ‘computertoepassingen in de geschiedenis’ doceren.’

‘We zijn zo aan gewend dat onze privé-gegevens bekend zijn dat we onze privacy blijmoedig te grabbel gooien’

‘De meeste historici vonden het natuurlijk maar niks, al die computers, maar er waren er ook die wel degelijk de mogelijkheden ervan inzagen. Zo konden we in 1989 beginnen met het aanleggen van het Nederlands Historisch Data Archief. Net als de Steinmetzstichting vijfentwintig jaar daarvoor haalden we overal oude ponskaarten en magneetbanden met onderzoeksgegevens vandaan om die te ‘ontsluiten’.

Een van de eerste echt grote bestanden die dat Steinmetzarchief had gekregen was van de Volkstelling van 1960: dat was de eerste waarbij de computer werd ingeschakeld. Weliswaar nog met ponskaarten – het CBS moest daarvan af, en ze werden toen onder verre van optimale omstandigheden opgeslagen in een ruimte van de Universiteit van Amsterdam – dus dat heeft later nog heel wat hoofdbrekens opgeleverd, maar dat soort data was wel een prachtige basis om te beginnen.

Pas veel later hebben we digitaal reddingswerk verricht aan de bestanden van de Volkstelling van 1971. De telling van 1981 ontbreekt: er was toen zoveel verzet uit de bevolking tegen ‘Big Brother’ dat men bang was dat door het aantal weigeraars het hele onderzoek zinloos zou worden. Later is de Wet op de Volkstelling aangepast, en nu worden er geen grote officiële Volkstellingen meer gehouden, tenminste niet op de klassieke manier.’

‘Dat verzet tegen die Volkstellingen had te maken met de vrees voor privacy – maar het interessante is dat al die data waar toen zoveel zorg over bestond, in feite nu nog veel nauwkeuriger bekend zijn. Door de koppeling van bestanden. Het CBS weet letterlijk alles van je – waar je woont, wat je doet, wat je verdient, je afkomst. Een ouderwetse volkstelling is daardoor helemaal niet meer nodig.

Ik moet er wel bij zeggen dat het CBS buitengewoon zorgvuldig is met zijn bestanden – alleen met een chipkaart en een vingerafdruklezer kun je vanuit een beveiligde ruimte op een speciale pc op de faculteit bij de CBS-bestanden op individueel niveau. En je krijgt je analyseresultaten pas nadat het CBS heeft gecontroleerd dat er geen data in zitten die tot individuen herleidbaar zouden kunnen zijn.

Grappig is dat de opvatting over privacy onder de bevolking ondertussen lijkt te zijn verschoven. We zijn er blijkbaar zo aan gewend geraakt dat onze privé-gegevens bekend zijn bij de Airmiles, bij *telemarketeers* en bij Google, dat we onze privacy blijmoedig te grabbel gooien. Het is onvoorstelbaar wat mensen geheel vrijwillig aan persoonlijke informatie op weblogs, Facebook en andere sociale netwerken zetten. En de zorgverzekeraars kunnen op sommige *community sites* zo hun risicogroepen van lijders aan nare ziekten traceren, daar hebben ze het elektronisch patiëntendossier straks niet eens voor nodig.’

‘Er wordt zo gehamerd op de verhoging van de AOW-leeftijd, terwijl dat maar een van de mogelijke oplossingen is‘

‘Het Historisch Data Archief werd in 1995 een instituut van de KNAW, en in 1997 samen met het Steinmetzarchief onderdeel van het NIWI. Ik wilde toen eigenlijk afscheid nemen van het data-archiveren en weer onderzoek gaan doen aan de universiteit, maar nadat bij het NIWI troebelen uitbraken en de hele leiding werd vervangen, besloot ik daar toch te solliciteren als afdelingshoofd. Het NIWI was een samenvoeging van zes instituten en dat was achteraf – en volgens sommigen ook al vooraf – niet zo’n goed idee, nee. Er waren toch te veel verschillende culturen, en er ontstond niet het krachtige instituut voor wetenschappelijke informatie dat bedoeld was. Er is door het interne gesteggel veel tijd verloren gegaan, juist in een periode dat het digitaliseren een hoge vlucht nam. Er is natuurlijk wel het een en ander tot stand gebracht, maar we hebben, vind ik, toen niet de grote sprong voorwaarts gemaakt die mogelijk was. In zekere zin zijn we toen overvleugeld door de Koninklijke Bibliotheek, die heeft toen het voortouw kunnen nemen.

Toen in 2005 DANS werd gevormd, was dat na het NIWI toch wel een bevrijding. Bij DANS gaat het nu simpelweg om toegang tot data voor de wetenschap, nu en op de lange termijn. We hebben in een paar jaar tijd een duidelijke positie opgebouwd, ook ten opzichte van de KB. Bij de KB staat de authenticiteit van de documenten voorop, wij zijn vooral geïnteresseerd in de herbruikbaarheid van de gegevens. Onderzoekers kan het meestal niet schelen in welk lettertype een dataset is gezet, of in welk handschrift een scheepsjournaal is bijgehouden, het gaat ze bijna altijd primair om de data zelf, om de informatie-inhoud. Die willen ze met moderne software te lijf kunnen.’

