E-DATA & RESEARCH

Jaargang 10 | nummer 2

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

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

Hoe maak je gearchiveerde websites bruikbaar voor de wetenschap?

Nationale webarchief onderzocht door WebART

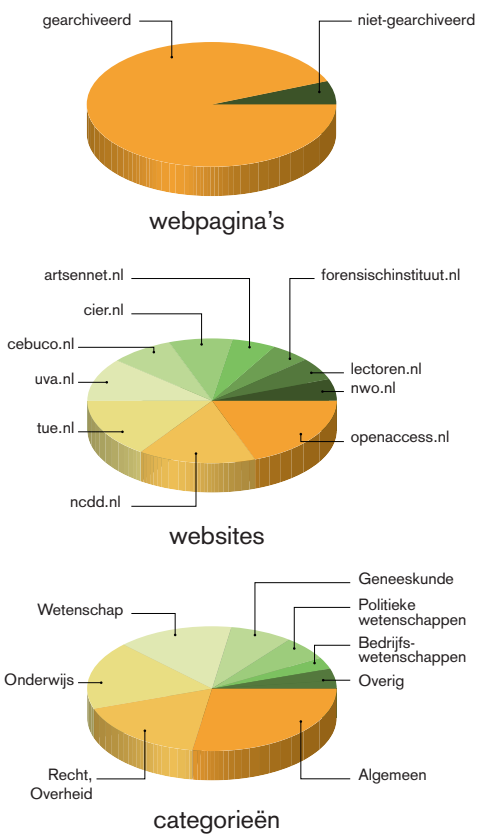
Het eerste grote onderzoeksproject in Nederland naar gebruik van gearchiveerde Nederlandse websites als primaire bron voor onderzoek sluit binnenkort de boeken. WebART-promovendus Hugo Huurdeman blikt terug.

Steven Claeyssens

Dit jaar ronden de laatste CATCH-projecten (Continuous Access to Cultural Heritage) hun werkzaamheden af en dus zet ook WebART (Web Archive Retrieval Tools) er een punt achter.
WebART was een samenwerking tussen de Universiteit van Amsterdam (UvA), het Centrum Wiskunde en Informatica (CWI) en de Koninklijke Bibliotheek (KB). Het WebART-team lichtte als eerste het Nederlandse nationale webarchief grondig door. Ze gingen daarbij na hoe zo'n heterogeen en omvangrijk *born-digital* archief voor onderzoeksdoeleinden bruikbaar kan zijn en bruikbaarder kan worden gemaakt.

10.000 websites

De KB archiveert sinds 2007 een immer groeiende selectie van Nederlandse websites. Op 1 januari van dit jaar stond de teller op 10.000 sites die met enige regelmaat worden *geharvest*. Het belang van dit *born-digital* archief voor onderzoek naar Nederlandse cultuur en samenleving zal naarmate de jaren verstrijken onvermijdelijk een steeds prominentere plaats opeisen. WebART onderschrijft dit belang en trok op onderzoek uit. Huurdeman: "In het WebART-project hebben we gekeken naar de onderzoeksvragen die wetenschappers aan webarchieven zouden willen stellen. Via een intensieve samenwerking met nieuwe media-onderzoekers hebben we vervolgens zoek- en onderzoekstools ontwikkeld die complexe onderzoekstaken kunnen ondersteunen. Denk bijvoorbeeld aan de initiële exploratie van het archief, het definiëren van een dataset en de analyse daarvan. Hiervoor was onderzoek nodig naar schaalbare extractie- en analysemethoden en naar bruikbare interfaces voor verschillende zoekstadia." Zo bouwde het team onder meer WebARTist, een interface waarmee onderzoekers op verschillende manieren het webarchief kunnen verkennen en bevragen.



Op basis van de zoekterm 'onderzoeksdata' toont WebARTist verschillende resultaten, waaronder deze grafieken. De bovenste grafiek laat de verhouding tussen de gearchiveerde en niet-gearchiveerde webpagina's zien, de middelste toont de belangrijkste websites voor deze zoekterm en de onderste grafiek vat de categorieën van de gevonden pagina's samen. De WebARTist-toolset biedt een veelheid aan mogelijkheden voor exploratie, analyse en visualisatie van de inhoud van het KB-webarchief. *credits WebART*

Ongearchiveerde websites

"Doordat webarchieven van nature incompleet zijn, vroegen wetenschappers ook om contextualisatie over wat er wel en niet in het archief zit. Dit heeft geleid tot verder onderzoek waarin we niet-gearchiveerde webinhoud hebben blootgelegd en gereconstrueerd." Zo slaagden Huurdeman en zijn mede-onderzoekers erin een fors aantal niet-gearchiveerde sites te identificeren op basis van verwijzingen in de vorm van URL's in het wel-gearchiveerde deel. Meer nog, door de afzonderlijke woorden uit deze URL's en de bijbehorende linkteksten te distilleren, maakten ze dit niet-gearchiveerde deel van het web tot op zekere hoogte toch vindbaar en daarmee ook onderzoekbaar.
"Deze informatie integreren we in de WebART-toolset. Helaas kan de toolset momenteel door auteursrechtelijke beperkingen nog niet volledig online worden aangeboden, maar de wens vanuit het projectteam om dit te bereiken, is er zeker."
webarchiving.nl

Principeakkoord open access VSNU en Elsevier

De Vereniging van Universiteiten (VSNU) en Elsevier hebben een principeakkoord bereikt waardoor Nederlandse wetenschappers toegang blijven houden tot de wetenschappelijke artikelen van Elsevier.

"Door deze overeenkomst," aldus prof. Gerard Meijer, hoofdonderhandelaar namens de VSNU en voorzitter van de Radboud Universiteit Nijmegen, "houden wetenschappers toegang tot Elseviertijdschriften en het biedt ze de mogelijkheid om in een selectie van die tijdschriften open access te publiceren. De universiteiten streven ernaar dat in 2018, het derde jaar van de overeenkomst, 30% van de Elsevierartikelen van Nederlandse auteurs open access beschikbaar is. Dit akkoord maakt dat mogelijk. Dit is echt geweldig nieuws en een 'big deal' voor open access." Philippe Terheggen, Managing Director Journals bij Elsevier: "Wij zijn content met deze overeenkomst, omdat blijvende subscriptietoegang tot onze hoogwaardige, 'peer-reviewed' wetenschappelijke artikelen essentieel is voor Nederland om zijn positie als één van de meest impactvolle onderzoekslanden te behouden. Daarnaast krijgen Nederlandse wetenschappers meer open access publicatiemogelijkheden om hun onderzoeksresultaten met de rest van de wereld te delen." De overeenkomst is in lijn met de ambitie van staatssecretaris Dekker (OCW), die wil dat artikelen van Nederlandse wetenschappers open access gepubliceerd worden. Blijf op de hoogte van deze en andere ontwikkelingen via de Open-Access nieuwsbrief van de VSNU, de Nederlandse universiteitsbibliotheken en de Koninklijke Bibliotheek. (VSNU)
vsnu.nl

OPROEP

Wint u de Nederlandse Dataprijs 2016?

Komend najaar wordt weer de Nederlandse Dataprijs uitgereikt. Een prijs voor een onderzoeker of onderzoeksgroep die extra bijdraagt aan de wetenschap door onderzoeksdata beschikbaar te maken voor aanvullend of nieuw onderzoek.
De winnaars van de voorgaande edities zijn in ieder geval enthousiast: "De jury noemt onze database een grote aanwinst voor zowel het Nederlands academisch als cultureel erfgoed. Dat is een bevestiging dat we op het goede spoor zitten," aldus Martine de Bruin, Nederlandse Liederenbank, winnaar van de Dataprijs humaniora en sociale wetenschappen 2014.
"Door het winnen van de Dataprijs kunnen we nu ook een paar grotere, al langer gewenste verbeterslagen maken," aldus Johan Molenbroek en Marijke Dekker, DINED, winnaars van de Dataprijs exacte en technische wetenschappen 2014.
Naast de winnaars waren ook de bijna 50 andere inzendingen van hoog niveau. De jury sprak over 'allemaal mooie voorbeelden van het toegankelijk maken en delen van onderzoeksdata'.
De organisatie van de Nederlandse Dataprijs is in handen van Research Data Netherlands, een samenwerkingsverband tussen 3TU.Datacentrum, DANS en SURFsara. Binnenkort staat meer informatie over de Dataprijzen 2016 op de website van RDNL. (HB)
researchdata.nl

GEHOORD & BIJGEWOOND

10-jarig bestaan voor Knowledge Exchange

Ingrid Dillo

Op 30 november en 1 december vond in Helsinki een conferentie plaats ter gelegenheid van het 10-jarig bestaan van de Knowledge Exchange (KE). Onlangs is een nieuw contract voor 3 jaar tussen de partners gesloten. De nieuwe visie op *open scholarship* wordt tijdens de conferentie verder ingevuld. De bedoeling van de conferentie is dit begrip verder in te vullen en de uitdagingen te definiëren.

