



E-DATA & RESEARCH

Jaargang 8 | nummer 3

Nieuwsbrief over data en onderzoek in de alfa- en gamma-wetenschappen.

E-data & Research verschijnt drie keer per jaar en wordt mogelijk gemaakt door: CentERdata, CLARIAH, DANS, Huygens ING, de Koninklijke Bibliotheek en het RIVM.

INHOUD

- 3 Een kennismaking met Linked Open Data
- 4 Deze datasets zijn sinds kort beschikbaar
- 5 Toine Pieters: "Het gaat om digital empowerment"



- 6 NPSO adviseert over goed onderzoek doen
- 6 Over vrouwen en het literaire veld
- 7 MigMap brengt verhuisgedrag in kaart
- 7 KB: open toegang tot ieder wetenschappelijk artikel
- 8 Piek Vossen over machines en mensentaal
- 8 Ewoud roept op tot een digitaliseringsregister



Scan deze QR code met een smartphone om de website van E-data te bezoeken.
www.edata.nl

'Wetenschappers die data niet delen, maken meer fouten'

OA-tijdschrift JOPD laat zien hoe het moet

Wat doe je, als tijdschriften in jouw vakgebied vrijwel onverschillig staan tegenover vrij toegankelijke data? Dan start je zelf een open access journal.

Marika de Bruijne, Heidi Berkhout

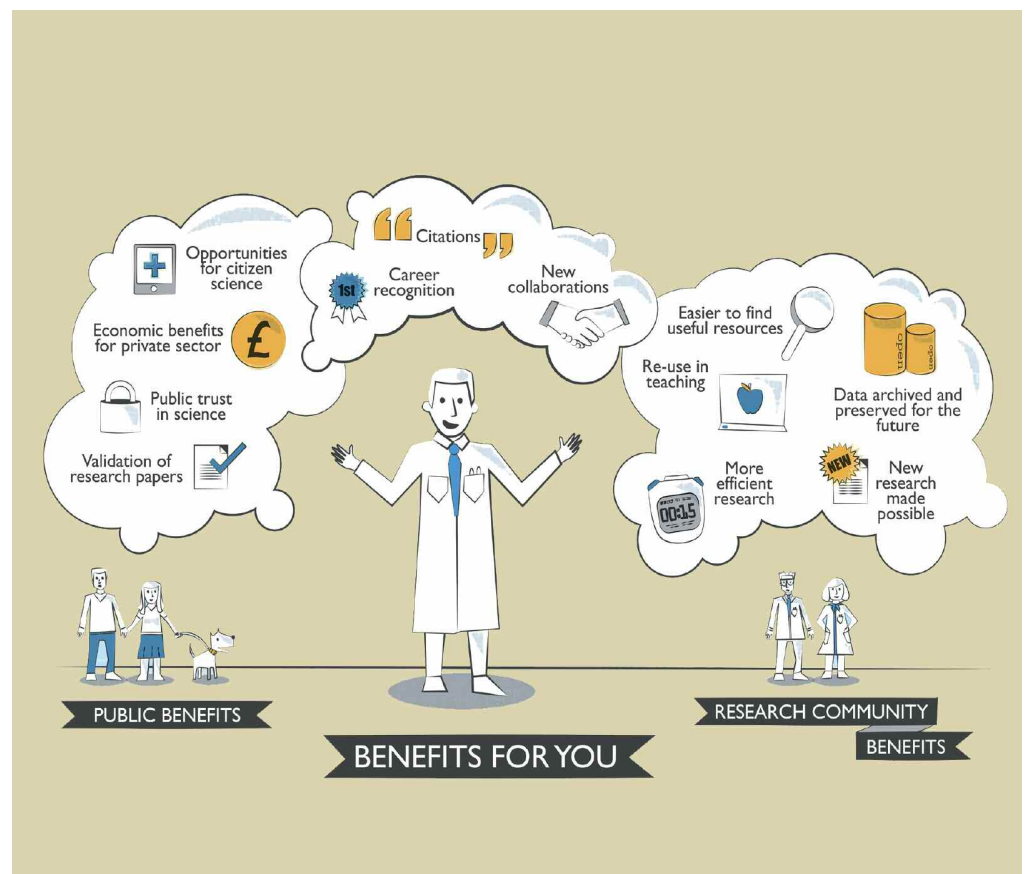
"Ik begon het tijdschrift uit frustratie." Aan het woord is Jelte Wicherts, Universitair Hoofd-docent aan Tilburg University, fervent aanhanger van open access en inmiddels ook hoofdredacteur van 'zijn' *Journal of Open Psychology Data (JOPD)*. "Tijdens het schrijven van mijn proefschrift had ik historische IQ-data nodig. Ik was verbijsterd dat de data wel waren verzameld, maar dat vrijwel niemand ze kon leveren. Voor mij zijn netjes werken, open zijn naar andere wetenschappers en data hergebruiken standaarden. Maar voor velen leek het eigen cv belangrijker", zucht hij. "Natuurlijk mag degene die de data verzamelt, de eerste vruchten plukken, dat is normaal. Maar het risico bestaat dat men de data op de eigen computer bewaart en dat men denkt er ooit nog een publicatie over te schrijven. Maar de tijd gaat voorbij, er komt een nieuwe computer, en weg zijn de data. Ze zijn niet netjes gedocumenteerd, want dat hebben de onderzoekers ook niet geleerd."

Iets niet helemaal goed

Uit onderzoek van Marjan Bakker, in april 2014 gepromoveerd aan de Universiteit van Amsterdam, blijkt dat in ongeveer de helft van de bestudeerde psychologische artikelen fouten stonden in het rapporteren van resultaten. Ze ontdekten ook dat wetenschappers die data niet delen, meer fouten maken. Het is volgens Wicherts nog veel erger. Als voorbeeld van dubieus onderzoek noemt hij een artikel dat in 2011 in *Journal of Personality and Social Psychology* verscheen dat bewees dat we in de toekomst kunnen kijken. "Dan is er toch iets niet helemaal goed in de psychologie", denkt Wicherts.

Beloon het delen van data

"Het is verbazingwekkend hoeveel weerstand er is tegen het wetenschappelijk belang: alles delen. Zowel vanuit uitgevers, universiteiten en subsi-



Het JOPD wordt uitgebracht door uitgever Ubiquity Press van University College London. De voordelen van het delen van open data worden in deze afbeelding in beeld gebracht bron Ubiquity Press Ltd.

dieverstrekters als wetenschappers zelf in machti-ge posities. Gelukkig is er inmiddels een ten-dens gaande om de beloningsstructuur te veran-deren. Zo is kwantiteit geen hoofdcriterium meer binnen de VSNU-criteria bij onderzoekvisaties. En verplicht de VSNU-gedragcode onderzoe-kers om data te delen als erom gevraagd wordt. Ook de American Psychological Association (APA) heeft een vergelijkbare gedragscode." Toch worden onderzoekers nog steeds niet beloond voor het delen van hun data. "En dus is het niet zo gek, dat data niet gedeeld worden", zegt Wicherts. "Het idee voor het JOPD was heel sim-pel: we moeten het delen van data belonen."

Het wauw-effect

Met het JOPD hoopt hij verandering te brengen. Het tijdschrift publiceert artikelen die onder-zoeksdata beschrijven. "De artikelen die we nu krijgen, zijn van de voortrekkers van open data. Zij zijn enthousiast. Het verkeer op de website is

boven verwachting, alle artikelen zijn meer dan 200 keer bekeken. We hebben verder de data op-gevraagd van alle 'meest geciteerde artikelen uit de psychologie van de afgelopen 25 jaar'. We gaan zelf achter auteurs aan. We geven presenta-ties en via-via proberen we de *talk of the town* te worden. Over een jaar wil ik elke maand twee à drie artikelen publiceren, meer (studerende) bezoekers op de website hebben en een paar data papers met een wauw-effect waaruit geciteerd wordt."

Maar nog liever ziet hij zijn tijdschrift overbodig worden. "De dag dat de APA met haar 50 tijd-schriften besluit dat data verplicht bij publicaties moeten staan, op die dag stop ik ermee", belooft Wicherts en hij legt uit waarom. "Want dan is het niet meer nodig. Tot die tijd, en zo lang er onderzoekers zijn die hun data bij zich willen houden, is ons werk waardevol. Simpelweg omdat we laten zien hoe het moet."

openpsychologydata.metajnl.com

OPROEP

Wint u de Nederlandse Dataprijs 2014?



Op 24 september wordt de Nederlandse Dataprijs uitgereikt. Een prijs voor een onderzoeker of onder-zoeksgroep die bijdraagt aan de wetenschap door onderzoeksdata beschikbaar te maken voor aan-vullend of nieuw onderzoek. Bent of kent u zo'n onder-zoeker met een goed voorbeeld van gedeelde data? **U kunt tot 1 juli uzelf of een ander via researchdata.nl nomineren.**

De winnaars van de voorgaande edities zijn in ieder geval enthousiast: "Het winnen van de Dataprijs, en zelfs al de nominatie daarvoor, onderstreept het be-lang van je werk. Dat voelt goed," aldus Maarten

Marx, en: "Het winnen van de Dataprijs geeft aan dat we de goede weg zijn ingeslagen," aldus Mark van Koningsveld, winnaars van de Dataprijs 2012. Kees Mandemakers, winnaar in 2010: "Laat maar zien hoe goed je database of dataset al is! De Dataprijs is een kroon op ons werk."