‘Een mooi project was dat trouwens, met die oude scheepsjournaals. Dat hebben we samen met het KNMI gedaan: alle weergegevens die in oude logboeken, van ongeveer 1750 tot 1850, waren vermeld, hebben we nageplozen. Het weer op die schepen werd elke dag heel nauwkeurig bijgehouden, en die database kun je prachtig gebruiken om het klimaat te reconstrueren voor een tijd dat er nog geen weerstations waren. Meer dan 300.000 waarnemingen zijn ingevoerd, en omdat Engeland en Spanje ook meededen, leverde dat een goed beeld van het weer op alle oceanen. Dat is toch waar het om gaat, vind ik, met oude data nieuwe dingen doen. Er zit in die scheepsjournaals nog zoveel meer: we weten met welke snelheid ze voeren, welke routes – je kunt ze zo op Google Maps uittekenen.

Van alles houden mensen gegevens bij, en die wil DANS beschikbaar maken voor wetenschappelijke analyse. Slavenhandelaren hielden registers bij, archeologen leggen hun opgravingen digitaal vast, we zijn nu ook bezig met de psychologen – ik vind alles interessant, als er maar data ter beschikking komen. Dat is ook wel een beetje kenmerkend voor mij: ik ben een omnivoor. In mijn geografieopleiding zaten stukken sociologie, antropologie en economie; en als historisch-geograaf heb ik jarenlang onderzoek gedaan met historici en archeologen in Griekenland.’

‘Overal haalden we oude ponskaarten en magneetbanden met onderzoeksgegevens vandaan om ze te ontsluiten’

‘En laat ik ten slotte nog zeggen, dat het af en toe ook buitengewoon actueel kan zijn, dat onderzoek naar historische data. Neem de discussie over de verhoging van de AOW-leeftijd die de laatste tijd speelt. Theo Engelen in Nijmegen heeft, op grond van historische reeksen uit onder andere de volkstellingen, laten zien dat er in feite nog nooit zoveel mensen aan het arbeidsproces hebben deelgenomen als tegenwoordig. Vroeger, toen de bevolkingspiramide nog een echte piramide was, en zelfs toen er nog kinderarbeid was, voor 1901, moest een veel kleiner deel van de bevolking dan nu het nationaal product bij elkaar verdienen.

Zo’n inzicht krijg je alleen als je naar lange reeksen kijkt, en niet alleen naar de laatste paar jaar. En dat heeft ook politieke consequenties. Er wordt maar gehamerd op de verhoging van de AOW-leeftijd van 65 naar 67, maar dat is maar een van de mogelijke oplossingen voor het probleem. Je kunt ook meer actieven krijgen door de arbeidsparticipatie van vrouwen te verhogen – en dat lijkt mij eerlijk gezegd een veel betere politieke koers.’

WIEBE KIESTRA

Peter Doorn: ‘Van alles houden mensen gegevens bij, en die wil DANS beschikbaar maken voor analyse’.

Peter Doorn studeerde sociale geografie in Utrecht en promoveerde daar in 1989. Tot eind jaren negentig 1998 doceerde hij computergebruik in de geschiedenis in Leiden. Daarna werd hij afdelingshoofd van het Nederlands Instituut voor Wetenschappelijke Informatie-diensten (NIWI) en directeur van het Nederlands Historisch Data Archief. Hij was betrokken bij veel projecten op het grensgebied van geschiedenis en informatica, zoals HGIN (naar een Historisch Geografisch Informatiesysteem voor Nederland), en Life Courses in Context. Doorn nam in 2005 de leiding op zich van het nieuwe instituut DANS.

Focus

Kohnstamm Instituut

THIJS HERMSEN

Het Amsterdamse Kohnstamm Instituut doet onderzoek naar onderwijs en opvoeding. Nadat het tientallen jaren onderdeel was van de afdeling Pedagogiek en Onderwijskunde van de Universiteit van Amsterdam (UvA), is het onderzoekscentrum afgelopen oktober verzelfstandigd.

CARO BONINK

Wetenschappelijk directeur van het Kohnstamm Instituut, Guuske Ledoux: 'Wij hechten veel belang aan maatschappelijke relevantie'

Het Kohnstamm Instituut is nu een zelfstandig bedrijf binnen de UvA, dat werkt in opdracht van derden, terwijl de afdeling Pedagogiek en Onderwijskunde het onderzoek uit eerste geldstroom doet. Onderzoek gefinancierd door organisaties als NWO doen beide partijen in samenwerking. Guuske Ledoux, de nieuwe wetenschappelijk directeur, benadrukt dat de nieuwe status niets verandert aan het type research: 'Wij deden altijd al veel beleidsgericht onderzoek naar onderwijs, opleiding, opvoeding en jeugdzorg in Nederland en dat blijven we op dezelfde manier doen. We blijven ook deel uitmaken van de UvA. De verzelfstandiging heeft vooral een financiële achtergrond.'

Ook de kwaliteit van het onderzoek blijft onveranderd hoog, verzekert Ledoux: 'Wetenschappelijkheid, onafhankelijkheid en betrouwbaarheid staan natuurlijk hoog in ons vaandel. Daarbij streven we naar een mix van toegepast en fundamenteel onderzoek. Wij hechten veel belang aan maatschappelijke relevantie: we doen graag onderzoek dat maatschappelijke kwesties verheldert en helpt oplossen.'

Het instituut heeft een staf van veertig medewerkers met verschillende achtergronden: opvoedkunde, onderwijskunde, psychologie, sociologie, taalwetenschappen. Een hoofdthema is de kwaliteit van onderwijs, opleiding, opvoeding en jeugdzorg. Ook talentontwikkeling en onderwijskansen van kinderen en jongeren zijn belangrijke aandachtsgebieden. 'We hebben een lange traditie in onderzoek naar onderwijsachterstanden en onderwijsongelijkheid,' aldus de wetenschappelijk directeur.