De dag wordt geopend door Kimmo Koski van CSC en Bas Cordewener die vanuit JISC voor KE werkt. Vervolgens houdt Sascha Friesike van het Alexander von Humboldt Institute for Internet and Society een provocerend en leuk verhaal over open scholarship. Wat verstaan we eigenlijk onder de term? Wat zijn de voordelen? Zijn conclusie is dat we teveel onder elkaar praten over de definities. We zouden veel meer naar de onderzoekers moeten gaan om met hen in discussie te gaan over de voordelen van *open scholarship* voor de onderzoekers zelf. Vervolgens start de plenaire sessie met vier korte intro's van experts. Interessant punt dat aan de orde komt: wat als je artikel wordt geweigerd vanwege slechte kwaliteit van de data, als dat data zijn die je niet zelf hebt verzameld maar die je hebt hergebruikt? Vervolgens is er een uur met acht *5-minutes-madness* presentaties. DANS presenteert de nieuwe *common requirements for basic certification*, gebaseerd op DSA en WDS. De interesse is groot. Daarna valt de groep in vier break-out sessies uiteen, 's avonds wordt doorgepraat tijdens een diner. De tweede dag komen in de afsluitende discussie vooral elementen naar voren die samenhangen met de druk om te publiceren en het ontbreken van directe *rewards* voor onderzoekers die *open scholarship* bedrijven. Data komen niet uitgebreid aan de orde. Wel worden er wensen uitgesproken voor *executable DMPs* en *trustworthy repositories* die tot in de eeuwigheid zorgen voor data.

http://www.knowledge-exchange.info/news/articles/21-12-2015

Tijdens het congres 'Getuige van Waarde' deelden wetenschappers en maatschappelijk partners ervaringen
foto Hans Tak

Boeiend congres getuigt van waarde wetenschap

Arjan Hogenaar

Op 2 december vond het congres 'Maatschappelijke Impact van Alfa- en Gammawetenschappen: Getuigen van Waarde' plaats, georganiseerd door de Universiteit Utrecht en ScienceWorks. Het was een boeiend congres. Vele aspecten van het alfa/gamma-onderzoek kwamen aan bod. Zo verklaarde José van Dijck (president KNAW) de cruciale rol van de alfa/gammawetenschappen, liefst in samenwerking met de beta-wetenschappen, bij het oplossen van wereldvraagstukken. Karl Dittrich (bestuursvoorzitter VSNU) gaf aan onderwijs als het belangrijkste valorisatie-aspect te beschouwen. Victor van der Chijs (Universiteit Twente) wees op het belang van maatschappelijke implicaties van technologische ontwikkelingen. Bernard ter Haar (Ministerie SZW) sprak de wens uit om, samen met de wetenschap, beleidsimplicaties van onderzoeksresultaten beter inzichtelijk te maken.

Het middagprogramma bevatte vele parallelsessies. Zo werd in een sessie over 'geletterdheid' door Frits van Oostrom (Universiteit Utrecht) het belang van ICT in tekstonderzoek toegelicht en beschreef Inge Molenaar (Radboud Universiteit) de samenwerking tussen leraren en universiteit bij de introductie van ICT in het onderwijs. Liesbeth van Zoonen (Erasmus Universiteit Rotterdam) liet tenslotte weten het project Urban Big Data als voorbeeld te zien van het delen en hergebruiken van data afkomstig van diverse disciplines en social media.

scienceworks.nl

Evaluatie toegankelijkheid microdatabestanden CBS

Marion Wittenberg

Het Centraal Bureau voor de Statistiek (CBS) en DANS werken al jaren samen aan het beschikbaar stellen van beveiligde microdatabestanden voor wetenschappelijk onderzoek. Omdat de statistieken van het CBS de laatste jaren dynamischer zijn geworden, onderzoeken zijn gewijzigd of stopgezet en andere zijn gestart, organiseerden het CBS en DANS op 24 november een gebruikersbijeenkomst om deze vorm van dienstverlening te evalueren.

Uit deze discussie bleek dat onderzoekers de beveiligde microbestanden positief evalueren. Een belangrijke reden hiervoor is dat de bestanden gratis beschikbaar en met de eigen computer te gebruiken zijn en dat men niet gehinderd wordt door verplichte outputcontrole. Dit in tegenstelling tot de Remote Access faciliteit, een andere vorm van dienstverlening van het CBS. Men vindt beide vormen van dienstverlening goed op elkaar aansluiten.

Kritiekpunten waren er echter ook. Onderzoekers zouden graag willen kunnen zoeken naar variabelen, wat momenteel niet mogelijk is. De procedure om toegang tot de data te krijgen, waarbij men elke keer een geheimhoudingsverklaring moet tekenen, vindt men nogal omslachtig en men vraagt zich af waarom de beveiligde microbestanden niet gebruikt mogen worden binnen het onderwijs. Met name researchmasterstudenten zouden toegang tot deze data moeten kunnen krijgen. Verder vindt men dat de criteria verruimd zouden moeten worden welke instellingen deze bestanden mogen gebruiken; ook niet universitaire onderzoeksinstituten zouden toegang tot de data moeten kunnen krijgen. Een ander kritiekpunt is het ontbreken van essentiële variabelen in bestanden, men zou graag inspraak willen hebben in de keuze van de variabelen.

DANS en het CBS gaan bekijken hoe ze de kritiekpunten kunnen oppakken en deze vorm van dienstverlening in de komende periode kunnen verbeteren.

http://dx.doi.org/10.17026%2Fdans-z5j-9bkf

Kennis- en onderwijsinstellingen, overheid en bedrijven bundelen krachten

Nieuwe big data-experts door GRIDS

TiU, TU/e, de provincie Noord-Brabant en de gemeente Den Bosch bundelen hun krachten: een Grand Initiative Data Science (GRIDS) is in de maak. Het Data Science Center Tilburg is één van de onderdelen. ***Patricia Prüfer***

Het Data Science Center Tilburg (DSCT) brengt bestaande wetenschappelijke expertise op het gebied van big data bij elkaar om onderzoek, onderwijs en toepassingen op het gebied van data science in de toekomst multidisciplinair te benaderen. DSCT is dan ook een gezamenlijk initiatief van vier faculteiten: Economics & Management, Law, Humanities en Social & Behavioral Sciences.

Unieke aanpak

De holistische visie op data science die DSCT hanteert, gaat uit van vier (kennis-)domeinen: methodisch/technisch, sociaal, juridisch en toegepast/innovatief. Onderzoek en onderwijs op het gebied van data science moeten met deze vier domeinen rekening houden. Daarnaast concentreert DSCT zich op een aantal gebieden: Human Capital & Labor Market, Smart Cities & Industries, Health, Consumer Behavior, Financial Institutions en Legal Analytics.

Focus op onderwijs

De focus van DSCT ligt in eerste instantie op onderwijs. "Want", aldus professor Arjan van den Born, Academic Director van DSCT: "naast weten hoe je met data moet omgaan, moet je ook weten hoe je goede vragen stelt en waarde kunt toevoegen met je onderzoek of project. Je hebt eigenlijk een T-shaped data scientist nodig die diepgaande kennis op één gebied heeft en daarnaast redelijke kennis op minimaal één ander gebied." Er komen vier onderwijsprogramma's: Data Science Master Business & Society (op de TiU), Data Science Master Engineering (op de TU/e), Data Science Master Entrepreneurship (op de Graduate School Mariënburg) en Bachelor Data Science (op alle locaties). De eerste masteropleiding Data Science & Governance is inmiddels op de TiU gestart, voor de andere opleidingen wordt de accreditatie aangevraagd. De nadruk binnen deze opleidingen ligt op het leren omgaan met data én het stellen van de goede vragen. Dit gebeurt in nauwe samenwerking met bedrijven die hun data beschikbaar stellen voor de analyses door de toekomstige Brabantse data scientists.

De rol van data wordt steeds groter en het omgaan met en gebruiken van data steeds belangrijker. Met het opzetten van een GRIDS dragen Brabantse kennis- en onderwijsinstellingen in samenwerking met de overheid en bedrijven bij aan de ontwikkeling van een data-gedreven maatschappij. Uiteindelijk zullen duizenden Brabantse data scientists worden opgeleid.

https://www.tilburguniversity.edu/research/institutes-and-research-groups/data-science-center/

COLOFON **Uitgever**: Stichting Uitgeverij E-data & Research Den Haag. **Redactieadres**: Postbus 93067, 2509 AB Den Haag, 070-3494450, edata@dans.knaw.nl, www.edata.nl. **Hoofd-/eindredacteur**: Heidi Berkhout. **Redactie**: Marika de Bruijne, Steven Claeyssens, Ilja Nieuwland, Rutger Nugteren, Erica Renckens, Marion Wittenberg. **Redactiesecretariaat**: Lucas Pasteuning. **Aan dit nummer werkten mee**: Ingrid Dillo, Arjan El Fassed, Guido de Groot (Kito), Wieke Hoeke, Arjan Hogenaar, Hugo Huurdeman, Edwin Klijn, Patricia Prüfer, Jos Rikers, Ewoud Sanders en Heiko Tjalsma.
Opmaak: Colette Sloots, Haarlem. **Productie**: Amsterdam University Press. **Druk**: Ten Brink, Meppel. **Webmaster**: Sonja Duijkers **Oplage**: 7500 papier, 1900 digitaal. **ISSN**: 1872-0374. We hebben getracht alle belanghebbenden met betrekking tot het gebruikte beeldmateriaal te benaderen.
Degenen die menen rechten te kunnen doen gelden, kunnen zich tot ons wenden. Toezending papieren en/of digitale versie is kosteloos aan relaties van de stakeholders en studenten in de alfa- en gammarichtingen.

Overname artikelen

Wilt u een artikel uit dit blad overnemen? Dat mag altijd, maar vermeld wel de bron (E-data & Research) en de naam van de auteur van het artikel. Neem ook contact op met de hoofdredacteur (zie colofon) om door te geven waar artikelen geplaatst worden.

