E-DATA & RESEARCH

Jaargang 6 | 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: DANS, CentERdata, CBS, CLARIN-NL, Huygens ING, Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica

INHOUD

COLOFON
Uitgever: Stichting Uitgeverij E-data & Research Den Haag
Redactieadres: Postbus 93067, 2509 AB Den Haag, 070-3446484 edata@dans.knaw.nl
Redactie: Inge Angevaare, Eric Balster, Heidi Berkhout, Ronald van der Bie, Peter Boot, Warna Oosterbaan (hoofd-/eindredacteur), Thijs Hermsen, René van Horik, Erica Renckens
Redactiesecretariaat: Lucas Pasteuning
Aan dit nummer werkten mee: Ingrid Dillo, Ilya van Marle, Ewoud Sanders, Steamwork Graphics, Maarten Streefkerk, Milja van Tielhof, Leo van Velzen, Rosa Vitalie
Opmaak: Colette Sloots, Haarlem
Productie: Amsterdam University Press
Druk: Ten Brink, Meppel
Oplage: 9100
ISSN: 1872-0374
E-data & Research is online te raadplegen op www.edata.nl. Toezending is kosteloos aan relaties van de stakeholders en op verzoek aan studenten in de alfa- en gammarichtingen.

Inscannen van oude handschriften maakt ze voor iedereen toegankelijk. Maar de computer kan er dan nog niet in zoeken. Honderden vrijwilligers die de documenten woord voor woord overtikken lossen dit probleem op. Twee praktijkvoorbeelden.

Zonder handwerk lukt het nog niet

Handgeschreven Militieregisters worden thuis overgetikt. *Erica Renckens*

Henny van Schie van het Nationaal Archief *foto Leo van Velzen*

De laatste jaren hebben archieven, bibliotheken en onderzoeksinstellingen veel tijd en geld gestoken in het digitaliseren van hun collecties. Hoewel de resultaten van Optical Character Recognition (OCR) lang niet perfect zijn, is veel van het gedrukte materiaal al redelijk goed te doorzoeken met behulp van deze techniek. Voor oude drukken en handschriften is de situatie een stuk minder rooskleurig. Je kunt ze wel inscannen en daarna op een scherm bekijken, maar de computer kan er dan nog niet in zoeken. Verschillende nieuwe technieken en *crowdsourcing*-projecten moeten hier verandering in brengen.
Afgelopen november lanceerde het Stadsarchief Amsterdam de website VeleHanden.nl. Via deze website kan iedereen helpen om de Militieregisters *online* doorzoekbaar te maken. In deze handgeschreven registers staan de namen en gegevens van alle dienstplichtige mannen van 1815 tot en met 1941. De gegevens van elke scan worden door twee vrijwilligers overgetypt, waarna een controleur zich ervan verzekert dat alles correct is overgenomen. Als dank krijgen de vrijwilligers toegang tot informatie uit de Militieregisters over hun eigen familie.
Eind december, amper twee maanden na lancering, had VeleHanden.nl al bijna 800 leden die gezamenlijk bijna 150.000 scans hadden ingevoerd. Het Stadsarchief is overweldigd door dit resultaat en is er van overtuigd dat deze vorm van *crowd sourcing* de toekomst is voor de culturele sector. Het Amsterdamse archief is niet de enige die de crowd inzet om data te ontsluiten; aan het Meertens Instituut lopen geregeld vrijwilligersprojecten, momenteel rond de Gekaapte Brieven.

Vrijwel foutloos

Het grote voordeel van de inzet van een grote hoeveelheid vrijwilligers is dat de gegevens zonder veel kosten volledig beschikbaar worden, en toch vrijwel foutloos worden getranscribeerd. Nadeel van het woord voor woord overtypen is de grote inspanning die geleverd moet worden: als 'Jansen' tienduizend keer voorkomt, zal hij ook tienduizend keer ingetypt moeten worden.
"Die arbeidsintensieve aanpak is voor zulke projecten onvermijdelijk, omdat je streeft naar volledigheid", aldus Henny van Schie, archivaris bij het Nationaal Archief. Zelf werkt hij binnen het CATCH-Plus-project Scratch4All aan een computerprogramma dat digitale handschriften met een relatief ge-