Het Kohnstamm Instituut heeft een brede mix aan klanten. Ledoux: 'Belangrijke opdrachtgevers zijn ministeries, gemeenten en brancheorganisaties als de raden voor primair en voortgezet onderwijs.' De variatie in opdrachtgevers en onderwerpen betekent samenwerking met veel verschillende partners, waarbij natuurlijk de afdeling Pedagogiek en Onderwijskunde van de UvA een constante factor is. Ledoux: 'Een partner waarmee we verder al lang samenwerken, bijvoorbeeld op het gebied van onderwijsachterstanden, is het Instituut voor Toegepast Sociaalwetenschappelijk onderzoek (ITS) van de Radboud Universiteit Nijmegen.' Een voorbeeld van een groot samenwerkingsproject is het Cohort Onderzoek Onderwijsloopbanen onder leerlingen van 5 tot 18 jaar (COOL 5-18). 'Daarin zijn wij partners naast het ITS, het toetsinstituut Cito in Arnhem, het Gronings Instituut voor Onderzoek van Onderwijs en het Centraal Bueau voor de Statistiek.'

In Nederland zijn veel instituten en bedrijven actief op het gebied van onderzoek naar onderwijs en opvoeding. 'Typerend voor de universitaire instituten is dat ze vaker de complexere onderzoeken doen waarbij de methodologie een stuk ingewikkelder is of waarin nieuwe dingen moeten worden bedacht,' aldus Ledoux.

Veel gegevens halen de onderzoekers zelf binnen door middel van surveys, toetsafnames, interviews en observaties. Ledoux: 'Een nieuw project heeft bijna altijd nieuwe gegevens nodig. Wij voeren veel projecten uit en bevragen en testen daarbij ook veel personen. We verzamelen dus ook heel veel data.' De databestanden worden niet afgeschermd voor buitenstaanders, benadrukt Ledoux, en dat verandert onder de nieuwe rechtsvorm niet: 'In principe kunnen alle wetenschappers onze databases gebruiken.'

'Al geruime tijd deponeren we grote databestanden, eerst bij het Steinmetzarchief en nu bij DANS. Dat geldt bijvoorbeeld voor databestanden van cohort-onderzoeken zoals COOL en PRE-COOL, een longitudinaal onderzoek naar de ontwikkeling van jonge kinderen vanaf twee jaar. Het gaat dan om tienduizenden kinderen die gedurende een langdurige periode verschillende keren worden bevraagd.' Alle data die het instituut verzamelt worden zorgvuldig bewaard en goed toegankelijk gehouden voor onderzoekers van buiten. 'De data die wij zelf beheren zijn natuurlijk ook goed gedocumenteerd. De technische rapportages zijn alleen nog niet in het Engels vertaald. Daar werken we nog aan.'

www.sco-kohnstamminstituut.uva.nl/

Repositories gaan meer samenwerken

De partners die betrokken zijn bij het Europese project DRIVER (Digital Repository Infrastructure Vision for European Research), hebben op 21 oktober besloten om hun samenwerking te intensiveren en uit te bouwen.

Waar DRIVER zich vooral richtte op de techniek van de repositories, digitale wetenschappelijke archieven, wil de nieuwe Coalition of Open Access Repositories DRIVER COAR ook de banden verstevigen op het gebied van organisatie en beleid.

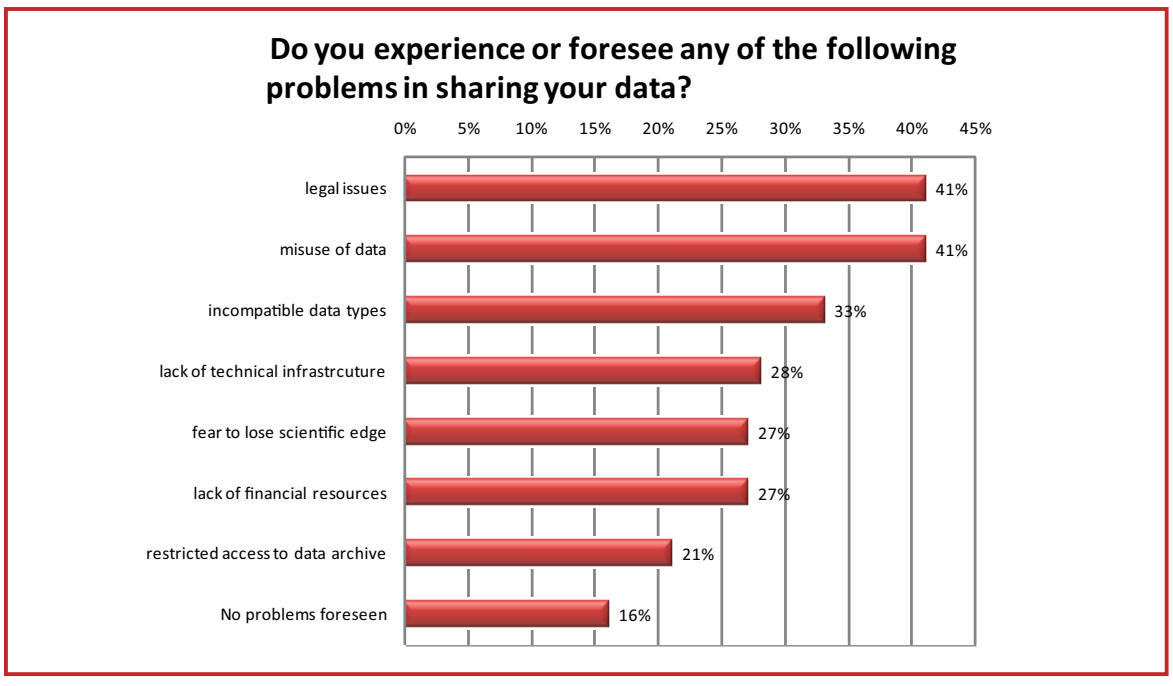
DRIVER COAR is een internationale not-for-profit vereniging die bevordert dat wetenschappelijke resultaten zichtbaarder worden en vaker toegepast kunnen worden met behulp van wereldwijde netwerken van open access digitale repositories. Om dit doel te bereiken worden in de eerste plaats de resultaten van het DRIVER project ingebed in een structurele organisatie. Daarnaast worden nieuwe initiatieven ontplooid om de repositories meer bekendheid te geven, ze verder te ontwikkelen en de zaak van Open Access op politiek niveau te bepleiten. DRIVER COAR laat ook de Europese wortels van DRIVER los: er wordt nu gestreefd nu naar een wereldwijd netwerk van zo'n duizend wetenschappelijke repositories die zich allemaal hard maken voor de kwaliteit van de onderzoeksresultaten zoals die in de repositories worden opgenomen en voor interoperabiliteit tussen de repositories.