JONG TALENT

Van Gompel winnaar CLARIN Young Scientist Award 2015

‘Ik ontwikkel alleen open source’

Maarten van Gompel ontving dit najaar de CLARIN Young Scientist Award. *Erica Renckens*

Van Gompel: “De toepassingen die ik ontwikkel, zijn altijd open source. Alleen zo kan onderzoek repliceerbaar zijn” foto Wieke Hoeke

De prijs wordt jaarlijks toegekend aan een veelbelovende jonge onderzoeker die bijdraagt aan het bouwen van taalbronnen, het ontwikkelen van tools en het delen van kennis. Hans Bennis (Meertens Instituut) en Walter Daelemans (Universiteit van Antwerpen) droegen Van Gompel voor. In hun aanbevelingsbrief noemen ze hem ‘een begaafd programmeur en onderzoeker’ die tools ontwikkelt ‘die aan de basis liggen van belangrijke ontwikkelingen’.

Die tools zijn met name FoLiA en CLAM. Van Gompel werkt daar al aan sinds hij na zijn master Human Aspects of Information Technology in Tilburg betrokken raakte bij onder andere het SoNaR-project. Het SoNaR-corpus bevat meer dan vijfhonderd miljoen woorden aan Nederlandse teksten.

‘Open source software voor onderzoekers’

Taalkundige annotaties

“FoLiA is een bestandsformaat voor geannoteerde corpora,” vertelt Van Gompel in zijn werkkamer bij de Radboud Universiteit. “Bij de ontwikkeling van SoNaR ontstond de noodzaak voor een goed formaat waarin taalkundige annotaties vastgelegd konden worden. Inmiddels wordt FoLiA ook door andere corpora gebruikt, zoals BasiLex en Nederlab.”

Toch is de tool nog altijd niet ‘af’. Van Gompel: “Ik ontwikkel nog steeds software voor FoLiA, zoals FLAT, waarmee een gebruiker heel makkelijk annotaties in het FoLiA-formaat kan toevoegen. En Frog voert taalkundige analyses uit met FoLiA als outputformaat.” Van Gompel typt razendsnel complexe commando’s terwijl hij zijn werk laat zien. Kan de gemiddelde onderzoeker wel met zijn programma’s uit de voeten? “Daarvoor heb ik CLAM ontwikkeld,” legt hij uit. “Dat vervangt de command-line en creëert automatisch een overzichtelijke gebruikersinterface. Die interface is een webservice, waardoor ook machines hem kunnen aanspreken. Een gebruiker kan zo ook een zelfontworpen interface gebruiken die weer via CLAM met de software communiceert.”

Alle codes vrij

Voorlopig is de jonge onderzoeker nog druk bezig met het afronden van zijn proefschrift, waarin hij de kwaliteit van automatische vertalingen probeert te verbeteren door te kijken naar de context. Maar daarna wil hij het liefst weer software ontwikkelen die onderzoekers nodig hebben. Van Gompel: “Ik probeer altijd een zo generiek mogelijke oplossing te vinden voor problemen. Zo kun je voorkomen dat mensen dubbel werk doen.” De toepassingen die Van Gompel ontwikkelt, zijn daarnaast altijd open source. “Daar ben ik heel principieel in. De code is altijd vrij te gebruiken door andere ontwikkelaars. Dat is de enige manier waarop onderzoek repliceerbaar kan zijn.”

De CLARIN Young Scientist Award bestaat uit een geldprijs van 500 euro en een certificaat. Van Gompel: “Ik werk veel vanuit huis, dus dat certificaat heeft daar een mooi plekje aan de muur gekregen.”

clarin.eu

Safe harbour-principe en Right to be forgotten uitgelegd

Issues privacy en open data ’16

Hoe verhouden privacy en open data zich tot elkaar? Welke issues spelen op dit moment? *Heiko Tjalsma*

Al jaren heerst een steeds sterkere drang om alle (onderzoeks)data zo onbeperkt mogelijk ter beschikking te stellen. Aan de andere kant wordt, óók onder invloed van de Europese Unie, de privacybescherming steeds strenger. Problematisch is dit wanneer personen het onderzoeksonderwerp zijn, zoals in de sociale wetenschappen, de medische wetenschappen en de humaniora. Dit spanningsveld tussen open access en privacybescherming werd ook tijdens de Amsterdam Privacy Conferentie (oktober 2015) geconstateerd. Twee issues, van belang voor wetenschappelijk onderzoek, worden in dit artikel besproken.

Persoonsdata over zee

Het *safe harbour*-principe gaat over het doorgeven van persoonsdata naar de Verenigde Staten. Recent is dit onrechtmatig verklaard door het Europese Hof van Justitie. Dit maakt de uitwisseling van persoonsdata van en naar de VS onmogelijk, ook voor onderzoekers. Er wordt, met hoge prioriteit, aan een nieuwe regeling gewerkt.

Recht om te worden vergeten

Het *right to be forgotten* gaat om de mogelijkheid voor individuen om hun geschiedenis, speciaal op internet, te wissen. Dit principe staat centraal in de door de EU voorgestelde nieuwe privacy-wet, de General Data Protection Regulation (GDPR). Deze wet gaat in de plaats komen van de huidige nationale wetten, zoals de Nederlandse Wet Bescherming Persoonsgegevens (WBP). De komst van de GDPR, en vooral het bijna absoluut genomen *right to be forgotten*, heeft tot grote ongerustheid bij onderzoekers geleid. Gevreesd werd voor een enorme verslechtering bij het gebruik van persoonsdata in wetenschappelijk onderzoek. Medio december 2015 is er eindelijk overeenstemming bereikt tussen de Europese bestuursorganen, naar verwachting wordt de wet in het voorjaar van 2016 aangenomen. Een voorlopige eerste conclusie is dat de wet er uiteindelijk minder alarmerend uitziet dan eerder leek. Er blijven uitzonderingen mogelijk voor onderzoek- en archiefdoeleinden. Voor actuele informatie zie http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

Informed consent essentieel

Voorlopig blijft de huidige praktijk van kracht, gebaseerd op de WBP. Essentieel is en blijft informed consent: het geven van toestemming door geïnformeerde proefpersonen of patiënten. Juist hier komen nadere aanvullende nationale regelingen. Hoe de nieuwe EU-wet precies zal uitwerken, zal pas over enige jaren duidelijk worden.

Meer weten over de GDPR en de gevolgen daarvan voor onderwijs en onderzoek? SURF heeft de afgelopen jaren een aantal praktische handleidingen gepubliceerd. Meer informatie staat op de website van SURF.

Heiko Tjalsma is juridisch adviseur bij DANS.
https://www.surf.nl/themas/beveiliging/beleidsondersteuning-privacy/index.html

KORT

LingOA biedt gratis online publicaties

De redacties van vijf taalwetenschappelijke tijdschriften, waaronder Lingua, hebben hun traditionele uitgever de rug toegekeerd. Voortaan publiceren zij hun artikelen via Ubiquity Press onder de voorwaarden van de zelf opgerichte stichting LingOA. Dit houdt in dat wetenschappers betalen voor hun publicaties, die vervolgens online vrij toegankelijk worden. Dankzij een garantie van VSNU, NWO en KNAW hoeven auteurs de eerste vijf jaar de publicatiekosten niet zelf te betalen.
De verwachting is dat na deze periode het publicatiemodel wereldwijd zal zijn veranderd van abonnementen naar open access. De komende tijd zullen naar verwachting meer linguïstische tijdschriften zich aansluiten bij LingOA. (ER)
lingoa.eu

Nieuwe cursusronde RDNL start in mei

Essentials 4 Data Support is een introductiecursus voor diegenen die onderzoekers (willen) ondersteunen bij het opslaan, beheren, archiveren en delen van hun onderzoeksdata. In mei gaat een nieuwe ronde van de cursus van start. De cursus

rdnl

bestaat uit twee groepsbijeenkomsten met begeleiding door coaches, presentaties van experts, online cursusmateriaal en opdrachten. Research Data Netherlands (RDNL) wil met deze cursus een bijdrage leveren aan de professionalisering van en de afstemming tussen datasupporters. Kijk voor inschrijving of de online-only variant van de cursus op de cursussite van RDNL. (HB)
datasupport.researchdata.nl

Subsidie Summer Program ICPSR

DANS biedt ook dit jaar een tegemoetkoming in de kosten voor deelname aan het Summer Program in Quantitative Methods of Social Research 2016 van het Inter-university Consortium for Political and Social Research (ICPSR) in de Verenigde Staten. Er is dit jaar één subsidie beschikbaar van € 2.000. Alleen researchmasterstudenten en PhD’s van instellingen die participeren in het Nationale Lidmaatschap van het ICPSR kunnen een aanvraag indienen. De uiterste datum voor het indienen van de aanvraag is 1 april.
Meer informatie over de aanvraagprocedure en het lidmaatschap van ICPSR staat op de site van DANS. (MW)
dans.knaw.nl

‘Trots op 50 DSA-seals wereldwijd’

Mary Vardigan (ICPSR) neemt na drie jaar afscheid van het internationale datakeurmerk DSA. Wat zijn de belangrijkste successen onder haar voorzitterschap geweest?
Ingrid Dillo

“Allereerst is daar natuurlijk de enorme groei die DSA in de afgelopen jaren heeft doorgemaakt. Toen ik voorzitter werd, waren er zo’n twintig repositories met een DSA-seal. Nu zijn dat er al meer dan vijftig, verspreid over de gehele wereld.”
Mary praat verder: “Wat ik ook een groot winstpunt vind, is de inrichting van een general assembly (GA) eind vorig jaar. De GA maakt DSA meer community gedreven en duurzamer en levert ons tegelijkertijd een grotere pool van reviewers op.”