De organisatie van de Nederlandse Dataprijs is in handen van **Research Data Netherlands (RDNL)**, een samenwerkingsverband tussen **3TU.Datacen-trum** en **DANS**. Kijk voor meer informatie over de prijs, zoals de twee categorieën, de prijzenpot en de criteria op de website van RDNL. **researchdata.nl**

GEHOORD & BIJGEWOOND

E-humanities ook voor historische taalkunde

Erica Renckens

Oude teksten vormen voor de computer een grote uitdaging, vooral vanwege de enorme spellingvariatie. Dat is echter geen argument om digitale onderzoeksmethoden links te laten liggen in de historische taalkunde, zo betoogde prof. dr. Nicoline van der Sijs tijdens haar oratie op 27 maart jongstleden aan de Radboud Universiteit. Ter ondersteuning van dit standpunt besprak ze verschillende mogelijkheden van de toepassing van computationele technieken in haar vakgebied. Zo bewijst een vergelijking tussen verschillende corpora dat de Statenvertaling al bij zijn verschijnen in 1637 archaïsch was: de tekst bevat meer verbogen naamvallen dan andere teksten uit dezelfde periode. Ook kan de computer taalveranderingen als neologismen en nieuwe grammaticale constructies opsporen door twee corpora te analyseren.



Nicoline van der Sijs tijdens haar oratie foto Gert Janssen

Om onderzoeksvragen met behulp van een computer te kunnen beantwoorden, moeten wetenschappers die juist formuleren. Om dat te trainen, organiseren faculteiten cursussen en workshops *e-humanities*. Daarnaast zal Van der Sijs in de loop van volgend jaar met collega's een handboek voor de digitale geesteswetenschappen publiceren. ru.nl/nederlands/wie_wat_waar/medewerkers/nicoline-sijs

Nieuwe tools voor Europees onderzoek

Tom van Nuenen

De universiteit van Würzburg ontving op 1 en 2 april jongstleden een twintigtal jonge wetenschappers uit Europa en Amerika voor de tweede workshop van het *Network for Digital Methods in the Arts and Hu-*

manities (NeDiMAH). Het netwerk, dat in 2011 werd opgericht, richt zich op computationeel onderzoek in de Europese humaniora.

Er was ruimte voor kennismaking met nieuwe tools – zoals *Stylo*, een stilometrische tool die verschillende classificatie- en clusteringalgoritmen combineert. Maar vooral het accent op samenwerking bleef terugkeren. De verschillende sprekers bleken in elk van hun landen bezig met vergelijkbare projecten om grote hoeveelheden literaire teksten te indexeren en analyseren. In Nederland wordt bijvoorbeeld *NederLab* ontwikkeld: een website die Nederlandstalige tekstcollecties onder één dak bijeen brengt en tekstanalyse-software aanbiedt om die collecties te ontleiden. In Duitsland, Polen en Amerika bleken soortgelijke projecten gaande. De vraag die daarbij telkens opnieuw aan de orde kwam: hoe zorg je ervoor dat die corpora geformaliseerd en uitwisselbaar zijn tussen die landen en talen?

Een eerste stap werd alvast gezet op de tweede dag, toen een op handen zijnde 'benchmark corpus' van *NeMiDAH* werd besproken. Organisatrice Karina van Dalen-Oskam (Huygens ING) legt uit: "Het gaat om een corpus van honderd Engelse talige romans, waaronder één roman waarvan het auteurschap betwist is en aan sommige van de andere auteurs in het corpus is toegeschreven. Doel van dit corpus is om auteursherkenningssoftware te kunnen testen en de resultaten te kunnen vergelijken. Als iemand een nieuwe tool ontwikkelt, en die op exact dezelfde data toepast, kun je meten hoe goed hij werkt."

nedimah.eu

EACL toont ook de muziek van de literatuur

Mike Kestemont

De *Conference of the European Chapter of the Association for Computational Linguistics* (EACL) is een prestigieuze conferentie binnen de computerlinguïstiek, waar onderzoekers de recente vooruitgang in de (vooral Europese) taaltechnologie laten zien. Naast het relatief technische hoofdprogramma kent de EACL een vruchtbare traditie van kleinere workshops die aan de conferentie voorafgaan. Deze studiedagen zijn meestal georganiseerd rond een nieuw en uitdagend onderzoeksthema binnen de computerlinguïstiek en haar toepassingen. Zo ook de workshop *Computational Linguistics for Literature* (CLfL), inmiddels al aan de derde editie toe.



Debora Thomas: "This digital age is like the Wild West: opportunities and unruly" foto Inge Angevaare

Aan deze workshop nam een groep van enthousiastelingen deel die zich interesseren voor taaltechnologische toepassingen binnen de hedendaagse literatuurstudie.

De deelnemers rapporteerden over onderzoek in uiteenlopende gebieden, zoals een sociolinguïstische analyse van personages in het oeuvre van Ibsen, intertekstuele verbanden in middeleeuwse Arabische literatuur of de modellering van het lineair verloop in literaire teksten. Keynotesprekers waren de Hamburgse narratoloog Jan Christoph Meister en de visuele artieste María Mencía. Hoogtepunt van de dag was de innovatieve bijdrage van Hannah Davis en Saif Mohammad, waarin zij *TransProse* presenteerden, een systeem dat automatisch muziek op basis van tekst genereert. Binnen hun intuïtieve aanpak wordt het overheersende sentiment van een tekst eerst benaderd op basis van een emotielexicon. De analyse van deze sentimentanalyse wordt vervolgens omgezet in muzikale patronen waarvan de toonhoogtes en het ritme de gevoelens van de tekst nabootsen. Het verbluffende resultaat van *TransProse* kan online worden geraadpleegd (transprose.weebly.com). Laat u ontroeren door de tragische tonen die het systeem distilleerde uit *Heart of Darkness*. aclweb.org/anthology/W/W14/#0900

Patronen en betekenis in grote tekstcollecties

Peter Boot

Hoe kunnen geesteswetenschappelijke projecten met *big questions* het beste omgaan met de *big data* van de grote tekstcollecties waarmee ze hun vragen willen beantwoorden? Dat was de centrale vraag tijdens

de workshop *Mining digital repositories*, gehouden op 10 en 11 april in de Koninklijke Bibliotheek. De workshop was mede georganiseerd door de projecten *Translantis* en *ASYMENC* (zie ook het interview met Toine Pieters op pagina 5).

Debora Thomas toonde de digitale krantencollectie van de Library of Congress (USA). De verzameling op zich is al een klein kunststuk, omdat collecties uit vijftig staten moesten worden geïntegreerd. De kranten kunnen online worden doorzocht en gelezen, maar zijn ook benaderbaar via een *application programming interface* (api): onderzoekers kunnen de krantenteksten analyseren met hun eigen programmatuur en zo onder andere *topic modelling* of *sentiment mining* doen. Zo is bijvoorbeeld bestudeerd hoe de kranten in 1918 verslag deden van de Spaanse griep. Alistair Dunning toonde vergelijkbare mogelijkheden in de *Europeana Cloud*, en Hildelies Balk gaf een preview van het researchlab dat de Koninklijke Bibliotheek aan het bouwen is. Ook de aanbieders van commerciële collecties hebben begrepen dat veel onderzoekers van zulke mogelijkheden gebruik willen maken.

Toch toonden veel geesteswetenschappelijke onderzoekers zich sceptisch. Wat is de status van een grafiek of een *word cloud*? Kunnen geesteswetenschappelijke vragen beantwoord worden met een beroep op getallen? Kunnen ze überhaupt worden beantwoord? En hoe betrouwbaar en representatief zijn de repositories? Organisator Joris van Eijnatten besloot dat er nog heel wat *fooling around* met de data nodig zou zijn voor we hierop een antwoord hebben.