De 28 organisaties die zich nu hebben aangesloten bij DRIVER COAR zullen daarvoor nog een lange weg moeten afleggen. Het deponeren van onderzoeksresultaten in de repositories van de universiteiten én in open access is in de wetenschap nog lang geen staande praktijk. Onderzoekers publiceren vaak liever in de belangrijke tijdschriften die door de grote uitgevers worden uitgebracht. En die brengen hun publicaties liever op de markt door middel van betaalde abonnementen. (IA)

www.driver-repository.eu/DRIVER-COAR.html

Europa niet duurzamer dan Nederland

Het PARSE.Insight project heeft de afgelopen maanden de duurzame toegang tot wetenschappelijke gegevens in Europa onderzocht. Wijkt de praktijk in Europa sterk af van die in Nederland, die onlangs werd beschreven in het rapport *Toekomst voor ons digitaal geheugen* van de Nationale Coalitie Digitale Duurzaamheid?



Het Europese project (Permanent Access to the Records of Science in Europe) was grootschaliger in opzet dan het Nederlandse en richtte zich alleen op de wetenschap. In totaal namen meer dan 1800 mensen deel aan drie enquêtes die tussen oktober 2008 en april 2009 werden verspreid onder onderzoekers, datamanagers en uitgevers. Uit de resultaten kan worden afgeleid dat op enkele uitzonderingen na duurzame toegang tot onderzoeksdata ook in Europa nog slecht is georganiseerd. Publicaties zijn er beter aan toe.

Een overgrote meerderheid van de respondenten is overtuigd van het nut van langetermijntoegang tot onderzoeksdata, vooral omdat men hierin mogelijkheden ziet om voort te bouwen op bestaand onderzoek. Ook zijn de meesten zich bewust van de bedreigingen die tot dataverlies kunnen leiden, zoals problemen met hardware en software of gebrek aan begeleidende informatie. De roep om een brede infrastructuur voor het bewaren en delen van onderzoeksdata is dan ook over de hele linie te horen, maar men is het niet eens over hoe een dergelijke infrastructuur er dan uit zou moeten zien. Genoemd worden een centraal geregeld systeem met open interfaces, maar ook een gedistribueerd netwerk. Over de rol van datamanagers en uitgevers in het geheel bestaat evenmin eenstemmigheid. Wel is men het snel eens over wie het geheel zou moeten financieren: dat zijn de nationale overheden.

Opvallend is dat onderzoekers graag gebruik maken van andermans data binnen de eigen discipline (interdisciplinair komt minder voor), maar aarzelen om hun eigen data te delen. Ze vrezen misbruik of zien juridische belemmeringen in de sfeer van privacy en auteursrecht.

De resultaten van het onderzoek werden bevestigd tijdens een PARSE.Insight workshop (zie pagina 2). Op basis van die resultaten, die in rapportvorm en als interactieve online grafieken beschikbaar zijn, wordt de PARSE.Insight roadmap voor een internationale science data infrastructure nu aangescherpt. Eind februari 2010 loopt het project ten einde en worden in een afsluitende conferentie de inzichten en aanbevelingen gepresenteerd. (Jeffrey van der Hoeven)

www.parse-insight.eu

DARIAH krijgt platte structuur

Op 19 en 20 oktober vond in Parijs een consortiumbijeenkomst plaats van het project DARIAH, Digital Research Infrastructure for the Arts and Humanities. Doel van het project is de bouw van een 'digitale werkbank' voor alfaweterschappers in Europa. Onderzoekers kunnen straks bij DARIAH terecht voor het vinden van data en tools, het archiveren van hun data, kennisuitwisseling en advies op het gebied van metadata en digitalisering.