Research Data Alliance

Als laatste belangrijke wapenfeit noemt Mary de samenwerking met het World Data System (WDS). “Om stakeholders nog beter te kunnen bedienen, heeft DSA het afgelopen jaar samenwerking gezocht met het WDS van het International Council for Science (ICSU). WDS biedt zijn datacentra een accreditatieprocedure die erg lijkt op de basiscertificering van DSA. Onder de vlag van de Research Data Alliance (RDA) hebben beide partijen gezamenlijk een catalogus van requirements ontwikkeld, die het beste van beide standaarden combineert. Deze catalogus zal dit jaar gefaseerd worden ingevoerd door beide organisaties.”

Nederlandse praktijk

Vanuit Nederland is DANS nauw betrokken bij de ontwikkelingen op het terrein van certificering. Het digitale archief van DANS is DSA- en WDS-gecertificeerd en heeft onlangs ook het nestorSeal verkregen. Daarnaast participeert DANS in de besturen van DSA en WDS en trekt de organisatie het Nederlandse certificeringsproject van de NCDD. Meer informatie over de certificering van digitale repositories in Nederland staat in de flyer Doe ik het goed? van de NCDD.
datasealofapproval.org

Internationaal raamwerk

Informatiebeheerders en andere betrokkenen kunnen terugvallen op een raamwerk van verschillende internationale certificeringstandaarden voor digitale repositories om de kwaliteit van hun werkprocessen en beheersystemen te toetsen en te verbeteren. Een ‘trustworthy digital repository’ (tdr) is een term die dan vaak wordt gebruikt. In toenemende mate van complexiteit en diepgang zijn de volgende drie instrumenten beschikbaar: het Data Seal of Approval (DSA), het nestorSeal (toetsing op DIN-standaard 31644) en de ISO-certificering (16363). De toetsing loopt in intensiteit uiteen van een ‘peer review’ van opgeleverde documentatie in het geval van DSA, tot een voorbereid ‘on-site’ bezoek van een extern audit team in het geval van ISO. Financiers, producenten en hergebruikers van data kunnen vertrouwen op een beherende instelling met een certificering volgens een van de omschreven standaarden.

Mary Vardigan neemt na drie jaar afscheid van het internationale datakeurmerk DSA *foto Umich*

SINDS KORT BESCHIKBAAR

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

CentERdata

• Data economische situatie van Nederlanders online beschikbaar

Al sinds acht jaar wordt de economische situatie van Nederlanders in beeld gebracht door Economic Situation. Deze studie maakt deel uit van de kernstudie van het LISS panel die de ontwikkeling van veel levensaspecten volgt. De studie bestaat uit drie modulen: Income, Housing en Assets. De eerste twee modulen worden elk jaar bevraagd, het laatste tweejaarlijks. De data bieden een breed inzicht in het Nederlands leven. Sinds kort zijn de data voor 2014 en 2015 beschikbaar en kunnen worden gedownload via LISS Data Archive.
lissdata.nl/dataarchive

Ook sinds kort beschikbaar:

Studies LISS panel

- Wetenschappelijk Bureau 50-Plus, februari 2015, Changing costs regarding care and pension
- Geijtenbeek, L.; Buser, T., maart 2014, Competition & sexual preference
- Leeuw, E.D. de; Conrad, F.G., juli 2014, Professional respondents in panels
- Vroege, L. de; Feltz-Cornelis, C.M. van der, april - mei 2014, Prevalence and relevant associations of alexithymia in a Dutch general population sample and in comparison to clinical patients with somatic symptom disorder (SSD)
- CentERdata, november - december 2014, Personality - Wave 7
- CentERdata, januari t/m december 2014, Initial Questionnaire - 2014
- CentERdata, april - mei 2015, Work and Schooling - Wave 8
- CentERdata, juli - augustus 2015, Health - Wave 8
- CentERdata, augustus - september 2015, Religion and Ethnicity - Wave 8

Studies Immigrant panel

- CentERdata, januari t/m december 2014, Initial Questionnaire - 2014
- Leeuw, E.D. de; Conrad, F.G., juli 2014, Professional respondents in panels

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

DANS

• Nieuw in EASY: dataset Hebrew Text Database ETCBC4b

Onlangs is de dataset Hebrew Text Database ETCBC4b gedeponeerd. Het gaat hier om de Hebreeuwse Bijbel, in de text van de Biblia Hebraica Stuttgartensia, taalkundig geannoteerd door het Eep Talstra Centre for Bible and Computer (ETCBC, VU Amsterdam) en gecureerd in het SHEBANQ-project. Dat de data al worden gebruikt, werd duidelijk op de Annual Meeting van de Society for Biblical Literature te Atlanta. Joshua Berman en Moshe Koppel van de Bar-Ilan universiteit Israel presenteerden een bètaversie van Tiberias, een systeem waarmee gebruikers relatief gemakkelijk datamining op het corpus van Bijbelse teksten kunnen uitvoeren.
http://dx.doi.org/10.17026/dans-z6y-skyh

Joshua Berman (rechts op de foto) en Moshe Koppel van de Bar-Ilan universiteit Israel presenteerden een bètaversie van Tiberias *foto Dirk Roorda*

Ook sinds kort beschikbaar:

- Centraal Bureau voor de Statistiek (2015): Enquête Beroepsbevolking - EBB - 2014. DANS. http://dx.doi.org/10.17026/dans-xqb-a38p
- Heijmans, Drs N. (Radboudumc Nijmegen) (2015): Social network composition of vascular patients and its associates with health behavior and clinical risk factors. DANS. http://dx.doi.org/10.17026/dans-zz6-fd4y
- Hense, Dr. E.H. (Radboud Universiteit Nijmegen) (2013): Thematische collectie - project Spiritualiteit en Maatschappelijke Vernieuwing. DANS. http://dx.doi.org/10.17026/dans-zuf-ck76
- Hoogendoorn-Lanser, Dr S. (KiM Netherlands Institute for Transport Policy Analysis) (2015): Mobiliteitspanel Nederland (MPN 2013). DANS. http://dx.doi.org/10.17026/dans-zyc-7qfv
- Kožuh, Dr. I.K. (University of Maribor) (2015): Community Building among Deaf and Hard of Hearing People on Social Networking Sites. DANS. http://dx.doi.org/10.17026/dans-xfw-qztc
- De Regt, dr. S. (Utrecht University) (2015): Onderzoek Nationale Dodenherdenking. DANS. http://dx.doi.org/10.17026/dans-x8u-fkzx
- Rezetko, dr. R.C. (Radboud University Nijmegen/ University of Sydney); Naaijer, drs. ir. M. (Vrije Universiteit Amsterdam) (2015): An Alternative Approach to the Lexicon of Late Biblical Hebrew-Dataset. DANS. http://dx.doi.org/10.17026/dans-256-4hcy-http://dx.doi.org/10.17026/dans-xn8-v6dy

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

Huygens ING

• Online versie Noord en Oost Tartarye van Nicolaas Witsen (1705)

Noord en Oost Tartarye is meer dan driehonderd jaar geleden geschreven door de Amsterdamse burgemeester en amateurgeleerde Nicolaas Witsen (1641-1717). Nooit eerder bracht iemand zoveel kennis bijeen over ‘Tartaria’ of ‘Tartarije’, het tegenwoordige Eurazië. Deze digitale uitgave bevat naast een inleiding in het Russisch en een toelichting in het Nederlands ook een uitvoerig register van zaken, persoonsnamen, geografische begrippen en etnografische namen. Ook een lijst van Witsens bronnen en van de gebruikte secundaire literatuur ontbreekt niet, evenals een toelichting op de illustraties. Hiermee is Noord en Oost Tartarye zo goed mogelijk toegankelijk gemaakt.
http://resources.huygens.knaw.nl/witsen

De digitale editie bevat ook deze afbeelding van Witsens ‘grote kaart’ van 1687 *bron Huygens ING*

Ook sinds kort beschikbaar:

- Documenten Molukse Kerk en School Ambon, Ternate en Banda: http://resources.huygens.knaw.nl/retroboeken/molukse_kerk/#view=homePane&page=0&accessor=toc
- Instrumenten van de macht. De archieven van de Staten-Generaal 1576-1796: http://resources.huygens.knaw.nl/retroboeken/instrumenten_macht/#page=0&accessor=toc&view=homePane
- Aanvullingen Willem de Clercq: http://resources.huygens.knaw.nl/retroboeken/declercq/#page=0&accessor=toc&view=homePane
- Ystroom: http://deystroom.huygens.knaw.nl/
- Clusius Correspondence: a digital edition-in-progress: http://clusiuscorrespondence.huygens.knaw.nl

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

KNAW-president José van Dijck:

‘Digital Humanities verfrissen onze blik op bestaande data’

De KNAW presenteerde onlangs CHAT, het Center for Humanities and Technology. E-data interviewt president José van Dijck over de kansen voor de geesteswetenschappen, nieuwe én oude stijl.