tinyurl.com/ketqa9m

KORT

Maak kans op de Digital Preservation Awards 2014

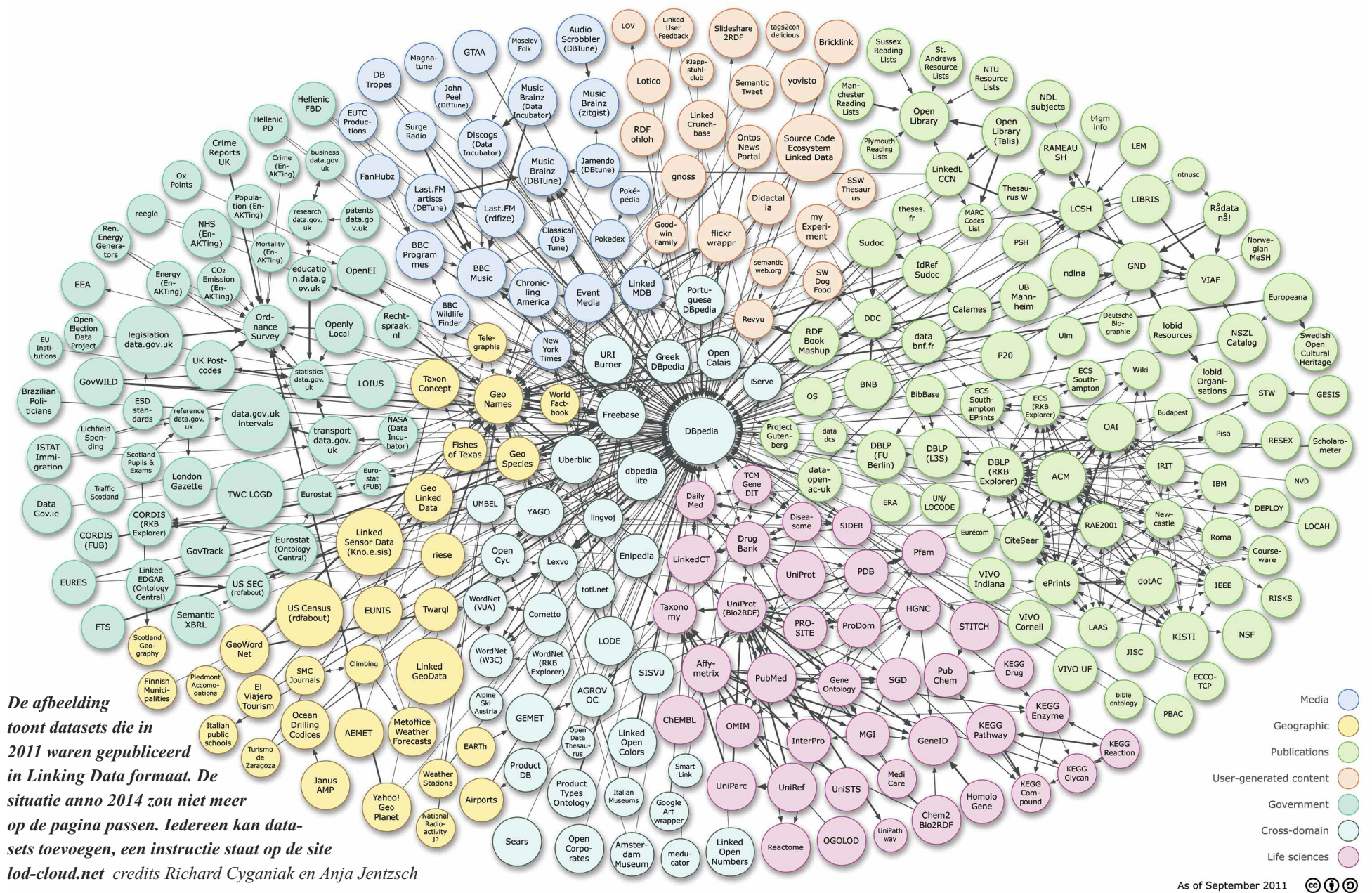
De Digital Preservation Awards worden uitgereikt aan initiatieven die bijdragen aan het duurzaam toegankelijk houden van digitaal erfgoed. Organisaties, projecten, personen en initiatieven kunnen worden genomineerd voor één van de prijzen. De prijsuitreiking vindt op 17 november plaats in Londen. De finalisten en winnaars begeven zich onder de beste projecten en initiatieven in een groeiende internationale community en zullen veel publiciteit krijgen.

De Digital Preservation Coalition (DPC) en de Nationale Coalitie Digitale Duurzaamheid (NCDD) organiseren samen de Digital Preservation Awards. Nominatie voor de Awards is open tot 28 juli via de website van de DPC. Kijk voor meer informatie op de website van de DPC. (HB) dpconline.org

Beheer Dutch Dataverse Network naar DANS

DANS heeft het beheer van het Dutch Dataverse Network (DVN) overgenomen van de Universiteit Utrecht. Het DVN is bedoeld voor onderzoekers en docenten van universiteiten en andere academische instituten in Nederland. Zij kunnen met deze dienst een grote verscheidenheid aan wetenschappelijke data digitaal opslaan in een online omgeving, op een zeer veilige en duurzame wijze. Anja Smit, directeur Universiteitsbibliotheek van de Universiteit Utrecht: "Het DVN is een wetenschappelijk platform ontwikkeld door Harvard. Na de installatie binnen de Universiteit Utrecht in 2010 werd de dienst in zeer korte tijd een gedeelde dienst van zeven instellingen. Dit succesvolle samenwerkingsverband is nu klaar voor de volgende stap. Wij hebben er alle vertrouwen in dat Dataverse door DANS verder ontwikkeld zal worden." Peter Doorn, directeur van DANS: "DANS bevordert de duurzame toegang tot digitale onderzoeksgegevens. Onze bestaande diensten (zoals NARCIS en EASY) faciliteren onderzoekers al voorafgaande aan en na afronding van hun onderzoek. Met de komst van Dataverse kunnen we onderzoekers ook tijdens hun onderzoek ondersteunen op datamanagementgebied. Bovendien zijn we verheugd met het gestelde vertrouwen in DANS door de deelnemers van Dataverse." (HB) dataverse.nl/dvn

Linked Open Data nader bekeken



Het Linked Open Data-principe is niet bij iedereen bekend. Een eerste kennismaking met voorbeelden uit de theorie en de praktijk. René van Horik

Een 'bank' kan verwijzen naar een financiële instelling of naar een meubelstuk. Wat de term betekent, wordt bepaald door de context. Als de term en de context volgens eenzelfde standaard zijn geformuleerd, kunnen deze informatiebronnen aan elkaar worden gekoppeld. En kunnen ze worden gebruikt voor bijvoorbeeld een nauwkeurigere zoekopdracht. Dit is het principe achter *Linked Open Data*. Door verbanden tussen informatie aan te brengen, ontstaan nieuwe (wetenschappelijke) inzichten. En dat biedt mogelijkheden voor archieven, bibliotheken en andere informatiebeheerders om de kwaliteit van informatiebronnen te verhogen.

Niet alleen maar rozegeur

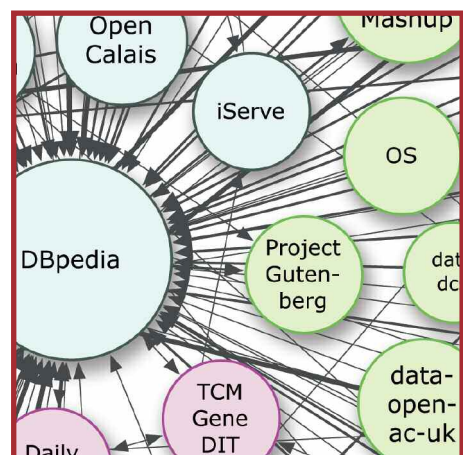
Linked Open Data zorgen ook voor nieuwe uitdagingen. Door de vele koppelingen tussen en doorverwijzingen naar verschillende informatiebronnen is de kwaliteitsgarantie en de verantwoordelijkheid voor die kwaliteit niet altijd even duidelijk. Ook veranderen informatiebronnen van inhoud, waardoor versiebeheer een punt van aandacht wordt. En er ontstaat nieuwe kennis. De vraag dient zich aan, wie de actualiteit van deze kennis bewaakt en garanties geeft voor de correctheid van de kennis. De reputatie van de dataleverancier komt hierbij in het geding.

Om *Linked Open Data* in de praktijk toe te kunnen passen en de uitdagingen het hoofd te bieden, is het essentieel dat betrokkenen gebruik maken van dezelfde standaarden. Het internationale orgaan voor webstandaarden, het World Wide Web Consortium (W3C), geeft leiding aan de ontwikkeling van deze standaarden. Op de website linkeddata.org staan informatiebronnen die als *Linked Open Data* beschikbaar en met elkaar verbonden zijn.

Langetermijntoegang

Ook op projectniveau wordt onderzoek gedaan naar de werkbaarheid van *Linked Open Data*. Een voorbeeld van zo'n project is *Preserving Linked Data (PRELIDA)*. Dit project richt zich op het verbeteren van de langetermijntoegang tot *Linked Open Data* en brengt onderzoekers bij elkaar die actief zijn op het gebied van digitale duurzaamheid en het semantisch web.

linkeddata.org
prelida.eu



Een praktijkvoorbeeld: 'hondsmoeilijk maar de moeite waard'

Linked Open Data in het juridische domein: hoe toegankelijk zijn rechterlijke uitspraken? We spreken Frank van Harmelen, hoogleraar kennisrepresentatie. Rutger Nugteren

"Onze vakgroep heeft enkele *Linked Open Data*-projecten gedaan op het gebied van toegankelijkheid van data binnen de rechtspraak. Het bleek hondsmoeilijk om binnen het domein van de rechtspraak te formaliseren en standaardiseren."