Groepsfoto van de conferentiegangers in Parijs

De voorbereidende fase van het project, *Preparing DARIAH*, die wordt gecoördineerd door Data Archiving and Networked Services (DANS) is nu precies één jaar oud. In die tijd is hard gewerkt aan een structuur voor de te bouwen infrastructuur. Tijdens de bijeenkomst in Parijs is vooral gesproken over te vormen virtual competence centres (VCC's). DARIAH krijgt geen pyramide-structuur, maar de belangrijkste taken worden verdeeld over een aantal virtuele centra die in verschillende landen worden gevestigd. De belangrijkste VCC's zullen zijn: research and education, e-infrastructure, content and legal, en management, promotion/dissemination and advocacy. Het voordeel van deze opzet is dat geen enkel land dominant zal worden, maar dat een gedistribueerde infrastructuur zal ontstaan met een beperkt aantal sterke 'hubs'.

Een jaar geleden is DARIAH van start gegaan met de volgende partners: Nederland: DANS (Data Archiving and Networked Services); Verenigd Koninkrijk: CeRch (Centre for e-research) King's College London; ADS (Archaeology Data Service) University of York; OUCS (Oxford University Computing Services); Frankrijk: CNRS (Centre National de la Recherche Scientifique); Duitsland: MPG (Max Planck Gesellschaft); UGOE (Universiteit van Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen); Griekenland: AA (Academie van Athene); DCU (Digital Curation Unit), Athena Onderzoeks Centrum; Slovenië: ICH (Institute of Contemporary History); Ierland: IRCHSS (Irish Research Council for Humanities and Social Sciences); Denemarken: NFI (Nordisk Forskningsinstitut), Universiteit van Kopenhagen; Kroatië: RBI (Ruder Boskovic Instituut); Cyprus: EUC (Europese Universiteit van Cyprus).

Inmiddels hebben zich vier geassocieerde partners aangesloten uit drie verschillende landen (Italië: FRD, Fondazione Rinascimento Digitale; Spanje: LaPa, The Heritage Laboratory, Spanish Research Council; Zweden: SND, Swedish National Data Service; HUMlab, Faculty of Arts Umea University). Een aantal andere organisaties heeft belangstelling getoond. Het groeiende aantal leden wordt door de betrokkenen beschouwd als een goede voorbode voor het tweede jaar van DARIAH. (Milena Piccoli)

Sinds kort beschikbaar

Het overzicht toont een aantal databestanden die recent voor onderzoekers beschikbaar zijn gekomen bij CBS en DANS. Een volledig overzicht van de CBS-bestanden is te vinden op www.cbs.nl/microdata. De bij DANS beschikbare databestanden komen van diverse andere onderzoeksinstellingen. Deze kunnen kosteloos worden gedownload vanuit DANS EASY: http://easy.dans.knaw.nl. Via DANS kunnen ook de beveiligde microdata van het CBS kosteloos geleverd worden: www.dans.knaw.nl/nl/data/cbs/overzicht/

Centraal Bureau voor de Statistiek	Periode
Sociaal Statistisch Bestand Overige Uitkeringen	2006
Bijstandsuitkeringen Statistiek	1995–2007
Centraal Administratie Kantoor Zorg met Verblijf	2007
POLS Module gezondheid	2008
Productiestatistieken Groothandel	1993–2007
Productiestatistieken Detailhandel	1993–2007
Productiestatistieken Arbodiensten en re-integratiebedrijven	2006–2007
Productiestatistieken Sociale werkvoorziening	2006–2007
Via DANS EASY	
Archeologie	
Vijf keer archeologie te Reusel-De Mierden (Amsterdams Archeologisch Centrum)	2008
The structure of Bronze Age settlements in the Dutch river area (Arnoldussen, S.)	2009
The physical landscape of the Dutch river area (Zijverden, W. Van)	2009
De Mortel De Leigraaf – Inventariserend Veldonderzoek door middel van Proefsleuven *(BAAC)*	2009
Geschiedenis	
Nederlandse SS'-ers. Erfgoed van de Oorlog; interviews afgenomen in het kader van het project Getuigen Verhalen *(Stichting Zuidenwind Filmprodukties)*	1940-1945
Sociale wetenschappen	
VO Monitor	2007
HBO Monitor	2007
WO Monitor	2007
BVE Monitor	2007
(ROA Universiteit Maastricht)	
International Social Survey NL – Nederlandse data	2003-2004
International Social Survey NL – Nederlandse data	2005-2006
International Social Survey NL – Nederlandse data	2007-2008
(H.B.G.Ganzeboom – Vrije Universiteit Amsterdam)	
Permanent Onderzoek Leefsituatie – POLS Basis *(CBS beveiligd microbestand)*	2008

Planets-dag: uitleg over bewaargereedschappen

De Koninklijke Bibliotheek (KB), het Nationaal Archief en de Nationale Coalitie Digitale Duurzaamheid (NCDD) organiseren op 14 december een Planetsdag in de aula van de KB. Onderzoekers van de KB en het NA leggen daar uit welke gereedschappen ze hebben ontwikkeld in het kader van het Planets-project en hoe die kunnen worden ingepast in het werk.

De KB en het Nationaal Archief nemen al enkele jaren deel aan het Europese project Planets (Preservation and Longterm Access through Networked Services), dat als doel heeft om praktische gereedschappen te ontwikkelen voor langetermijnbewaring van digitale objecten. Nu het project op zijn einde loopt wordt het tijd om de balans op te maken: wat heeft het project opgeleverd, en hebben we ook echt iets aan de ontwikkelde instrumenten? Zo is binnen Planets bijvoorbeeld een tool ontwikkeld, SIARD genaamd, dat de oorspronkelijke relaties tussen verschillende databasecomponenten in stand kan houden. Dit is belangrijk voor het toegankelijk houden van bevolkingsonderzoeken die in de jaren tachtig onder het databaseprogramma dBase werden opgeslagen. SIARD is ontwikkeld door het Zwitserse Federale Archief. Door de databasetabellen inclusief de relaties om te zetten naar XML, maakt SIARD het mogelijk om met gebruikmaking van moderne databasesoftware de oude gegevens te bevragen.