Erica Renckens

“Als ik mijn promotieonderzoek nu opnieuw zou mogen uitvoeren, dan zou er zó veel meer mogelijk zijn,” verzucht José van Dijck in haar onlangs gerenoveerde kantoor in het Amsterdamse Trippenhuis. “Ik onderzocht het publieke debat rondom in-vitrofertilisatie (ivf). De eerste reageerbuisbaby werd in 1978 geboren en zeven jaar later zat ivf in het ziekenfondspakket. In het begin waren mensen fel tegen, maar na een tijdje werd het toch geaccepteerd. Ik vroeg me af hoe zo’n proces verloopt en wat de rol van de media daarin is.”

Om dit te kunnen onderzoeken, moest Van Dijck de archieven in. “Hele krantenarchieven heb ik doorgeploegd op dit debat, dat was bijna niet te doen. Ik moest het onderwerp noodgedwongen heel klein houden, want als ik dat niet deed, werd het een onmenselijke klus,” vertelt Van Dijck. “Nu zou ik veel meer data over een veel langere periode kunnen doorzoeken. Ik zou toegang hebben tot gedigitaliseerde kranten, radio- en televisieopnames en gestructureerde data, en zo veel breder naar het debat kunnen kijken.”

Zoeken in de ondertiteling

Tijd om haar onderzoek met de huidige middelen nogmaals uit te voeren, heeft Van Dijck echter niet. Naast hoogleraar Media en Cultuur aan de Universiteit van Amsterdam en president van de KNAW is zij ook een van de aanvragers van CLARIAH, een consortium dat zich richt op de ontwikkeling van een digitale infrastructuur voor de geesteswetenschappen. “De drie focusgebieden van CLARIAH ontwikkelen tools voor hun eigen data: taalkunde voor tekstbestanden, mediastudies voor audiovisuele bronnen en sociaaleconomische geschiedenis voor gestructureerde data,” vertelt Van Dijck. “Maar het leuke is dat die tools vervolgens ook bruikbaar zijn in andere geesteswetenschappelijke disciplines. Zo kunnen we straks taalkundige tools gebruiken voor semantisch zoeken in automatisch gegenereerde ondertiteling bij audiovisuele bestanden.”

Netwerken met CHAT

Van Dijck: “De KNAW ziet aan CLARIAH dat de digital humanities echt een opkomend gebied zijn, van belang voor alle geesteswetenschappen. Daarom is begin december het Center for Humanities en Technology (CHAT) gelanceerd. Hierin zitten zeven KNAW-instituten en acht faculteiten Geesteswetenschappen, van alle grote Nederlandse universiteiten.”

Met CHAT wil de KNAW een landelijk netwerk vormen. “We brengen met CHAT mensen uit uiteenlopende vakgebieden bij elkaar. Zij hebben hele verschillende onderzoeksvragen, maar zijn tegelijkertijd wel heel geïnteresseerd in elkaars methodologieën. Je kunt elkaars data en tools vaak goed gebruiken in een heel ander vakgebied.”

INTERVIEW

‘CHAT brengt mensen, data en tools bij elkaar’

Publiek belang bij open data

Voorwaarde voor het gebruik van elkaars tools is dat de data waarop ze losgelaten kunnen worden ook vrij toegankelijk zijn. “Dat is inderdaad nog een groot obstakel,” geeft Van Dijck toe. “Veel historische data zijn nog altijd niet gedigitaliseerd. Ik geloof dat slechts 5 tot 7 procent van de archieven van KNAW-instituten gedigitaliseerd zijn. Bij de KB en het Instituut voor Beeld en Geluid zal dat iets meer zijn. Het toegankelijk maken van al die data zal nog heel wat geld en manuren kosten.”

“En dan spelen er ook nog copyright-problemen. Dat is een belangrijk onderwerp dat ook op de agenda moet staan. Van wie zijn die data nu eigenlijk en hoe mogen ze naar buiten gebracht worden? Dat is hele grote problematiek, maar daar schrikken we niet voor terug. Ik kan niet voorspellen waar we uit zullen komen, maar zeker is dat het alleen maar verder gaat en we kunnen die ontwikkeling wel stuwen,” aldus Van Dijck.

José van Dijck

José van Dijck (1960) doet onderzoek naar sociale media, mediatechnologieën en digitale cultuur. Ze studeerde aan de Universiteit Utrecht en promoveerde aan de Universiteit van Californië in San Diego. Ze was universitair docent journalistiek aan de Rijksuniversiteit Groningen en hoofddocent media en visuele cultuur aan de Universiteit Maastricht.
In 2001 werd Van Dijck benoemd tot hoogleraar bij het departement Mediastudies aan de Universiteit van Amsterdam, waarvan ze van 2002 tot 2007 voorzitter was.
Van 2008 tot 2011 was ze decaan van de Faculteit Geesteswetenschappen van de Universiteit van Amsterdam.
Op 18 mei 2015 volgde ze Hans Clevers op als president van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW).

“We hebben een publiek belang bij open data” *foto Milette Raats*

Publiek belang bij open data

“Ik begrijp het sentiment van de geesteswetenschapper die zijn data niet wil delen heel goed,” zegt Van Dijck. “Als je vroeger als historicus naar een archief ging, was wat je daarin vond jóuw schat en daar wilde je eerst zelf induiken. Maar in het digitale tijdperk is dat toch wat moeilijker vol te houden, want steeds meer data zijn open beschikbaar. Tegelijk zien we ook een trend naar juist meer gesloten data. Vier jaar geleden kon je als onderzoeker nog zo Twitter-data gebruiken, nu vragen ze er geld voor. Data zijn veel geld waard, daarom moeten we als onderzoekers bijblijven en het niet aan bedrijven overlaten om bijvoorbeeld zoekalgoritmes te ontwikkelen. We hebben een publiek belang bij open data.”

Financiers gezocht

“CLARIAH is erg blij met de 12 miljoen euro die ze van NWO heeft ontvangen om de komende vijf jaar tools mee te ontwikkelen, maar er is nog onvoldoende geld om ook onderzoek mee te doen,” vertelt Van Dijck. “Kijk naar de fysici, die hebben CERN, met die infrastructuur doen ze hun onderzoek. Of de astronomen, als zij voor 50 miljoen een mooie telescoop hebben, moeten ze nog beginnen met onderzoek doen. Bij ons is dat precies hetzelfde.”

“Onze grote wens is dus dat er meer financiering beschikbaar komt. Daarvoor kijken we smekend naar NWO en andere financiers. Als individuele vakgebied binnen de geesteswetenschappen zijn we maar klein, dus daarom vormen we met CHAT een coalitie om grote aanvragen te doen. Bovendien zie je dat landen die zich sterk organiseren in de digital humanities veel sterker staan bij calls.”

Positieve impuls

Gaat de financiering van de nieuwe, digitale geesteswetenschappen ten koste van de geesteswetenschappen ‘oude stijl’? “Nee, absoluut niet, het is geen kwestie van of-of,” reageert Van Dijck. “Het zal juist een impuls geven aan sommige gebieden waar we nu dingen kunnen zien die we voorheen niet konden zien. Daarnaast zien we dat onze cultuur in hoog tempo digitaliseert en je moet je instrumenten afstemmen op de data waar je mee werkt. Maar digital humanities is geen vervangende methodologie, je hebt nog steeds interpretatieve methoden nodig.”

knaw.nl

ONDERZOEK

Welk cijfer geeft u E-data & Research?

In maart vindt een onderzoek onder lezers van dit blad plaats, uitgevoerd door CentERdata. Vragen over de inhoud, de leesbaarheid en bijvoorbeeld de verspreiding van E-data & Research laten we aan bod komen.
We stellen het zeer op prijs als u mee wilt werken aan dit onderzoek.

U ontvangt hierover binnenkort een e-mail. Het onderzoek zal een paar minuten in beslag nemen, uw antwoorden zijn voor ons erg waardevol. We zijn benieuwd naar de resultaten en berichten hier natuurlijk graag over in een volgend nummer. Alvast bedankt voor uw medewerking! (HB)

Met een online informatie-infrastructuur brengt het Landelijk Coördinatiepunt Research Data Management wel structuur aan voor onderzoekers.

illustratie Kito/Kitocartoons.com

Landelijk Coördinatiepunt Research Data Management

LCRDM gaat voor samenhang

Datamanagement leeft. Er worden veel projecten opgezet door landelijke werkgroepen en binnen universiteiten. Maar eenheid in het beleid, laat staan de praktijk, ontbreekt. Het LCRDM moet samenhang brengen. ***Marika de Bruijne***

LCRDM staat voor Landelijk Coördinatiepunt Research Data Management. Om de vele initiatieven op dit gebied bij elkaar te brengen, vroeg de VSNU aan SURFsara om een landelijk coördinatiepunt op te richten voor wetenschappelijk onderzoek in Nederland. Eén van de eerste activiteiten van het LCRDM was een bijeenkomst van experts, met onder andere leden van de SIG Research Data en de UKB Werkgroep Research Data. “Er moet niet nóg een vergaderclubje bij,” riep één van de aanwezigen. En daarmee schetste hij in één keer de ontstaansreden van het coördinatiepunt. Het moet bestaande activiteiten en resultaten op het gebied van Research Data Management (RDM) zichtbaar maken en een gemeenschappelijke aanpak van belangrijke vraagstukken bevorderen. De rol van het LCRDM is faciliterend. Ingeborg Verheul, aanspreekpunt van het LCRDM bij SURFsara, vertelt: “De input komt van deskundigen van deelnemende instellingen.”