Regels van de taal

Dit blijkt ook uit het proefschrift van Marc van Opijnen. Hij stelt dat wetsteksten ongestructureerd zijn en in allerlei combinaties en stijlen voorkomen. Er zijn wel standaarden om te structureren, maar die worden niet overal toegepast. Dit komt mede doordat gerechtelijke uitspraken zijn opgesteld in de natuurlijke taal. De regels van deze taal

zijn afgeleid uit het gangbare gebruik, in plaats van opgesteld op formele wijze, zoals bij computertaal. En de schrijvers van wetsteksten nemen aan dat de lezers allerlei voor de hand liggende zaken zelf weten. Voor een computer is dat echter niet duidelijk. Wanneer je bijvoorbeeld in een formele taal moet definiëren wat een voertuig is, dan moet je vastleggen dat een voertuig gebruikt kan worden door een mens, voordat je een ongeluk met een mens in een voertuig kan beschrijven. In het medische domein zijn dit soort formele relaties vastgelegd in bestaande taxonomieën, terminologieën en classificaties. De mate van standaardisatie is daar veel groter, toch is het niet fundamenteel anders.

Nog hooguit 5 jaar

Volgens Van Harmelen zijn semantische webtechnieken dan ook toepasbaar binnen rechterlijke uitspraken, het zal alleen aanzienlijk meer moeite en tijd kosten. Dat is

ook het beeld dat Van Opijnen oproept.

Een ander obstakel bij rechterlijke uitspraken is de beschikbaarheid van brongegevens. In het proefschrift van Van Opijnen staat: "door de aanwezigheid van steeds meer data, en de beschikbaarheid ervan in voor hergebruik geschikte formaten, zijn de onderzoeksmogelijkheden in een paar jaar explosief toegenomen." "Echter," merkt Van Harmelen op: "we zijn er nog lang niet. Juist de beschikbaarheid van *Linked Open Data* laat nog zeer te wensen over. Veel context ontbreekt doordat uitgevers hun artikelen nog niet 'open' op internet zetten. De verwachting is dat dit snel zal veranderen. Onze vakgroep ontwikkelt voor Elsevier prototypes die het mogelijk maken om van artikelen naar data te linken. Het zal nog hooguit vijf jaar duren voordat de uitgeverwereld om is, en artikelen en data beschikbaar stellen en aan elkaar linken," aldus Van Harmelen.

SINDS KORT BESCHIKBAAR

Dit overzicht toont databestanden die recent beschikbaar zijn gekomen bij CentERdata, Data Archiving and Networked Services en Huygens ING.

CentERdata

• Morele bezwaren bij aankoopbeslissingen

Hoe en waarom consumenten morele bezwaren over dierenwelzijn meenemen bij hun aankoopbeslissingen, bestudeerde Janneke de Jonge in haar postdoctoraal onderzoek aan Wageningen Universiteit. In de studie ‘Consumer heterogeneity with respect to morality in consumption decisions and perceptions of animal welfare’ zijn verschillen tussen consumenten in hun percepties en voorkeuren ten aanzien van diervriendelijk vlees geïdentificeerd. Het onderzoek beoogt bij te dragen aan de ont-



Respondenten maakten in het onderzoek een serie keuzes voor kipfilet met wisselend dierenwelzijn en prijsniveau foto Janneke de Jonge

wikkeling van nieuwe welzijnsconcepten die voldoen aan de behoeften van de geïdentificeerde consumentensegmenten. Het project bestaat uit meerdere keuze-experimenten en vragenlijsten die zijn afgenomen in de maanden oktober 2012, november 2012 en mei 2013 in het LISS panel. De eerste twee delen van het onderzoek zijn beschikbaar via LISS Data Archive.

Ook sinds kort beschikbaar:

Studies LISS panel

- Antoun, C., september 2013, Nonresponse and Measurement Error in Mobile Web Surveys - Baseline

- Dalen, H. van; Henkens, K., januari/april 2013, Changing images of older workers
- Evers, E.R.K.; Lindenberg, S.M., augustus/september 2013, Collecting as Self-Regulation
- Landsheer, J.A., januari 2011, Prosocial and Antisocial Behavior
- Swan, E.; Koelen, M.; Bouwman, L.; Aarts, N., januari 2013, How is obesity perpetuated or resisted in different contexts?
- Boekel, L. van; Garretsen, H.F.L.; Brouwers, E.P.M.; Goor, L.A.M. van de, februari 2012, Perceptions and beliefs about people with a substance use disorder
- CentERdata, mei/juni 2013, Personality - Wave 6
- CentERdata, januari/februari 2013, Religion and Ethnicity - Wave 6
- CentERdata, maart/april 2013, Family and Household - Wave 6
- CentERdata, juni/juli 2013, Economic Situation: Income - Wave 6
- CentERdata, april/mei 2013, Work and Schooling - Wave 6
- Pieters, R.; Leenheer, J., maart 2013, Tilburg Consumer Outlook Monitor
- Studies Immigrant panel
- Zavala-Rojas, D.; Saris, W., maart/mei 2013, Method Effects of Language in a Survey



Deze bestanden zijn kosteloos beschikbaar via www.lissdata.nl/dataarchive. Bezoek deze site of scan de QR-code

DANS

• Wat deden archeologen 100 jaar geleden?

Vanuit het *Odyssee*-project, een initiatief van NWO, worden vele archeologische vondsten uit de periode 1900 tot 2000 in kaart gebracht en in historisch perspectief geplaatst. Dit leidt tot nieuwe wetenschappelijke inzichten binnen de archeologie. De digitale archivering van de opgravingsdocumentatie vindt plaats via het e-depot voor de Nederlandse archeologie (EDNA).



Odyssee brengt nieuwe wetenschappelijke inzichten foto Evert van Ginkel

Zo is de data, waaronder veldtekeningen, dia's, foto's en analyses, duurzaam en online toegankelijk via <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:34359>. Het overzicht van de resterende onuitgewerkte opgravingen, de werkvoorraad, is door DANS in kaart gebracht binnen het project *Odyssee in de breedte* waardoor onderzoekers en studenten zich nu over het uitwerken van oud archeologisch onderzoek kunnen ontfemen.

Ook sinds kort beschikbaar:

Sociale Wetenschappen

- Karsten, Dr. MAN., Tilburg University, Majesteitelijk en magistratisch: De Nederlandse burgemeester en de staat van het ambt. Persistent Identifier: urn:nbn:nl:ui:13-5w9g-ga
- Stichting Kiezersonderzoek Nederland; Centraal Bureau voor de Statistiek; Kolk, Dr. H. van der; Tillie, Prof.dr. J.N.; Erkel, P. van; Velden, M. van der; Damstra, A., 2012, Dutch Parliamentary Election Study 2012; Nationaal Kiezersonderzoek 2012. Persistent Identifier: urn:nbn:nl:ui:13-93iu-8p
- Sociaal en Cultureel Planbureau, 2012, Vrouwen In Besluitvorming 2012. Persistent Identifier: urn:nbn:nl:ui:13-gkpo-8j



Via easy.dans.knaw.nl zijn deze bestanden beschikbaar. Bezoek deze site of scan de QR code.

Huygens ING

• Database verzekeringsfondsen 19e eeuw online

In de 19e eeuw organiseerden de burgers, en niet de overheid, de Nederlandse sociale zekerheid. Ze richtten duizenden instellingen van onderlinge hulp op, zoals ziekenfondsen, begrafenisfondsen en weduwefondsen. De gegevens van 1.500 van deze verzekeringsinstellingen uit de periode 1827-1890 zijn uit zeer diverse bronnen bijeengebracht in jarenlang research door onderzoekster Loes van der Valk. Ze zijn nu online gepubliceerd in de *Database zieken- en begrafenisfondsen, weduwefondsen en levensverzekeringen*, een website van het Huygens ING. De database bevat voor elk fonds basisgegevens zoals de plaats van vestiging, jaar van oprichting, de bestaansduur en de voorzieningen voor hun leden. resources.huygens.knaw.nl/verzekeringsfondsen



Deze publicaties zijn beschikbaar via www.historici.nl. Bezoek deze site of scan de QR code.



De begrafenisfondsen refereerden heel direct aan de dood, zoals blijkt uit het embleem van het Haarlemse Mijn glas loopt ras bron B. Sliggers, 'In alles ghetrou' (Haarlem 1994) 14

AGENDA

12 - 13 juni • Den Haag
DHBenelux – Conference for Digital Humanities Research

Deze conferentie op het gebied van digitale geesteswetenschappen wordt voor het eerst georganiseerd. De call for proposals meldt dat ook onderzoekers van buiten de Benelux voorstellen kunnen indienen. dhbenelux.org/dhbenelux-2014-conference

2 - 5 juli • Riga (Litouwen)
LIBER Conference

LIBER is de Europese organisatie van onderzoeksbibliotheken. Haar jaarlijkse conferentie is in Riga. Het thema luidt: Research libraries in the 2020 information landscape. lnb.lv/en/liber2014

8 - 12 juli • Lausanne
Digital Humanities

Dit is de jaarlijkse internationale

conferentie van de Alliance of Digital Humanities Organizations. Het thema van DH2014 is Digital Cultural Empowerment. Naast het wetenschappelijke programma worden ook workshops en tutorials georganiseerd. dh2014.org

6 - 11 september • Dresden
CIDOC 2014

CIDOC is een internationale organisatie op het gebied van museumdocumentatie. De conferentie richt zich op de toegang tot museumcollecties en heeft als thema: Access and Understanding. Networking in the Digital Era. cidoc2014.de

8 - 12 september • Londen
Digital Libraries 2014

DL2014 is het podium voor de belangrijkste uitkomsten van Digital Library-onderzoek. De conferentie wordt georganiseerd door de Joint Conference on Digital Libraries en Theory and Practice of Digital Libraries. dl2014.org