Maar Planets heeft aanzienlijk meer opgeleverd, zoals de planningstool PLATO die helpt bij het maken en documenteren van een keuze voor de juiste migratie- of emulatiesoftware. Of de *Core Registry* die technische informatie bevat over de eigenschappen van digitale objecten en digitale duurzaamheidstools. En natuurlijk het Testbed, dat gebruikers in staat stelt karakteriserings-, migratie of emulatie-experimenten in een gecontroleerde omgeving uit te voeren. Meer informatie en een aanmeldingsformulier is te vinden op de website van de NCDD. (Frank Houtman)

www.ncdd.nl/toolkit.php

Nieuwe GIS-toepassingen in alfawetenschappen

Nadat eerder dit jaar het project AlfaGeo, bedoeld om het gebruik van geo-informatie in de alfawetenschappen te bevorderen werd afgerond, verschijnt deze maand mede op basis daarvan het boek 'Tijd en Ruimte; Nieuwe toepassingen van GIS in de alfawetenschappen'.

Binnen AlfaGeo ontplooide DANS een aantal initiatieven, zoals studiedagen waarin onderzoekers voorbeelden toonden van het gebruik van geografische infiormatiesystemen in wetenschappelijk onderzoek. De presentaties op die studiedagen vormen een belangrijke basis van het boek, dat onder redactie van Onno Boonstra (Radboud Universiteit Nijmegen) en Anton Schuurman (Universiteit Wageningen) tot stand is gekomen en wordt uitgegeven door uitgeverij Matrijs. Het boek vormt de kroon op het AlfaGeo project. Hoewel de toepassing van geografische informatieysstemen en het gebruik van de informatie eruit in de alfawetenschappen een opvallende bloei doormaakt, is kennis erover slechts bij een beperkte groep onderzoekers aanwezig. Ook bestaat er geen overzicht van de bronnen en toepassingen waardoor veel mogelijkheden onbenut blijven. In het boek worden de nieuwste ontwikkelingen beschreven om te laten zien hoe GIS binnen de alfawetenschappen ingezet kan worden. De verschillende functies die GIS kan vervullen vormen de kapstok voor zestien artikelen van auteurs afkomstig uit verschillende wetenschapsgebieden. De bijdragen laten zien op welke wijze de analyse van gegevens met een ruimtelijke component, zoals het kadaster en de volkstellingen, door middel van GIS-systemen tot nieuwe inzichten kan leiden. Het boek bevat vele fraaie afbeeldingen en richt zich, mede door de vormgeving en tekstredactie, op een breed publiek. (Rene van Horik)

** Onno Boonstra, Anton Schuurman (red.): Tijd en Ruimte; Utrecht, Historische uitgeverij Matrijs; verschijnt in december 2009*

COLOFON

e-data&research is het kwartaalblad in Nederland over data en onderzoek in de alfa- en gammawetenschappen. Het verschijnt onder auspiciën van DANS, het Huygensinstituut, het Internationaal Instituut voor Sociale Geschiedenis, 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: 8800. 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)3494450 **f** (070)3494451 **e** edata@dans.knaw.nl

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

Aan dit nummer werkten mee: Marjolein van den Dries, Jeffrey van der Hoeven, René van Horik, Frank Houtman, Esther Jansma, Peter Kanne, Hans van Maanen, Liesbeth Koenen, Douwe Zeldenrust.

Redactiesecretariaat: Lucas Pasteuning, Jetske van der Schaaf

Vormgeving en opmaak: Ellen Bouma

Productie: Uitgeverij Aksant, Amsterdam

Druk: Thieme Almere

ISSN: 1872-0374

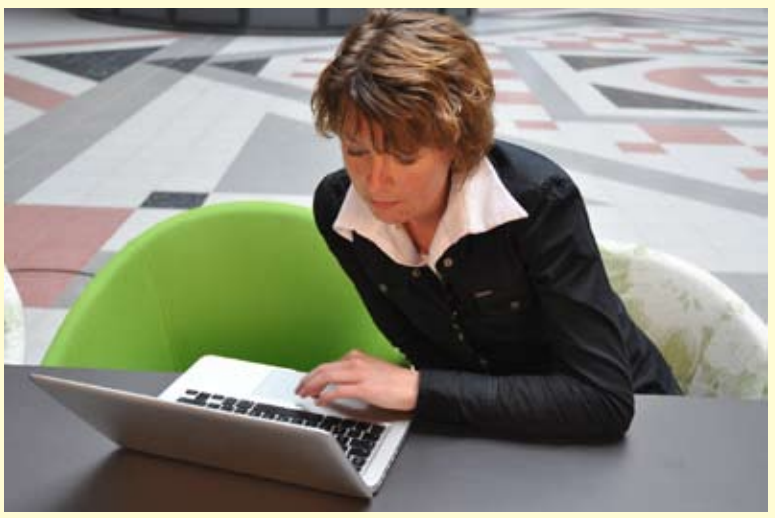
INGEZONDEN

Wat is slecht onderzoek?