Online informatiebronnen

Verheul: “We bieden de Nederlandse onderzoeker een online informatie-infrastructuur, bestaande uit een website en een online platform voor werkgroepen. De website bevat informatie over datamanagement, ook bedoeld voor het algemene publiek. Op het platform kunnen onderzoekers specifieke vragen stellen, richtlijnen en best practices vinden en contactgegevens opzoeken van bijvoorbeeld een juridische specialist die veel weet over zeggenschap. Ook komt er een digitale nieuwsbrief die eens per maand wordt verzonden naar geïnteresseerden.”

Roadmap RDM

Tevens heeft het LCRDM de zogenoemde roadmap Research Data Management opgesteld die de huidige situatie van RDM in Nederland beschrijft. Daarvoor heeft SURFsara een dertigtal deskundigen bij universiteiten en onderzoeksinstellingen om input gevraagd. Verheul: “Uit deze interviews bleken vijf vraagstukken voor veel betrokkenen belangrijk: Bewustwording, Ondersteuning, Juridische Aspecten & Zeggenschap, Financieel en Faciliteiten & Data-Infrastructuur. In de roadmap beschrijven we wat het LCRDM op deze gebieden de komende drie jaar gaat doen.”

De roadmap is besproken met de Stuurgroep Onderzoek en Valorisatie van de VSNU en de vijf vraagstukken worden nu opgepakt door werkgroepen. “Voor de werkgroep Juridische Aspecten en Zeggenschap zijn al twee trekkers gevonden: Marlon Domingus van de EUR en Esther Hoorn van de RUG. Zij zijn al begonnen.”

Verder heeft Faciliteiten & Data Infrastructuur prioriteit. “Dat werd als belangrijkste vraagstuk gezien omdat iedere universiteit daar nu - in verschillende stadia - mee bezig is.” Het LCRDM werkt momenteel aan de samenstelling van deze werkgroep.

Met zijn aanpak wil het LCRDM een internationaal voorbeeld zijn. Het project loopt in ieder geval tot eind 2017. Per 2020 moet RDM een vanzelfsprekend onderdeel zijn van het Nederlands wetenschappelijk onderzoek en onderwijs, aldus de roadmap.

ingeborg.verheul@surfsara.nl
https://www.surf.nl/innovatieprojecten/duurzame-data.html

Centrale aanpak met centrale voorzieningen

RDM bij de Open Universiteit

De Open Universiteit heeft in 2014 een Research Data Managementbeleid vastgesteld. Een centrale aanpak met centrale voorzieningen is volgens de betrokkenen de sleutel tot succes. ***Jos Rikers***

Onze centrale aanpak start met het implementeren van het beleid door een projectgroep. Naast de projectleider uit de centrale organisatie bestaat deze groep voornamelijk uit ervaren onderzoekers en DANS, extern expert op het gebied van datamanagement en data archivering. Vervolgens hebben we de prioritering van het project bepaald. De kennisachterstand moest worden ingelopen en onderzoekers met acute vragen (wat is een datamanagementparagraaf; hoe schrijf ik een datamanagementplan; waar kan ik data archiveren) werden als eerste geholpen. Met DANS werd gesproken over mogelijke trainingen en cursussen, waaronder de RDNL-cursus voor data librarians.

Informatie op intranet

Uit deze ervaringen werd geput bij de volgende stap: het formuleren van de ondersteuning voor onderzoekers aan de hand van de verschillende fasen in het onderzoekproces. Het resultaat van deze stap wordt verwoord in een intranet. Bij de informatievoorziening houden we rekening met verschillen tussen wetenschapsgebieden (zoals verschillen in metadata standaarden) en verwijzen we naar externe bronnen, opleidingen e.d. Onze intentie is om van dit intranet een internetsite af te leiden (beschikbaar eind 2016).

De laatste stap is het ontwikkelen van een interactieve tool die de onderzoeker in de diverse fasen van een onderzoek voorziet van informatie op het gebied van Research Data Management. Deze interactieve tool houdt rekening met de context van het onderzoek en geeft actief tips en richtlijnen aan de onderzoeker, die dan niet meer op zoek hoeft naar informatie. Ook deze tool willen we graag delen als daar belangstelling voor is.

Enkele acute vragen van onderzoekers hebben we al kunnen beantwoorden. OU-onderzoeker Eric Kluijfhout: “Zo heeft het Europese RAGE-project (rageproject.eu), waarvan wij lead partner zijn, gekeken of de datasets die uit het project voortkomen, bij DANS kunnen worden opgeslagen. Een eerste set is bij wijze van test gedeponeerd. Het bleek echter nog niet praktisch uitvoerbaar om datasets gedurende het project bij DANS op te slaan en dan onderzoekers uit diverse landen toegang tot die datasets te geven voor het uitvoeren van analyses. Wel heeft DANS ons uitstekend geholpen bij onze verplichting een datamanagementplan te ontwikkelen in de eerste fase van het project.”

Prettig en professioneel

We zijn tevreden over ons centraal Research Data Managementbeleid. Het beleid geeft onze onderzoekers duidelijkheid. Ze kunnen in hun interactie met partners, subsidieverstrekkers en anderen op dit beleid terugvallen. En dat werkt erg prettig en professioneel.

Jos Rikers (MSc) is senior beleidsmedewerker onderwijs en onderzoek bij de Open Universiteit.
Jos.Rikers@ou.nl

AGENDA

22 - 25 februari • Amsterdam
Digital Curation Conference
Het thema van deze conferentie is ‘Visible data, invisible infrastructure’.
dcc.ac.uk/events/idcc16

1 - 3 maart • Tokio
Research Data Alliance
De RDA werkt aan het realiseren van open toegang tot onderzoeksdata. Twee keer per jaar is er een bijeenkomst om de voortgang van resultaten van de werkgroepen te presenteren.
rd-alliance.org

13 maart • Amsterdam
Paradisolezingen 2016
Voor de Paradisolezingen heeft de KNAW acht topwetenschappers uitgenodigd te komen spreken over cruciale vragen die de wetenschap de komende jaren denkt te gaan beantwoorden. De lezingen vinden plaats op zondag van 11.00 tot 13.00 uur in Paradiso te Amsterdam.
knaw.nl/nl/actueel/agenda

22 - 23 maart • Amersfoort
ICT.OPEN 2016
Tijdens deze jaarlijkse conferentie voor onderzoekers op het gebied van ICT en micro-elektronica ontmoet de wetenschap elkaar.
www.ictopen.nl

4 april • Amsterdam
InterScience
De Jonge Akademie organiseert een serie publieksbijeenkomsten waarbij verschillende leden vanuit hun eigen vakgebied eenzelfde thema belichten.
dejongeakademie.nl

20 mei • Amsterdam
Bijeenkomst eHumanities
Tijdens deze bijeenkomst wordt teruggeblikt op 5 jaar Computational Humanities: wat hebben we geleerd?
ehumanities.nl

31 mei - 3 juni • Bergen
IASSIST 2016
Het thema van dit jaarlijkse IASSIST-event is ‘Embracing the ‘data revolution’: opportunities and challenges for research’.
iassistdata.org/conferences

7 - 9 juni • Göttingen
Internationale Conferentie over Electronisch Publiseren
Wetenschappers, uitgevers, docenten, librarians, ontwikkeleraars en andere stakeholders komen samen tijdens dit event en delen ervaringen vanuit de eigen context.
meetings.copernicus.org/elpub2016

9 - 10 juni • Belval
DHBenelux - Conference for Digital Humanities Research
Deze conferentie op het gebied van digitale geesteswetenschappen wordt voor de derde keer georganiseerd.
dhbenelux.org

Arjan El Fassed, directeur Open State Foundation:

‘Kennis is macht, mits gedeeld’

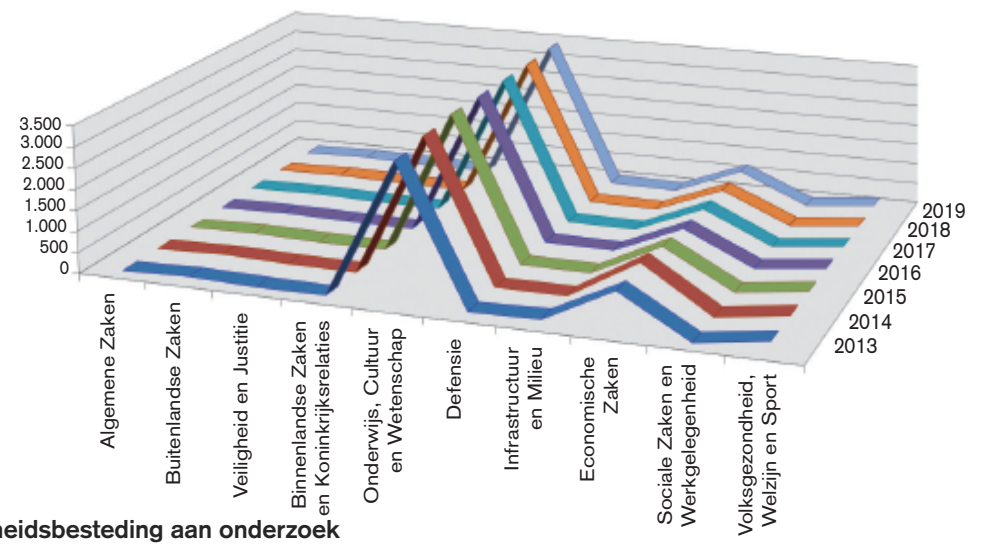
Overheden huren vaak externe adviesbureaus in voor onderzoek. Het resultaat van die onderzoeken wordt beschikbaar gesteld, maar de onderliggende data vaak niet. Wat betekent dit in de praktijk? *Arjan El Fassed*

Met bijna 5 miljard euro per jaar geven overheden opdracht voor grote en kleine onderzoeken. Deze onderzoeken herbergen een schat aan data die met publieke middelen zijn bekostigd en die hergebruikt zouden kunnen worden. Maar doordat onderliggende data niet beschikbaar zijn, is het voor derden onmogelijk de kwaliteit van deze onderzoeken te verifiëren en te valideren, terwijl het resultaat van die onderzoeken van grote invloed is op overheidsbeleid.