22 - 24 september • Amsterdam
Research Data Alliance Plenary 4
Tijdens plenaire bijeenkomsten van de RDA worden resultaten en activiteiten met betrekking tot onderzoeksdata besproken. Deze vierde plenaire bijeenkomst van de RDA vindt plaats in Amsterdam. rd-alliance.org/rda-fourth-plenary-meeting.html

24 september • Amsterdam
Uitreiking Nederlandse Dataprijs
De Nederlandse Dataprijs geeft waardering aan onderzoekers die data bij elkaar brengen, documenteren en toegankelijk maken voor anderen. Op 24 september wordt de prijs uitgereikt. researchdata.nl/nl/activiteiten/dataprijs

6 - 10 oktober • Melbourne
International Conference on Preservation of Digital Objects
iPRES is een jaarlijkse conferentie op het gebied van digitale duurzaamheid. In 2014 wordt deze

conferentie voor het eerst in Australië gehouden. ipres2014.org

8 - 11 oktober • Austin
International conference on Dublin Core and metadata applications
De jaarlijkse conferentie gerelateerd aan de Dublin Core standaard heeft als thema: Metadata Intersections: Bridging the Archipelago of Cultural Memory. dcevents.dublincore.org/IntConf/dc-2014

11 - 15 oktober • Girona
Archives and Cultural Industries
Deze conferentie is een bundeling van bijeenkomsten, waaronder de Europese conferentie van de International Council on Archives (ICA) en een gerenommeerd seminar op het gebied van beeldcollecties. girona.cat/web/ica2014/eng

Pirate Bay-blokkade heeft weinig effect op downloaden

De Universiteit van Amsterdam en CentERdata hebben longitudinaal onderzoek uitgevoerd naar de effecten van het blokkeren van *The Pirate Bay* op het downloaden van films uit illegale bron (Poort, Joost, Jorna Leenheer, Jeroen van der Ham & Cosmin Dumitru (2014). *Baywatch: Two approaches to measure the effects of blocking access to The Pirate Bay. Telecommunications Policy. 38(4), 383–392*). Daarbij zijn metingen gedaan onder consumenten in het CentERpanel (zelfrapportages) en deze zijn vergeleken met gedragsdata die zijn verzameld via een nieuw ontwikkelde technologie om BitTorrent-activiteit direct te monitoren. De bevinding dat de Pirate Bay-blokkade nauwelijks effecten genereert, is in een rechtszaak actief gebruikt bij de argumentatie van een rechterlijke uitspraak. Inmiddels mogen Ziggo en XS4ALL de blokkade opheffen. (JL)

Toine Pieters doet samen met collega's in Utrecht grootschalig transnationaal onderzoek. Big data maken dat mogelijk. Maar er doemen nieuwe grenzen op: gaten in collecties, slecht leesbare teksten. Hoe gaat de wetenschap daarmee om? Inge Angevaere

“Wij bestuderen referentieculturen, dat zijn culturen die over nationale grenzen heen een voorbeeldfunctie hebben. In de zeventiende en achttiende eeuw fungeerde de Republiek der Nederlanden als een belangrijke referentiecultuur voor Europa en de Verenigde Staten. In de twintigste eeuw groeiden de VS zelf uit tot een invloedrijke voorbeeldcultuur. Wat ons interesseert, is hoe wij onze eigen (Nederlandse of Europese) identiteit ontwikkelen met verwijzing naar voorbeeldculturen. De grote databanken met gedigitaliseerde kranten zijn een onmisbare bron voor dit onderzoek. Kranten kunnen beschouwd worden als een klankbord voor wat wij ‘de publieke sfeer’ noemen. Niet alleen artikelen, maar ook ingezonden brieven, cartoons en advertenties leveren ons belangrijke informatie over de wisselwerking tussen wetenschap, cultuur en economie.”

“In het Europese HERA-project *AsymEnc* (*Asymmetrical Encounters*) zoeken we naar de historische dynamiek van culturele concepten zoals ‘metropolis’ in grote digitale krantenbanken in Nederland, Duitsland en Engeland. Voor het eerst proberen we systematisch een grootschalig longitudinaal onderzoek te doen in diverse landen. Zonder big data en text mining zou dat bijzonder lastig zijn.”

Big data, new questions

“Maar we lopen ook tegen grote nieuwe vragen aan. Over hoe we dit materiaal wetenschappelijk kunnen en mogen gebruiken. Want er zitten grote gaten in de digitale krantencollecties. Wat hebben die voor effect op onze onderzoeksresultaten? Ook de uitgelezen teksten (OCR) zijn vaak van slechte kwaliteit. Kunnen de bibliotheken daar iets aan doen of moeten we daarmee leren leven? En hoe dan?”

“Afgelopen april organiseerde ik met collega

Jaap Verheul de workshop ‘*Mining digital repositories: challenges and horizons*’ in de Koninklijke Bibliotheek. Daar zijn die vragen uitgebreid aan de orde gekomen. Tussen onderzoekers onderling, maar ook in een dialoog met de bibliotheken en commerciële partijen die de krantencollecties beheren en ontwikkelen.”

“Tijdens die workshop hebben we natuurlijk niet alle vragen kunnen beantwoorden. Maar wat mij betreft was het een uitstekend begin van de dialoog die moet leiden tot een nieuwe wetenschappelijke cultuur rond big data. En ook tot goede afspraken met bibliotheken. Zo willen wij als onderzoekers méér informatie over wat er in de databanken zit en wat niet, en over de wisselende kwaliteit van de OCR.



Toine Pieters in zijn werkkamer foto Inge Angevaere

Toine Pieters over grootschalig transnationaal onderzoek:

‘We moeten nog leren werken met big data’

Zodat we daar rekening mee kunnen houden bij ons onderzoek.”

Crowdsourcing belangrijk

“Bibliotheken zijn openbare instellingen met beperkte budgetten. Dat is natuurlijk een gegeven. En technisch verandert de situatie voortdurend. De machines waarmee nu OCR wordt gemaakt, zijn veel beter dan een jaar of tien geleden. Maar dat oudere OCR-spul zit nog steeds in de databanken. Ik denk dat er

maar één betaalbare manier is om die situatie te verbeteren: *crowdsourcing*. Onderzoekers, studenten en amateurhistorici toestaan om de uitgelezen teksten te verbeteren. Bibliotheken hebben daar niet altijd open voor gestaan. Maar nu lijkt het tijd voorzichtig te keren.

Ik bespeurde tijdens de workshop bij de KB althans een positieve grondhouding.”

“En dan natuurlijk het auteursrecht. Er is zo-

veel materiaal waar we als onderzoekers nog niet bij kunnen, dat is doodzonde. Het krantenarchief van de British Library, bijvoorbeeld, is helemaal in commerciële handen. Persoonlijk vind ik dat je individuele onderzoekers niet lastig moet vallen met al die auteursrechtenkwesties. Dat bibliotheken het op zich moeten nemen om toegang voor wetenschappelijk onderzoek te regelen met de rechthebbenden.

Liefst in Europees verband. En als het moet onder strikte voorwaarden dat het materiaal niet verder verspreid wordt. Daar staan we best open voor.”

Omgaan met taalkwesties

“Taalkwesties zijn een andere factor in transnationaal onderzoek waarmee we moeten leren werken. Je kunt concepten niet zomaar vertalen en er dan van uitgaan dat die vertaalde termen dezelfde lading hebben als het origineel. Woordbetekenissen veranderen ook voortdurend. Wat op het ene moment een neutraal woord is, kan tien jaar later een positieve of een negatieve bijklank hebben. We moeten tools ontwikkelen om die finesse uit de databestanden te kunnen halen. Tijdens onze workshop vroeg iemand of je kunt analyseren waarover niet gesproken wordt in kranten. Tja, dat gaat nog weer een stap verder.”

“We vroegen tijdens de workshop ook aan collega's hoe hun ‘digitale utopia’ eruit zou zien. De een wilde vooral de OCR verbeteren, de ander wilde méér materiaal. Het mooiste vergezicht kwam denk ik van mijn Utrechtse

‘*digital history*’ collega Joris van Eijnatten. Hij sprak van een ‘Globiana 5.0’: een wereldwijd, transnationaal digitaal archief vol mate-

riaal in standaardformaten en gekoppeld aan tweetalige en meertalige woordenboeken. Die moeten we dan wel kunnen onderzoeken met een netwerk van wetenschappers die goed kunnen omgaan met big data. Collega Rens Bod van het Amsterdamse Centre for Digital Humanities benadrukte dat hier nog veel werk te verrichten is. Veel onderzoekers missen nog big data-vaardigheden, en dat zijn lang niet alleen de ouderen onder ons. Het gaat hier letterlijk om *capacity building* met als doel ‘*digital empowerment*’ in de geesteswetenschappen.”

tal Humanities benadrukte dat hier nog veel werk te verrichten is. Veel onderzoekers missen nog big data-vaardigheden, en dat zijn lang niet alleen de ouderen onder ons. Het gaat hier letterlijk om *capacity building* met als doel ‘*digital empowerment*’ in de geesteswetenschappen.”