VERVOLG OP PAGINA 3

De buit bestaat uit bits en bytes

Gekaapte brieven en verhoren toegankelijk gemaakt voor onderzoek. *Ingrid Dillo*

The National Archives (TNA) in Kew (Londen) bewaart het archief van de High Court of Admiralty. Dit archief bevat naar schatting veertigduizend in het Nederlands geschreven brieven en documenten. Dat is een deel van de buit die Engelse kapers in de zeventiende en achttiende eeuw veroverden op Nederlandse schepen. Die buit bevat brieven van en aan zeelieden en kooplieden en hun familieleden, scheepsjournalen, aanbevelingsbrieven, ladingboekjes en kwitanties. Eeuwenlang lagen ze in de donkere kelders van het High Court of Admiralty, zonder dat iemand ernaar omkeek. Sommige brieven zijn tot op de dag van vandaag niet geopend.

Papieren erfgoed

Een klein deel van dit materiaal is in het kader van het nationale programma voor het behoud van het papieren erfgoed, Metamorfoze, gedigitaliseerd. Dit heeft een kleine negenduizend scans opgeleverd. Taalkundige Nicoline van der Sijs heeft samen met het Meertens Instituut en met subsidie van het Prins Bernard Cultuurfonds het initiatief genomen deze scans te ontsluiten en te transcriberen. "Het werk wordt uitgevoerd door een grote groep vrijwilligers. Een kleine honderdvijftig enthousiaste medewerkers is inmiddels aan de slag."
Volgens Van der Sijs is deze vorm van *crowd sourcing* een groot succes en zijn de resultaten van uitstekende kwaliteit, mede doordat vrijwilligers elkaars werk controleren. "Om de scans te ontsluiten voor wetenschappelijk onderzoek, worden ze in het project voorzien van inhoudelijke metadata. Daarnaast transcriberen de vrijwilligers zoveel mogelijk scans. Zo ontstaat er een interessant corpus voor onderzoek en wordt het bijvoorbeeld mogelijk het achttiende-eeuwse taalgebruik van de gewone man te vergelijken met de meer formele taal die werd gebezigd in bijvoorbeeld literaire teksten."
Waren de verschillen echt zo groot of hebben we dat wellicht altijd overschat? De waarheid kan binnenkort boven tafel komen", zegt Van der Sijs.

http://tinyurl.com/cmwnrtv

Het Prize Papers Project

De Engelse kapers beperkten hun buit niet tot Hollandse schepen. In twee eeuwen zijn tienduizenden Franse, Spaanse, Portugese, Deense, Zweedse, Duitse, Italiaanse en Amerikaanse schepen in Engelse havens opgebracht. Om te bepalen of een schip rechtmatig was veroverd werd er een proces gevoerd en werd de bemanning ondervraagd. Functionarissen van het Britse High Court of Admiralty zorgden voor de administratieve begeleiding. De verhoren van de gevangen genomen bemanningsleden vonden plaats aan de hand van een standaard vragenlijst van achttien en later zelfs vierendertig vragen over herkomst, route, bestemming, tonnage, lading en de eigenaren van het schip en over herkomst, nationaliteit, leeftijd en migratiegeschiedenis van de bemanning. De antwoorden werden door tolken vertaald en in het Engels genoteerd.
Dit materiaal levert een schat aan informatie op die zeer waardevol is voor de maritieme geschiedenis, maar bijvoorbeeld ook voor migratiestudies en de vroege Amerikaanse geschiedenis.
De oudste Nederlandse academische uitgeverij, Koninklijke Brill NV, heeft het initiatief genomen deze internationale bron te digitaliseren en te voorzien van inhoudelijke metadata. Brill-directeur en maritiem historicus Perry Moree vindt dat de bron uitstekend past in Brill's online research collecties. In totaal zullen er ongeveer 300.000 scans worden gemaakt. Het Prize Papers Project zal in fasen worden uitgevoerd en start met de achttiende eeuw. Het is de bedoeling dat het eerste deel van het materiaal in de tweede helft van 2012 als betaalde online database beschikbaar is voor wetenschappelijk onderzoek.

De inhoud van één van de dozen uit het archief van de High Court of Admiralty. De documenten bevatten de verhoren van door Engelse kapers gevangen genomen zeelieden.