Wake up call

In de wetenschap is data delen al jaren gemeengoed, zoals onder kwantitatieve sociologen, of is dat recent geworden, zoals onder archeologen. Onder psychologen was de affaire Stapel een wake-up call. In een gesloten datacultuur bestaan echter geen mogelijkheden data te controleren en kan manipulatie met onderzoeksgegevens gemakkelijk onopgemerkt blijven. Bovendien betekent een gebrek aan open data cultuur ook, dat nuancering ontbreekt. Toen twee jaar geleden het rapport Werk aan de Wijk van het Sociaal en Cultureel Planbureau (SCP) verscheen, kopte De Volkskrant dezelfde dag: ‘Aanpak Vogelaarwijken mislukt’. Een te snel getrokken conclusie, want in een tijd van economische crisis was het juist een succes dat zwakkere wijken niet verder achteruit gaan.

Directe uitgaven voor R&D per departement (op kasbasis) in miljoenen euro’s per jaar



Overheidsbesteding aan onderzoek

Arjan El Fassed: “Deze tabel laat zien dat ministeries miljarden investeren in onderzoek, gefinancierd met publiek geld. De enorme hoeveelheid data die uit dit onderzoek vrijkomen, leveren economisch en maatschappelijk rendement op, maar dan moeten we de data wel structureel delen”

bron Rathenau Instituut

Vijf tips voor het delen van data

Open State Foundation maakt met open data en hergebruik publieke informatie digitaal transparant. Vijf tips voor onderzoekers:

1. Weet wat je hebt. Doe een data inventarisatie.
2. Maak de data toegankelijk via open machine-leesbare formaten.
3. Maak gebruik van open licenties, bij voorkeur cc0.
4. Zorg ervoor dat de data vindbaar is.
5. Communiceer erover en stimuleer het hergebruik ervan.

Democratisch gat

Op de website van het SCP stond het desbetreffende rapport (als PDF-bestand) en een bijlage met een uitleg over de gebruikte databestanden. Zo was voor het onderzoek gebruik gemaakt van de Woonmilieudatabase, de Integrale Veiligheidsmonitor, de Leefbaarometer en het Woononderzoek Nederland (WoON). Weliswaar kon je wel doorklikken naar een uitleg over deze bestanden, maar kon je niet bij de data zelf. En dat terwijl al deze bestanden voortkomen uit onderzoek gedaan in opdracht van en gefinancierd door ministeries (VROM, Binnenlandse Zaken en Justitie) en diensten als de Rijksplanologische Dienst (RPD) en het Centraal Bureau voor de Statistiek (CBS).

Het ontbreken van data voortkomend uit onderzoek in opdracht van de overheid betekent naast een democratisch gat ook dat duplicatie plaats vindt. Tel eens alle onderzoeken op dat jaarlijks in opdracht van gemeenten gedaan wordt. Denk bijvoorbeeld aan onderzoeken op het gebied van publieke dienstverlening of het terrein van decentralisaties. Dat behelst veel data in de vorm van survey onderzoek. Deze data zouden op een eenvoudige manier hergebruikt kunnen worden, ware het niet dat deze data niet beschikbaar komt.

Meer mogelijk

Welke mogelijkheden en nuanceringen ontstaan, als de uitkomsten onderling vergelijkbaar zouden zijn? En valt er niet van alles te besparen als onderzoeksdata herbruikbaar zouden zijn? Hoe betrouwbaar zijn onderzoeken eigenlijk en hoeveel overlap zit tussen de onderzoeken? Het beschikbaar maken van onderzoek data stimuleert ook nog eens innovatie. Uit onderzoek in onder andere Denemarken en het Verenigd Koninkrijk blijkt dat het delen van onderzoek data een enorm economisch voordeel kan opleveren voor midden en klein bedrijven, waardoor ook economisch rendement te behalen valt.

Goed nieuws

Kortom: het openstellen van onderzoekdata creëert een economische en maatschappelijke meerwaarde, stimuleert innovatie en hergebruik, en zorgt voor sneller profijt van wetenschappelijke ontdekkingen. Het goede nieuws is dat Algemene Rijksvoorwaarden voor het verstrekken van opdrachten tot het verrichten van diensten (ARVODI) bestaan. Met deze voorwaarden wordt de overheid eigenaar van data die voortkomen uit onderzoek waar ze zelf opdracht toe heeft gegeven. Sommige ministeries schrappen deze bepaling echter uit contracten. En als het al in de contracten staat, wordt het niet altijd nageleefd. Ook inkopers van onderzoek op decentraal niveau zijn zich vaak niet bewust dat overheden dergelijke voorwaarden kunnen opleggen. Dat is zonde en hiermee gaat veel waarde verloren en wordt de overheid en de belastingbetaler op kosten gejaagd.

Arjan El Fassed is directeur Open State Foundation
openstate.eu/nl

Dutch Techcentre for Life Science

Expertiseplatform voor de life sciences

Met het Dutch Techcentre for Life Sciences (DTL) is Nederland een expertiseplatform van onderzoeksorganisaties uit de life sciences rijker. Ruben Kok, directeur DTL, vertelt. *Rutger Nugteren*

“Met inmiddels 35 partners willen we een duurzaam en samenhangend netwerk vormen van lokale expertisegroepen en hun geavanceerde onderzoeksfaciliteiten, biobanken en databanken. We brengen experts uit het brede veld van de life sciences bij elkaar,” aldus Ruben. “We helpen bijvoorbeeld biomedische en klinische wetenschappers, maar ook onderzoekers uit de (agro)genomics sector, voedingsonderzoek en biotechnologie, actief bij het vinden van zowel expertise als infrastructuur. Ook bundelen we de kennis die binnen de partners aanwezig is ten aanzien van gemeenschappelijke uitdagingen, zoals data stewardship (rentmeesterschap) en methoden en standaarden voor het combineren van data.”

Ruben Kok
foto Thijs Rooijmans

Alle partners beseffen dat bepaalde aspecten van onderzoek beter in gezamenlijkheid gedaan kunnen worden. Ruben: “We leren veel van elkaar en voorkomen dat iedereen het wiel opnieuw uitvindt. Synergie ontstaat ook door verschillende disciplines bij elkaar te brengen. DTL werkt hierin nauw samen met onderzoeksfinanciers zoals de NWO en ZonMw.”

Data4lifesciences

Een ander voorbeeld is de samenwerking tussen UMC’s, NFU, SURF en DTL. Samen werken ze aan ‘Data4lifesciences’, een data- en ICT-programma, met als doel een gezamenlijke onderzoeksdata infrastructuur. “De rol van DTL ligt met name op het versterken van het interne expertise-netwerk van de UMC’s en op de aansluiting met instituten en initiatieven buiten de UMC’s. Biobanken en cohorten, experimentele faciliteiten, databanken uit onderzoek en zorg: allemaal noodzakelijke ingrediënten voor een samenhangende infrastructuur die essentieel zal zijn voor toekomstig onderzoek op het vlak van gezondheid. Dit vereist een hoogwaardige computationele omgeving die niet gebouwd is op grote centrale capaciteit, maar meer op een aanpak van gedistribueerde data analyse. DTL buigt zich hier graag over.

Wilt u zich aansluiten bij het DTL platform? Heeft u vragen over DTL of over FAIR data? Neem dan contact op met Ruben Kok.
ruben.kok@dtls.nl

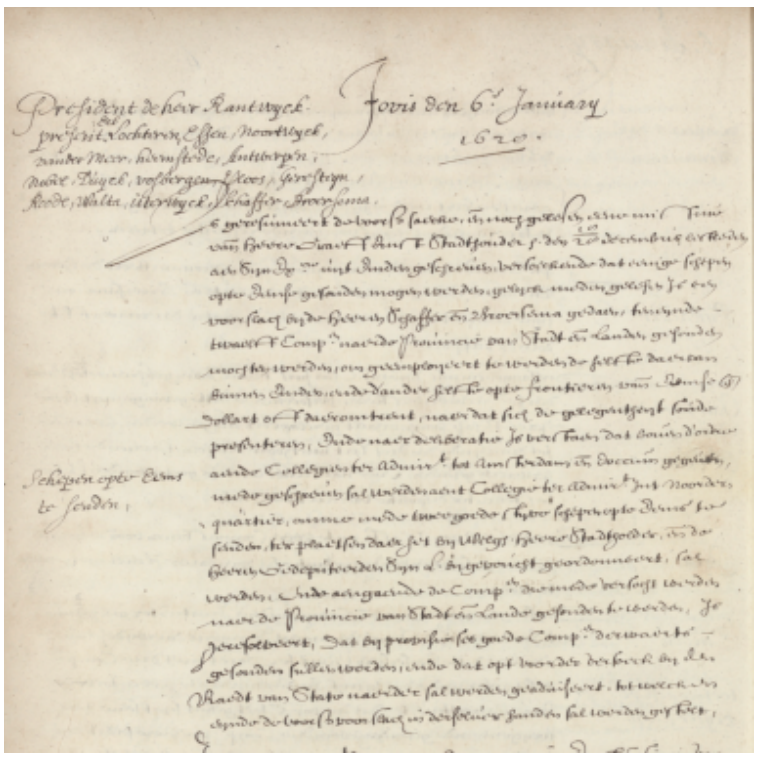
FAIR-data aanpak

DTL hanteert in haar werk de FAIR-data aanpak, afgeleid van de internationaal ontwikkelde FAIR data principles (datafairport.org).
FAIR staat voor Findable (vindbaar), Accessible (toegankelijk), Interoperable (uitwisselbaar) en Reusable (herbruikbaar). “Door data goed te beschrijven, verbanden te leggen en betekenis en context aan datasets toe te voegen, worden datasets extra waardevol omdat ze met informatie uit andere databronnen verrijkt kunnen worden. Hierdoor kunnen nieuwe inzichten worden verworven uit data, en kan nieuwe kennis worden verkregen.”