Datalabs

“Onderdeel van die ‘*digital empowerment*’ is ook dat bibliotheken een andere rol nemen ten opzichte van de data die ze beheren. De British Library heeft daarin het voortouw genomen en de Koninklijke Bibliotheek volgt binnenkort. Ze stellen zich niet langer op als passieve leverancier van hun digitale data, maar zoeken actief de samenwerking met wetenschappers om méér met die data te kunnen doen. In zogenaamde ‘datalabs’ werken wetenschappers samen met data-experts van de bibliotheken. Met voornamelijk opensource-tools worden de mogelijkheden van big data verkend. Van die samenwerking worden beide partijen rijker.”

tinyurl.com/ketqa9m

INTERVIEW

“Hoe ziet uw digitale utopia eruit?”

“Het gaat om ‘digital empowerment’ in de geesteswetenschappen”

Toine Pieters

Toine Pieters is verbonden aan het Descartes Centrum voor Wetenschapsgeschiedenis en -filosofie van de Universiteit Utrecht. In april maakte hij de overstap van de afdeling Metamedica van het VU Medisch Centrum naar een aanstelling als hoogleraar/directeur van het Freudenthal Institute for Science and Mathematics Education. Hij is projectcoördinator van Translantis, ASYMENC en projectleider van het digitale erfgoedproject Timecapsule.

Het NPSO adviseert over hoe je survey-onderzoek goed uitvoert. Met voorzitter Vera Toepoel spreken we over de status en toekomst van survey-onderzoek.

Marike de Bruijne

“Iedereen kan tegenwoordig een survey doen, en doet dat ook”, begint Toepoel nadrukkelijk. Ze heeft er een reden voor. Er worden tegenwoordig te veel vragenlijsten afgenomen van wisselende kwaliteit waardoor mensen moe worden van enquêtes. “*Do-it-yourself-surveys* kunnen door iedereen worden afgenomen, zoals de buurvrouw of mensen van de sportvereniging. Vervolgens worden algemene uitspraken gedaan, terwijl er geen kanssteekproef is gebruikt”, legt Toepoel uit. “Wat is nu het verschil tussen goede en slechte onderzoeken?”

Checklist

Als antwoord op deze vraag heeft het NPSO een checklist ontwikkeld. Met deze checklist kunnen geïnteresseerden inschatten of een onderzoek deugt of niet. De checklist is beschikbaar op de NPSO-website. Ook organiseert het NPSO regelmatig bijeenkomsten en lezingen; deelname is kosteloos. Dergelijke acties vormen de kern van het NPSO. “We willen onderzoekers en iedereen die te maken heeft met survey-onderzoek helpen en advies geven over hoe je goed onderzoek kunt doen. Op onze website staat bijvoorbeeld wie je met welke expertise kunt benaderen bij vragen. Ook is er een LinkedIn-groep waar je je voor kunt aanmelden.”

Driehoek

Het NPSO is een initiatief van partijen uit de wetenschap, overheid en commercieel onderzoek. Zo’n drie-



Vera Toepoel: “Met de door ons ontwikkelde checklist kunnen geïnteresseerden inschatten of een onderzoek deugt of niet” foto Bart van Vliet

NPSO adviseert in goed survey-onderzoek

hoek biedt een kans om van elkaar te leren en kennis te verspreiden. Toepoel: “Hoewel de wetenschappers vaak achter de nieuwste ontwikkelingen aan lopen, kunnen ze de markt helpen bij het valideren van onderzoeksmethoden of door opleidingen te bieden. Het marktonderzoek kan weer ervaringen vanuit de praktische uitvoering delen. Overheid en semi-overheid kunnen vertellen wat hun behoeften zijn en wat er in de maatschappij speelt, en bovendien advies van de andere groepen gebruiken.”

Toekomst surveys

Maar terug naar de respondent. Hoe kun je hem of haar tegenwoordig

nog bereiken? Toepoel denkt na. “Het invullen van een survey is een kosten-batenafweging. Als onderzoeker moet je beseffen dat je met elke survey de kostbare tijd van de respondent claimt”, zegt ze. “Om ervoor te zorgen dat een respondent deelneemt, is het van belang om surveys interessanter en vermakelijker te maken en uit te leggen wat het belang is van een onderzoek.”

Meer vooronderzoek

Verder ziet ze nieuwe technieken zoals mobiele surveys als een noodzakelijke toepassing voor deze tijd. “Over een aantal jaar heeft bijna iedereen een mobiele telefoon. Het volstaat dan niet meer om vragenlijsten alleen voor desktops te ontwikkelen. Een respondent zit niet meer achter een scherm, maar staat bij een bushalte een korte vragenlijst in te vullen.”

Volgens Toepoel moeten we ons bij

Nederlandstalig Platform Survey Onderzoek

Doel: adviseren over hoe je survey-onderzoek goed uitvoert.
Middel: het Nederlandstalig Platform voor Survey Onderzoek (NPSO).
Idee: een initiatief van wetenschappers, marktonderzoeksbureaus en overheidsinstellingen.
Voorzitter: Vera Toepoel, Universitair docent bij Methoden en Statistiek aan de Universiteit Utrecht.
Meer informatie: wil je op de hoogte blijven van de activiteiten van het NPSO? Meld je dan aan op nps0.net.

KORT

Een database van wetenschappelijke instrumentmakers

Het Digitaal Wetenschaps-historisch Centrum van Huygens ING heeft samen met Museum Boerhaave biografische gegevens over Nederlandse makers van wetenschappelijke instrumenten bijeengebracht. Eén van de verworven inzichten in de wetenschapsgeschiedenis van de laatste decennia is het gegeven dat in de vroegmoderne tijd de kruisbestuiving tussen ambachtslieden en geleerden van vitale betekenis is geweest voor de ontwikkeling van de natuurwetenschap. Instrumentmakers, bijvoorbeeld, opereerden op het snijvlak van ambachtelijke en geleerde kennis en hebben door hun specifieke expertise soms belangrijke bijdragen geleverd aan de ontwikkeling van wetenschappelijke instrumenten. Hun inbreng is echter vaak onderbelicht, reden waarom de Amerikaanse wetenschaps-historicus Steven Shapin ze als *invisible technicians* heeft aangeduid. De database borduurt voort op het werk van Maria Rooseboom (1909-1978), directrice van de voorganger van het huidige Museum Boerhaave. In de instrumentenmakersdatabank worden deze gegevens aangevuld en ontsloten. Zo wordt onderzoek gefaciliteerd naar mensen die traditioneel niet binnen de categorie van ‘wetenschappers’ worden gerekend. De database is een onderdeel van het *Biografisch Apparaat*, een database van wetenschappelijk actieve personen in Nederland, die later dit jaar zal worden gepubliceerd. (HZ, IN)

www.dwc.knaw.nl/biografie/scientific-instrument-makers/

Over vrouwen en het literaire veld

Binnen het literatuurhistorisch onderzoek ontbreekt het aan een digitale structuur waarbinnen de receptie van auteurs kan worden geanalyseerd. Twee projecten werken aan een oplossing. Suzan van Dijk en Walter Ravenek

Het ontbreken van zo’n structuur was vooral een probleem met betrekking tot vrouwelijke auteurs. Voor velen van hen zijn recensies te vinden in de achttiende- en negentiende-eeuwse pers, maar hun namen zijn vaak afwezig in moderne literatuurgeschiedenissen. Biografieën bieden geen oplossing, die lichten ons niet in over de belang-

rijkste vragen inzake het vrouwelijk auteurschap: de vragen naar (1) de mate van integratie van de dames in de literaire wereld en (2) de representativiteit van individuele gevallen.

Database WomenWriters

De behoefte aan een brede aanpak van deze problematiek werd ruim tien jaar geleden al duidelijk. NWO digitaliserings- en internationaliseringsgelden leidden tot de Europese *COST Action Women Writers in History* (geleid vanuit Huygens ING, 2009-2013) en tot het doorontwikkelen van een aanvankelijk simpele database getiteld *WomenWriters*. *COST Action*-leden uit 25 landen hebben deze gebruikt en uitgetest

met het oog op voor hen relevante vraagstellingen. Dit leidde tot twee Europese samenwerkingsprojecten: *Travelling Texts 1790-1914*. *The Transnational Reception of Women’s Writing at the Fringes of Europe* (2013-2016) en *COBWWWEB: Connections Between Women and Writings Within European Borders* (2013-2014).

Europese schrijfsters

Het HERA-project *Travelling Texts* bestudeert de internationale invloed van Europese schrijfsters aan de hand van hun receptie in vijf landen (Finland, Nederland, Noorwegen, Slovenië, Spanje). Gegevens over de interesse van lezers voor hun werk - blijkend uit bijvoorbeeld de

aankoop van boeken, uit commentaren in pers of privé-correspondentie, uit vertalingen - vormen de basis voor het onderzoek naar het functioneren van schrijfsters in hun eigen tijd en naar hun historische belang.