In het septembernummer van *e-data&research* pleitte Hans van Maanen voor een Enquêtekeurmerk. Hij stelde voor dat de leden van het Nederlandstalig Platform voor Survey-Onderzoek in ieder persbericht over een onderzoek vermelden hoe betrouwbaar de gegevens zijn volgens de maatstaven van het NPSO, en in een voetnoot uitleggen hoe ze tot dit oordeel zijn gekomen.


NATIONALE BEELDBANK

Dit is niet zo makkelijk als het lijkt. De kwaliteit van een onderzoek heeft vele aspecten en hangt onder meer samen met het doel. Eigenlijk is het makkelijker om te zeggen wat slecht onderzoek is dan wat goed onderzoek is. Als je als respondent de volgende vraag voorgelegd krijgt, weet je bijvoorbeeld zeker dat het een slecht onderzoek is.


NATIONALE BEELDBANK

Vindt u dat de regering meer aandacht moet geven aan het snel afbouwen van de staatsschuld of meer geld besteden aan zaken voor het land en de burger?
Ja; Nee; Weet niet/geen mening

En als een onderzoek naar het gemak van elektronische belastingaangifte wordt uitgevoerd onder een steekproef van doorgewinterde internetgebruikers, is ook duidelijk dat dit geen representatief beeld geeft van de hele bevolking. Misschien geeft het wel een beeld van het oordeel van fervente internetgebruikers. Dat moet de onderzoeker dan wel duidelijk in zijn rapportage vermelden en de journalist in zijn artikel. Een onderzoek kan dus voor het ene doel slecht zijn en voor het andere doel misschien wel goed.

In de wereld van de surveys is langzamerhand overeenstemming bereikt over kwaliteitsmaatstaven. Een onderzoek moet het onderwerp goed dekken, de resultaten moeten accuraat zijn, de gegevens moeten tijdig beschikbaar komen en toegankelijk zijn en het gebruik van standaardvragen voor achtergrondgegevens wordt aanbevolen. De accuraatheid van resultaten neemt in het algemeen toe naarmate de steekproef groter is, maar dat geldt alleen bij een random steekproef. En helaas kan hoge nonrespons ook bij een grote random steekproef tot een foutief resultaat leiden. Lastig is dat verschillende componenten van kwaliteit elkaar niet kunnen compenseren. Als de vragen niet deugen, is de hoogte van de respons irrelevant. En als je de resultaten snel nodig hebt (Vond men dat er nieuwe verkiezingen zouden moeten komen als Balkenende president van Europa zou zijn geworden?) heb je geen tijd voor een uitgekiend responsverhogend programma.

Een keurmerk is dus niet zo makkelijk. De tweede suggestie van Van Maanen is eenvoudiger uitvoerbaar. Vermeld bij iedere dataverzameling de omvang van de steekproef, hoe de respondenten zijn geselecteerd, de wijze van data verzamelen en de nonrespons en geef een link zodat de geïnteresseerde lezer op haar gemak de hele vragenlijst kan doorlezen.

Dat zou een belangrijke eerste stap zijn. In de tussentijd kunnen we binnen de NPSO verder discussiëren over nut en noodzaak van een keurmerk, en ook alvast een overzicht geven van alle bestaande richtlijnen en kwaliteitsoverzichten die we kennen. Wil je mee doen: meld je aan op www.npso.net. We komen op dit onderwerp graag nog eens terug.

Ineke Stoop is mede-oprichter en lid van de kerngroep NPSO (www.npso.net)

Column
Liesbeth Koenen

De hinderpaal taal


MARCO FRIGERIO

Droombeeld: alle wetenschappelijk onderzoek van alle tijden en talen staat on line. Liefst samen met lekker veel bronnen en ruwe gegevens. Dus niet alleen boeken en artikelen, maar ook kleitabletten en papyrusrollen, oude dagboeken en labaantekeningen, en nog oneindig veel meer. En dan tik of spreek je wat zoektermen in, en floep, daar verschijnt in je eigen taal wat de Japanners erover te zeggen hebben, wat er in Engeland in de fameuze zeventiende eeuw over is opgeschreven en welke conclusies Indiase onderzoekers trekken. Je komt vanzelf uitsluitend terecht bij gegevens die ertoe doen, en overal vind je handzame samenvattingen.

Nu de echte wereld. Een Telegraafbericht op internet over een zware zeebeving toont direct daarnaast een advertentie voor een fijn cursusweekend overleven op de Rucphense heide. Laatst vond een vriendin een brief op de mat, gericht aan 'Lieve inwoners van de busje Eeghenlaan en het busje Eeghenstraat '. Ook de rest van het schrijven ('de inbreker ging door onze rug tuin', 'verhuur alstublieft ons weten') begon pas enigszins begrijpelijk te worden toen ze zich realiseerde dat het om een computervertaling uit het Engels ging (van is busje, back achter en rug, en let zowel laat als verhuur).

De automatische omzetting van achttiende-eeuwse artikelen uit De Vaderlandsche Letteroefeningen met Optical Character Recognition (te vinden bij e-laborate.nl) maakt van 'het menschelyk geslagt' 'het menfchelyk geflagt' en achter 'geraoedsgefieldheid' blijkt bij controle gewoon 'gemoedsgesteldheid' schuil te gaan. Verbaasd lees ik deze week in een mailtje 'Ok zit Nietzsche boot niks op 40000 vortex te mailen.' De volgende dag volgt gelukkig de uitleg: het is wat je na autocorrectie van iPhone overhoudt van 'Ik zit niet voor niks op 40000 voet te mailen.' Spellingcheckers vinden tegenwoordig flessentelefoon, geelhork en opmuizen prima woorden. En ook tegen een computer praten, levert nog altijd hilarische resultaten op.