EU-project Transcriptorium ontwikkelt HTR-tool

Automatisch herkennen van handschriften weer dichterbij

Om historische tekstbronnen geschikt te maken voor digitaal gebruik, moeten ze machineleesbaar zijn. Het EU-project Transcriptorium experimenteerde met Handwritten Text Recognition. ***Edwin Klijn***



Een pagina uit de Resolutie van de Staten-Generaal *bron Huygens ING*

Voor handgeschreven teksten volstaat OCR-technologie niet. In het EU-project Transcriptorium (met als partners onder andere het Huygens ING, de Universiteit van Innsbruck en het Instituut voor Nederlandse Lexicologie) is drie jaar geëxperimenteerd met Handwritten Text Recognition (HTR). Deze technologie probeert interactief en voorspellend handgeschreven gedigitaliseerde teksten in machineleesbare teksten om te zetten. Op een workshop, gehouden op 27 november in Den Haag, werden de eindresultaten besproken én uitgeprobeerd, in een hands-on sessie rondom de transcriptietool Transkribus.

60% goed

Véronica Romero (Universitat Politècnica de València) introduceerde het Transcriptorium-project, waarbij HTR-technologie als test is ingezet op handgeschreven materiaal van de Engelse filosoof Jeremy Bentham en schrijfster Jane Austen. De resultaten laten zien dat al veel kan worden bereikt met HTR-technologie door het inzetten *prior knowledge methods*, zoals layout analyse, tekstregeldetectie en -extractie en lexical and language modelling.
Binnen het Transcriptorium-project is ook een pilot met HTR-technologie uitgevoerd op het handgeschreven deel van de Resoluties van de Staten-Generaal. Jesse de Does, computationeel taalkundige bij het Instituut voor Nederlandse Lexicografie (INL) legt uit: “De eerste resultaten waren schrikbarend: 68% van de woorden was incorrect! Na finetuning van de software slaagde men erin de Word Error Rate (WER) terug te brengen tot 40,4%. Het is een goed teken dat experts tijdens een workshops van het project zich vooral bekommeren om de WER. Want 40% fout kun je net zo goed zien als 60% goed.”

Meerdere toepassingen

Walter Ravenek (Huygens ING) vertelde hoe door toepassing van tools van onder meer de Stanford Natural Language Processing Group gedigitaliseerde corpora beter toegankelijk kunnen worden gemaakt op onder meer datum, geografische locatie en personen. Günther Mühlberger (Universiteit van Innsbruck) introduceerde de opvolger van het Transcriptorium-project: READ (Recognition and Enrichment of Archival Documents). READ richt zich sterk op de toepassing van HTR-technologie bij het digitaal toegankelijk maken van archiefcollecties. Het project wil nadrukkelijk oplossingen bieden die toepasbaar zijn op grote hoeveelheden documenten. READ gaat op basis van de Transkribus-tool verder bouwen aan een cloud-service waarin diensten worden aangeboden op het gebied van HTR, lay-outanalyse, document understanding en language modelling. Ook gaat er geëxperimenteerd worden met automatische handschriftherkenning (Famous Hands).

Enorme sprong mogelijk

De workshop Automated Handwritten Text Recognition liet zien dat een van de uitdagingen is om oplossingen te ontwikkelen die relatief goedkoop zijn en kunnen worden geïntegreerd in het productieproces van massadigitaliseringsstraten. Als men de beperkingen van de huidige technologie accepteert, is het mogelijk om met relatief kleine investeringen een enorme sprong te maken in het toegankelijk maken van archieven.

Edwin Klijn werkt bij het Nederlands Instituut voor Oorlogsdocumentatie (NIOD)
huygens.knaw.nl

GELEZEN

Beyond Open Access to Open Publication and Open Scholarship, John W. Maxwell (Simon Fraser University, Canada)
Dirk Roorda

Dit artikel maakt duidelijk dat het bij Open Access niet alleen om gaat dat het lezen van artikelen gratis wordt, maar dat het digitale paradigma een revolutie aan het bewerkstelligen is in de wetenschappelijke communicatie. Het artikel bevat een aantal catch-phrases die aangeven wat er aan de hand is.

Hier zijn er alvast twee:
1. Lenige wetenschap (Agile scholarship). Vroeger betekende publiceren dat een werk afgerond was en vervolgens openbaar gemaakt werd. Nu gebruiken groepen ook elkaars tussenresultaten, die dan wel openbaar moeten zijn. Zo is het afronden losgekoppeld van het openbaar maken.
2. Publiceren is publiek verzamelen (gathering an audience). In de digitale wereld is het een klein kunstje om iets openbaar te maken. De grote kunst is anderen zover te krijgen dat ze het aandacht geven. Publiceren is nu meer dissemineren geworden. Actief netwerken, met een zichtbare rol voor collega’s en het publiek. Al met al helpt dit artikel om de eigen (discipline-specifieke) activiteiten in een breder kader te zien.

http://src-online.ca/index.php/src/article/view/202

COLUMN

Zo eenvoudig is metadateren niet in de praktijk

Voor een historisch letterkundige studie die ik aan het schrijven ben, heb ik de afgelopen twee jaar een paar honderd boeken en artikelen moeten lezen. Ik las ze op papier en digitaal. Terugkijkend is het lezen van fotokopieën mij het slechtst bevallen. Fotokopieën ga ik te lijf met een potlood en markers in verschillende kleuren. Met het potlood maak ik aantekeningen in de marge, met de markers maak ik een samenvatting. Ik highlight eerst de grote lijn van het verhaal, plus passages die me om een of andere reden nuttig lijken. Vervolgens vat ik de highlights samen in een andere kleur.
Dat lijkt een redelijk efficiënt systeem, maar het komt erop neer dat je, als je iets wilt naslaan, de hele tijd in stapels kopieën aan het bladeren bent. Dan geef ik toch de voorkeur aan een boek – dat bladert makkelijker. Hoewel ik uiteindelijk vrijwel alles digitaal heb gelezen, zou ik de echte boeken alleen al om die reden niet willen missen.

Digitaal lezen doe ik op m’n iPad. Ik lees boeken het liefst in pdf-formaat, omdat je er van alles mee kunt. Je kunt een pdf bijvoorbeeld makkelijk dupliceren, zodat je een schoon exemplaar kunt bewaren naast een exemplaar om digitaal aantekeningen in te maken. Als ik maar een of twee hoofdstukken uit een boek nodig heb, verwijder ik de andere hoofdstukken uit het duplicaat.
De grootste winst van digitaal lezen zit wat mij betreft in de annotatiemogelijkheden. Ik gebruik daar een buitengewoon handige app voor – iAnnotate – waarmee je diverse soorten aantekeningen aan een pdf kunt toevoegen: beeld, geluid, teksten, hightlights in alle kleuren van de regenboog, uitroeptekens, vraagtekens, stempels – je kunt het zo gek niet bedenken.

Toen mij eenmaal duidelijk was welke onderwerpen ik in mijn studie wilde opnemen, ben ik mijn digitale bronnen gaan verrijken met metadata. Zo zette ik bij alle theologische verhandelingen bijvoorbeeld het woord ‘theotag’. Vervolgens kun je alle bronnen op dat woord doorzoeken – al dan niet via een index – wat veel tijd scheelt.
De theorie achter metadateren is relatief simpel, maar de afgelopen twee jaar heb ik ondervonden hoe weerbarstig de praktijk kan zijn. Om goede metadata te kunnen maken, moet je eerst patronen in je bronnen herkennen. Maar om heldere patronen in je bronnen te herkennen, moet je er eerst veel hebben gelezen. En moet je ze grondig hebben gelezen.

foto Leo van Velzen

Ik zou hier graag vertellen hoe schoon en helder gestructureerd mijn digitale bronnenverzameling eruitziet, maar dan zou ik liegen. Ik zweer nog altijd bij digitaal lezen en metadateren is echt buitengewoon handig, maar ik kom pdf’s tegen met highlights in vier kleuren, met inconsistente metadata en met aantekeningen die me ooit helder waren, maar die ik nu niet meer begrijp.
Echt heldere metadatering van letterkundige bronnen vraagt niet alleen zeer veel discipline, maar ook een flinke dosis helderziendheid. Pas als je van tevoren weet wat je in die bronnen gaat aantreffen, kun je je onderzoek beginnen met een consistente, heldere set metadata, verankerd in een gedegen theoretisch kader.
Wie weet hoe je dat aanpakt met historisch letterkundige bronnen waar tot nu toe nauwelijks onderzoek naar is gedaan, moet het mij een keer uitleggen.

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