Gendersolidariteit

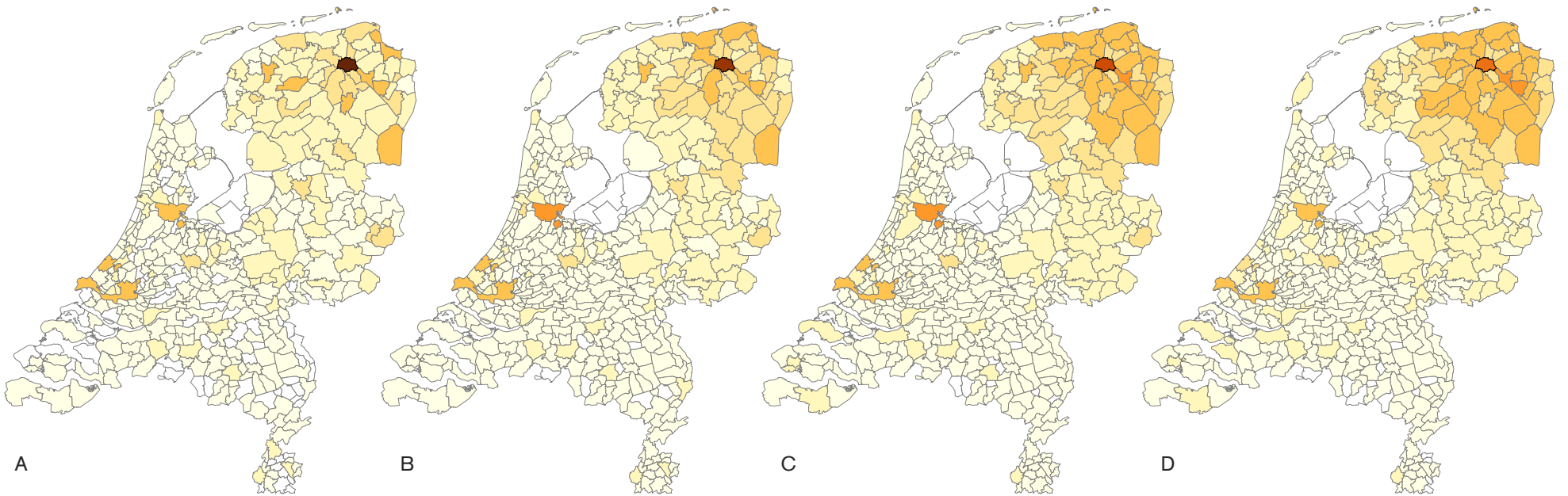
Door dit op internationale schaal te bestuderen, kan duidelijk worden gemaakt dat transnationale gendersolidariteit als een motor kan hebben gefunctioneerd. Het CLARIN-project *COBWWWEB* (Nederland, Noorwegen, Servië, Zweden en Zwitserland) bekijkt hoe het technisch mogelijk is om vrouwelijke auteurs te behandelen als onderdeel van het totale literaire veld (m/v). Het onderzoeksmateriaal wordt op-

geslagen in een CLARIN-compliant data-repository, gebruik makend van de door Huygens ING ontwikkelde software Timbuctoo. Hierin kunnen verschillende projecten verschillende varianten van gegevens opslaan. Zo kunnen bijvoorbeeld meerdere interpretaties van recensies van romans naast elkaar bestaan; elk project heeft zijn eigen versie en kan interpretaties van andere onderzoekers al dan niet overnemen. De repository bevindt zich in de opbouwfase; naast de gegevens uit de oude database worden nu Noorse en Servische gegevens toegevoegd. Na datacuratie zullen de gegevens voor iedereen te raadplegen zijn.

databasewomenwriters.nl
huygens.knaw.nl/timbuctoo

Vier generaties migratie in beeld

Hoe honkvast zijn Groningers?



De herkomstkaarten uit Migmap geven de geboorteplaats aan van de huidige inwoners van de gemeente Groningen in de leeftijd 30 tot en met 50 jaar (kaart A), de herkomst van hun ouders (kaart B), van hun grootouders (kaart C) en van hun overgrootouders (kaart D). Op deze kaarten is duidelijk te zien dat de voorouders van de huidige Groningers tussen de 30 en 50 jaar hoofdzakelijk uit dezelfde regio komen. In de online applicatie geeft een mouse-over het exacte percentage per gemeente aan en is in een tabel de herkomst van Nederlanders uit het buitenland opgenomen. *credit Jan Pieter Kunst*

De Migmap-applicatie van het Meertens Instituut brengt het verhuisgedrag van de Nederlandse bevolking van de afgelopen honderd jaar in kaart.

Erica Renckens

Volgens onderzoeker Gerrit Bloothoof is dit het opvallendste resultaat dat is gevonden met behulp van de Migmap-applicatie. “De verspreiding van de bevolking is beperkt; men hecht blijkbaar aan de eigen sociale omgeving.”

Migmap maakt gebruik van de gegevens over

geboorteplaats, woonplaats en generatie uit de *Gemeentelijke basisadministratie persoonsgegevens (GBA)*. Op basis hiervan genereert de online applicatie kaarten waarop staat hoe generaties uit een bepaald gebied zich verspreid hebben. De gebruiker kan kiezen voor een gemeente, regio, dialectgebied of provincie, waarna de resultaten in percentages per gemeente getoond worden. Eventueel kan hij er ook voor kiezen alleen de vrouwelijke of juist de mannelijke lijn te volgen.

Twee soorten online kaarten

Herkomstkaarten geven aan waar de huidige bewoners oorspronkelijk vandaan komen, verspreidingskaarten tonen waar de nakomelingen van de bewoners van een eeuw geleden te-

recht zijn gekomen. “Het pluspunt van Migmap is dat de resultaten niet op een steekproef gebaseerd zijn, maar op de hele bevolking,” aldus Bloothoof, die de algehele leiding had over het door CLARIN-NL gesubsidieerde project.

“Heel bijzonder is dat de verspreiding is gekoppeld aan familierelaties tot vier generaties diep.” Het door de GBA geleverde gegevensbestand komt uit 2006 en bevat de gegevens van ruim 22 miljoen mensen met de Nederlandse nationaliteit. De GBA is sinds 1994 digitaal. “Omdat we niet verwachten dat er van vóór 1994 betere woon- en verhuisgegevens beschikbaar komen, is de applicatie voorlopig afgerond,” licht Bloothoof toe. Migmap is voor iedereen online te gebruiken, maar de

applicatie is hoofdzakelijk bedoeld voor taalkundigen en sociaal-economische onderzoekers, aldus de projectleider. “Voor hen kan het van belang zijn inzicht te krijgen in bevolkingsveranderingen, zoals de herkomst van de huidige bevolking of de verspreiding van de bevolking van een eeuw geleden. Een gebied met een stabiele bevolking kan duiden op behoud van tradities, bijvoorbeeld in dialect, en een stabiele werkgelegenheid.”

De kaarten die Migmap genereert zijn in groot formaat te downloaden en af te drukken. Daarnaast kan de gebruiker de kaarten ook als KML-bestand downloaden zodat de migratiekaart gecombineerd kan worden met andere kaarten in Google Earth.

meertens.knaw.nl/migmap

De Koninklijke Bibliotheek onderzoekt informatiebehoefte

Vrije toegang tot de wetenschap voor iedereen?

De Koninklijke Bibliotheek heeft de behoefte aan wetenschappelijke informatie onder de Nederlandse bevolking onderzocht. Projectleider Astrid van Wesenbeeck vertelt over de resultaten van dit onderzoek. Inge Angevaere

“Veel wetenschappelijke artikelen zijn momenteel alleen tegen betaling aan de uitgevers beschikbaar, tenzij je toegang hebt via de bibliotheek van je universiteit. In het licht van de open access-discussie wilden we weten of er ook buiten de wetenschap behoefte is aan die informatie.”

Voor eigen interesse

De KB peilde in samenwerking met TNS NIPO de mening van 44.000 Nederlanders van 25 jaar en ouder. Van Wesenbeeck: “13% van de onder-

vraagden gaf aan interesse te hebben in het lezen van wetenschappelijke artikelen en 61% van deze groep leest al circa tien wetenschappelijke artikelen per jaar. Men gaf aan voornamelijk voor eigen interesse te lezen; daarna volgden werk en studie. Medische wetenschappen en sociale wetenschappen werden als meest populaire vakgebieden genoemd door de lezer die voor eigen interesse leest, gevolgd door geschiedenis. Belangstellenden die lezen voor hun werk waren opvallend vaak geïnteresseerd in exacte wetenschappen.”

Belanghebbenden

“Nu Staatssecretaris Sander Dekker van Onderwijs, Cultuur en Wetenschap heeft aangedrongen op vrije toegang tot wetenschappelijke artikelen, zijn verschillende belanghebbenden, zoals uitgevers, biblio-

theken en universiteiten, met elkaar in gesprek hoe de overgang naar vrije toegang kan worden gerealiseerd. Ook de KB is hierbij betrokken, mede omdat vanaf 2015 een deel van het werkterrein van de openbare bibliotheken onder het mandaat van de KB valt. In Engeland heeft een soortgelijke ontwikkeling geleid tot de website *Access to Research*, een tweejarige pilot waarin vrije toegang tot wetenschappelijke artikelen door de uitgevers wordt aangeboden in openbare bibliotheken.”