Ze liggen voor het opscheppen. Bij de weidse vergezichten op een ideale digitale wereld staat kortom een ding gigantisch in de weg: taal. Computers begrijpen er nog steeds geen bal van. Intussen wordt ons al sinds 1945 'binnen vijf jaar' de vertaalcomputer beloofd. Er wordt gesleuteld en gedaan, en soms lijkt het nog heel wat. Maar het fundament ontbreekt. We weten zelf veel te weinig over ons ingenieuze taalvermogen.

Het rare is intussen dat alleen taalkundigen dat lijken te beseffen. Taal is ons kennelijk zo eigen, is zo gewoon, dat het voor niet-ingevoerden meestal een makkie lijkt. Of weet iemand een betere verklaring voor het onbegrijpelijke feit dat taalonderzoek niet allang op elke universiteit vooraan staat zodra de onderzoeksgelden worden verdeeld?

Liesbeth Koenen is taalkundige en wetenschapsjournalist

Gelezen

Peter Boot: *Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship*; Amsterdam, Amsterdam University Press, 2009; Proefschrift Universiteit Utrecht, november 2009; ISBN 978-90-89641-87-8
De belangrijkste ontbrekende functionaliteit in huidige digitale edities is die van annotatie. Digitale edities zouden aan onderzoekers de mogelijkheid moeten bieden om gestructureerde en ongestructureerde observaties met betrekking tot de uitgegeven tekst vast te leggen.
Dit boek bespreekt een aantal benaderingen van annotatiesystemen. Het doet dat in het kader van de studie van het embleem, het zestiende en zeventiende eeuwse literair genre waarin een afbeelding, een motto en een vaak moraliserende epigram met elkaar verbonden werden.
Bij een juiste aanpak kan annotatie uitgroeien tot mesotekst: tekst gepositioneerd tussen de geannoteerde teksten en de wetenschappelijke artikelen en monografieën waarvoor de annotaties de argumenten leveren. In een digitale context moet het mogelijk zijn om heen en weer te navigeren tussen geannoteerde tekst, annotatie en artikel.

M. Senten (redactie): *Experiment NL 2, wetenschap in Nederland*; Diemen/Den Haag, G+J Publishing, 2009; ISBN 978-94-6044-014-4
Het vierde NWO-publieksboek Experiment NL 2 beschrijft in begrijpelijke taal de meest opmerkelijke, briljante, leuke en spannende onderzoeksprojecten van het afgelopen jaar. Onderzoekers uit alle disciplines komen aan het woord en tonen hun bevindingen aan mensen die geïnteresseerd zijn in wetenschap, vernieuwing en de wereld om hen heen. *Verwondering*, *Op onderzoek* en *Experiment NL deel 1* gingen deze uitgave voor.

Henk Vinken: *Verkenning Jeugddata. Raadplegingen van jeugdonderzoekers met betrekking tot hun datawensen.* Den Haag, Data Archiving and Networked Services (DANS), 2009; Studies in Digital Archiving 3; ISBN 978-94-90531-01-0
Jeugd heeft nu veel aandacht in wetenschap, overheidsbeleid en samenleving. Er wordt veel jeugdonderzoek gedaan door diverse onderzoekinstituten en universiteiten, zowel landelijk als regionaal. Tot op heden is er echter weinig systematisch inzicht in aard en kwaliteit van dit jeugdonderzoek en van de onderliggende onderzoeksgegevens. DANS (Data Archiving and Networked Services) heeft daarom in de eerste helft van 2009 een Dataverkenning Jeugd uitgevoerd. Deze publicatie, uitgebracht in de reeks Studies in Digital Archiving, doet verslag van die verkenning.

Karl Fogel: *Open source-software produceren, met succes een open source-softwareproject runnen*; Utrecht, Stichting Kennisnet, SURFfoundation, SURFdiensten en SURFnet; Juli 2009
Het boek *Producing Open Source Software* van de Amerikaanse Open Source-deskundige Karl Fogel geldt in de internationale Open Source-community als een standaardwerk en is inmiddels in diverse talen beschikbaar. Sinds afgelopen zomer ook in het Nederlands. SURF en Kennisnet hebben de vertaling, met de titel *Open Source-software produceren*, financieel mogelijk gemaakt.

Het boek is bedoeld voor ontwikkelaars en managers van open software die overwegen een open source-project te beginnen, en voor degenen die al een project gestart zijn en zich afvragen hoe ze verder moeten. Het kan ook handig zijn voor mensen die alleen willen participeren in een open source-project, maar dit nooit eerder gedaan hebben. De lezer hoeft geen programmeur te zijn, maar moet wel de technische basisconcepten van software kennen, zoals 'broncode', 'compilers' en 'patches'. Ervaring met open source-software, als gebruiker of ontwikkelaar, is niet nodig. Mensen die al aan open source-softwareprojecten hebben meegewerkt, zullen sommige delen van het boek waarschijnlijk als een open deur beschouwen en deze waarschijnlijk willen overslaan. Omdat er grote verschillen kunnen zijn wat betreft de ervaring die de lezers hebben, zijn de secties van duidelijke koppen voorzien en wordt vermeld wanneer iets overgeslagen kan worden door degenen die al bekend zijn met deze materie.
De pdf-versie van de vertaling is kosteloos beschikbaar:
www.surf.nl/nl/OverSURF/Publicaties