GEHOORD & BIJGEWOOND

De loden letters en digitale dartels van Stronks

Peter Boot

Op tien januari hield onder grote belangstelling Els Stronks haar inaugurele rede als hoogleraar Vroegmoderne Nederlandse letterkunde aan de Universiteit Utrecht. Het onderwerp van haar rede was de betekenis van digitale technieken voor de studie van historische letterkunde. In hoeverre gaan de 'loden letters' – het langzame lezen en de trage bezinning – samen met wat Stronks de 'digitale dartels' noemt – de speelsheid en de dynamiek van de nieuwe digitale mogelijkheden? Aan de hand van nagelaten onderzoeksaantekeningen van een beroemde voorganger op haar leerstoel, W.A.P. Smit, laat Stronks zien dat er in de traditionele onderzoekspraktijk een grote afstand is tussen analyse en interpretatie. De grote hoeveelheden digitale tekst maken het steeds makkelijker patronen en regels te vinden. Maar de letterkundig onderzoeker is juist vaak op zoek naar de afwijkingen van de regels. Daar komen de ironie, de ambiguïteit of de paradoxen in een tekst naar voren. De uitdaging is er voor te zorgen dat het creatief en interpreterend lezen in de digitale context niet verloren gaan. Stronks pleit in haar oratie voor digitale geletterdheid van de historisch letterkundige. Letterkundigen hebben te vaak het proces van digitaliseren aan anderen overgelaten, en dreigen daardoor met onbruikbaar materiaal hun werk te moeten doen. Letterkundigen moeten aandringen op goed doorzoekbare en foutloze teksten, die voldoen aan internationale standaards en die zijn voorzien van adequate metadata.

Goed gedigitaliseerd materiaal maakt onderzoek mogelijk waarin verschillende bestanden worden gecombineerd, waarbij nu eens van grote hoogte wordt gezocht naar de patronen, en vervolgens kan worden ingezoomd op betekenisvolle details. Dat helpt het onderzoek, maar het wordt daarmee ook voor de niet-onderzoeker mogelijk om kennis te nemen van ons verleden, en bijvoorbeeld iets te leren van hoe we in het verleden omgingen met culturele diversiteit.

www2.hum.uu.nl/onderzoek/lezingenreeks/pdf/Stronks_Els_oratie.pdf

EDDI-conferentie met veel XML en Questasy

Maarten Streefkerk

December jl. vond de derde editie van de EDDI-conferentie (European DDI Users Group Meeting) plaats. Het SND (Swedish National Data Service) was de organiserende partij en de University of Gothenburg bood ruimte aan bijna 100 deelnemers. DDI (Data Documentation Initiative) is een XML standaard om (meta)data en de levensloop van onderzoek (van onderzoeksvoorstel tot archivering/disseminatie) te documenteren.

Tijdens de conferentie werden ervaringen gedeeld, voorbeelden van webapplicaties van online-archieven en gerelateerd onderzoek getoond.

Ook demonstreerden softwareontwikkelaars applicaties zoals DDI-editors, tools voor archivering van metadata en conversie-hulpmiddelen. Een aantal applicaties zijn Open Source of kunnen gratis worden gebruikt. CentERdata toonde tijdens de bijeenkomst recente vernieuwingen aan webapplicatie Questasy, een disseminatietool, welke is gebaseerd op de DDI 3 standaard. De belangrijkste vernieuwingen zijn multi-panel support (waardoor gegevens van verschillende panels naast elkaar kunnen worden gepubliceerd) en de 'variable basket', waarmee gebruikers in staat zijn om datasets samen te stellen uit verschillende studies, inclusief documentatie. Belangrijke discussiepunten waren de beperkingen van DDI 3.1 voor ontwikkelaars en het gebruik van relationele databases (en ook XML databases) om metadata op te slaan. Tijdens de bijeenkomst werd bekendgemaakt dat DDI versie 3.2 in de loop van 2012 ter beschikking komt.

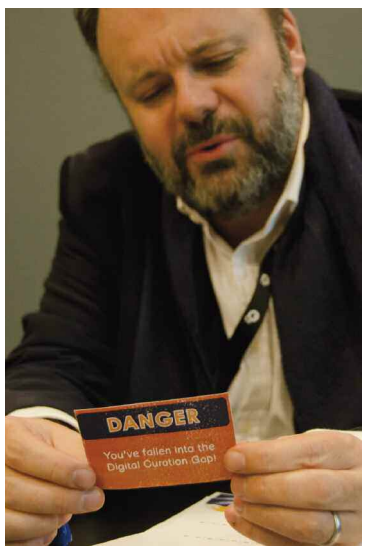
www.iza.org/conference_files/EDDI2011/

Mist erfgoed aansluiting bij digitale cultuur?