“Het zou mooi zijn als Nederland zo’n voorbeeld zou volgen. Maar we moeten niet uit het oog verliezen dat de echte meerwaarde bestaat in plaats- en apparaatonafhankelijke toegang. 100% open access dus. Als de ambities van Dekker waar worden gemaakt, zijn we in 2024 zo ver.”

accessstoresearch.org.uk



Astrid van Wesenbeeck foto Inge Angevaere

Communiceren machines straks in menselijke taal?

Vrijdag 6 juni zijn de Spinoza-premies 2014 uitgereikt, de hoogste onderscheiding in de Nederlandse wetenschap. Taaltechnoloog Piek Vossen was vorig jaar één van de laureaten. *Erica Renckens*



Taaltechnoloog Piek Vossen, winnaar van de Spinozapremie 2013 foto Studio VU

Naast eeuwige roem ontvangt de winnaar van de Spinoza-premie ook 2,5 miljoen euro. “Daarmee heb ik vier projecten gedefinieerd met als rode draad de vraag hoe computers natuurlijke taal kunnen begrijpen,” vertelt Piek Vossen, die sinds 2006 werkzaam is aan de Vrije Universiteit. “Het is nog een groot mysterie hoe we dat computers kunnen leren, terwijl wij mensen er ogenschijnlijk geen moeite mee hebben.”

Eerder maakte Vossen voor de Europese Unie *wordnets* in acht talen: netwerken van woorden die op basis van hun betekenis met elkaar verbonden zijn. Soms zijn de woorden synoniemen van elkaar, soms juist tegenstellingen, en soms hebben ze alleen met elkaar gemeen dat ze met dezelfde kleur geassocieerd worden. Met behulp van zulke *wordnets* kan een computer betekenis afleiden en zo spelling controleren, zoekresultaten verbe-

teren en teksten interpreteren.

Betekenis is voor computers een lastig aspect van taal. “Een woord als ‘slag’ kan in het Nederlands achttien verschillende betekenissen hebben, maar als we een tekst lezen zien we er maar één,” aldus Vossen. Mensen leiden aan de hand van zinsbouw en context moeiteloos af welke betekenis bedoeld is, maar een computer heeft deze kennis niet.

Semantiek centraal

“In het eerste project gaan we het probleem van de meerduidigheid van woorden aanpakken.” Ook in

de overige drie projecten staat de semantiek van natuurlijke taal centraal. “Het is de bedoeling dat we meer fundamenteel inzicht krijgen in de manier waarop taal betekenis krijgt,” legt Vossen uit. Daarvoor kijken de onderzoekers in zijn groep onder andere naar de relatie tussen woorden, concepten, perceptie en het brein, naar de invloed van het wereldbeeld van de auteur op teksten en naar de rol van de lezer bij het begrijpen van teksten. “Uiteindelijk hopen we machines te kunnen maken die met mensen kunnen communiceren aan de hand van natuurlijke taal. Daarbij

richten we ons niet direct op spraak, maar wel op de betekenis.”

Resultaten delen

Vossen: “Ieder project kijkt naar andere deelaspecten, maar er zijn ook veel verbanden waarbij projecten elkaars deelresultaten kunnen gebruiken. We hebben tien mensen aangesteld, onze groep is verdubbeld. We werken enthousiast samen aan een grote uitdaging, ik hoop dat de goede sfeer en samenwerking die we nu hebben ook in een grotere groep blijft bestaan.”

vossen.info

GELEZEN

Managing and Sharing Research Data. A guide to good practice. Louise Corti, Veerle van den Eynden, Libby Bishop & Matthew Woollard. Uitgeverij Sage. 2014. René van Horik

Vier auteurs, allen werkzaam bij het UK Data Archive, schreven dit handboek over het beheren en delen van onderzoeksdata. Men richt zich zowel op de onderzoeker (van student tot professor) als op de dataprofessionals, de ondersteuners. Het handboek biedt een actueel en compleet overzicht. Jammer dat het niet als open access publicatie beschikbaar is. Het zou dan veel meer lezers bereiken. Niet alle elf hoofdstukken zijn van belang voor alle doelgroepen. Dus het is zaak goed de inhoudsopgave te bekijken en te bepalen welk onderdeel bruikbaar is. Elk hoofdstuk bevat een aantal praktische oefeningen en een uitgebreide literatuurlijst. Ook is er een begeleidende website met achtergrondinformatie en databronnen die behandeld worden. Het boek is sterk gericht op de Engelssprekende wereld en mist daarom een aantal belangrijke niet-Engelstalige initiatieven. Er is relatief veel aandacht voor *legal and ethical issues in sharing data*, met goed uitgewerkte oefeningen, bijvoorbeeld manieren om onderzoeksdata te anonimiseren. Voor de onderzoeker zijn met name de laatste drie hoofdstukken interessant. Ze hebben betrekking op collaborative research, gebruik van onderzoeksdata van anderen en het publiceren en citeren van onderzoeksdata. ukdataservice.ac.uk/manage-data/handbook

COLUMN

Waarom is er nog steeds geen digitaliseringsregister?

In 2004 was ik bij een bijeenkomst in de Koninklijke Bibliotheek over massadigitalisering. Eén van de sprekers was Astrid Verheusen, programmaleider digitale bibliotheek. Met enige gêne stipte Verheusen één van de centrale problemen van massadigitalisering aan, namelijk het ontbreken van een register waarin wordt vastgelegd wie wat aan het digitaliseren is.

Zij liet afbeeldingen zien van enkele reeksen die door diverse bibliotheken en instellingen waren gedigitaliseerd. Steeds dezelfde reeks, steeds met overheidssubsidie.

In de zaal ging een besmuikt gelach op – het probleem was algemeen bekend.

Dat was tien jaar geleden.

Een paar maanden geleden was ik bij een internationale bijeenkomst in de Koninklijke Bibliotheek over textmining. Een Amerikaanse spreker vroeg onder meer: ‘Weet iemand of er een centraal register bestaat waarin je kunt nazien wie wat heeft gedigitaliseerd? Ik heb de indruk dat er veel dubbel werk

wordt gedaan en dat lijkt mij zonde van het geld en van de inspanning.’ In de zaal bleef het stil.

Is het werkelijk van belang dat er zo’n register komt? Ja, en niet alleen omdat het zonde is van geld en tijd dat er – nog steeds – zoveel veel dubbel werk wordt verricht. Een andere reden houdt verband met een voortkabbelende discussie in E-data.

In oktober schreef ik in E-data dat ik iemand ken die af wil van een collectie van 13.000 boeken. Het gaat om een met smaak en kennis opgebouwde collectie boeken uit en over de 19de eeuw. Je zou zo’n specialistische collectie aan de KB moeten kunnen schenken, stelde ik, onder voorwaarde dat je

’m binnen een bepaalde tijd in digitale vorm terugkrijgt. De redactie van E-data legde dit idee voor aan Jan Bos, hoofd collecties van de KB. Zijn antwoord, kort samengevat: de KB zou die dienst graag willen leveren, maar kan dit niet. Bos: “Met name de selectie is veel duurder dan je zou denken. Je moet eerst nagaan wat we al gedigitaliseerd hebben, en van wat er overblijft moet je alle rechten napluizen.”

Inderdaad, het napluizen van rechten kan een tijdrovende en dus kostbare zaak zijn. Maar nazien wat er al gedigitaliseerd is, zou een fluitje van een cent moeten zijn. Niet alleen de KB, maar alle instellingen die digitaliseren, zouden dit met bepaalde technische specificaties moeten vastleg-

gen in een register. In PiCarta of WorldCat bijvoorbeeld. Als niet alleen bibliotheken dat wereldwijd zouden doen, maar bijvoorbeeld ook Google, zou dit heel veel dubbel werk kunnen voorkomen.

Waarom is het in tien jaar niet gelukt om zo’n centraal register op poten te zetten? Desinteresse, eigenwijsheid en onwil, vermoed ik. ‘Eigen instel-

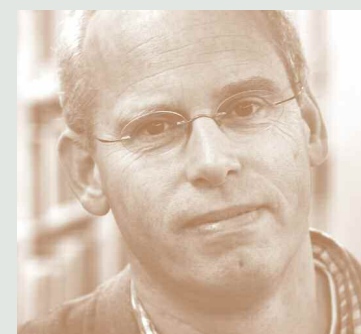


foto Leo van Velzen

ling eerst’ is het motto van velen, het algemeen belang staat veel lager op de ladder.

Ik heb weleens voorgesteld om instellingen alleen geld voor digitalisering te geven als ze kunnen aantonen dat ze meewerken aan een centraal register. Dat was in 2011, naar aanleiding van een stuk in NRC Handelsblad getiteld ‘Het digitale drama’. Uit dat stuk bleek onder meer dat niemand wist hoeveel geld er tussen 2004 en 2011 aan digitalisering was uitgegeven. De schattingen liepen uiteen van 50 tot 200 miljoen euro.

En waarom was en is daar zo slecht zicht op te krijgen? U voelt ’m al aankomen: mede vanwege het ontbreken van een centraal register.

Stuitend hoe lang het soms duurt voordat een idee dat zo voor de hand ligt, wordt ingevoerd.

Ewoud Sanders

Taalhistoricus en journalist. Sanders is vaste medewerker van onder meer NRC Handelsblad en Onze Taal.