Inge Angevaare

'Digital Strategies for Cultural Heritage' (DISH) is de tweejaarlijkse internationale conferentie die Erfgoed Nederland en Digitaal Erfgoed Nederland december jl. organiseerden. De conferentie is bedoeld om erfgoedinstellingen te stimuleren nieuwe strategieën te ontwikkelen die passen in de digitale cultuur.

Deze aflevering van DISH gaf een



Shawn Day, Digital Humanities-Specialist, tijdens een digitaal spel *foto's Inge Angevaare*

iets ander beeld dan de organisatoren wellicht hadden bedoeld. In vijf plenaire toespraken (door internationale cultuurexperts als Amber Case, Samuel Jones, Charles Leadbeater) werd vooral het cultuurbeeld anno 2012 geschetst zonder dat duidelijk werd wat erfgoedinstellingen daar precies mee kunnen. De cultuur anno 2012 ontwikkelt zich razendsnel, het is een van trend naar trend hoppende chaos en waar binnen 'our cell phones change us into cyborgs' (Amber Case).

Van dit alles wordt maar weinig bewaard en voor de wetenschap toegankelijk gemaakt. Traditionele erfgoedinstellingen weten nog niet goed raad met de overdaad aan cultuur die het internet biedt. Tekenend was een parallelsessie over 'nationale infrastructuren'. In Nederland hebben een aantal partijen, waaronder de Koninklijke Bibliotheek, Beeld en Geluid en het Nationaal Archief, de handen ineen geslagen om hun digitale collecties gecoördineerd te aggregeren naar de Europese Europeana portal. Maar het gaat slechts om metadata waar je nog niet in kunt zoeken. DEN-directeur Marco de Niet zei erover: 'Dit hadden we tien jaar geleden al moeten doen.'

Maar Clifford Lynch van de Coalition for Networked Information concludeerde: 'The digital shift is disrupting our organizations in fundamental ways.' Het lijkt alsof de sector daar eerst nog aan moet wennen.

www.dish2011.nl

Nog meer slimme tools voor geesteswetenschap

Peter Boot

'Supporting the Digital Humanities' is de gezamenlijke missie van de projecten CLARIN en DARIAH, en de naam van de conferentie die de projecten op 17 en 18 oktober jl. hielden in Kopenhagen. Onder het ambitieuze motto 'Answering the unaskable' kwamen infrastructuurspecialisten en digitaal onderzoekers bijeen om van gedachten te wisselen over een digitale infrastructuur voor de geesteswetenschappen. Er was een grote Nederlandse delegatie, met vertegenwoordigers van onder ander het Meertens Instituut, DANS, Huygens ING en onderzoekers uit Twente.

Tijdens de conferentie werden voornamelijk datacollecties en slimme tools gepresenteerd, en de infrastructuur waarin data en tools kunnen worden ingebed. Voor wat betreft de tools was een gemeenschappelijke noemer het probleem van 'alignment': het vinden van parallelle plaatsen in bijvoorbeeld een tweetalig corpus, in vergelijkbare melodieën of in twee versies van eenzelfde tekst.

In de slotsessie kwamen de toekomstperspectieven voor CLARIN en DARIAH aan de orde. Voor beide projecten zijn aanvragen ingediend voor erkenning als Europees Research Infrastructure Consortium (ERIC). Veel belangstelling bestond er voor het Nederlandse initiatief om CLARIN en DARIAH samen te voegen tot CLARIAH.

http://cst.ku.dk/sdh2011/

AGENDA

6 - 10 februari • Leiden

Biblical Scholarship and Humanities Computing: Data Types, Text, Language and Interpretation

What are the requirements for text data bases to allow for the systematic study of ancient texts, especially Hebrew, Aramaic or Greek biblical texts? The question to be discussed by biblical scholars and ICT specialists is: how to deal with a historically grown and changed set of literary and linguistic data?

www.lorentzcenter.nl/lc/web/2012/480/info.php3?wsid=480

7 - 9 februari • Kopenhagen

Hackathon: A Practical Approach to Database Archiving

The large and growing volume of data held in an increasing variety of relational databases presents a huge challenge to the archiving community. This hackathon, organised by the Open Planets Foundation, is designed to bridge the gap between digital preservation practitioners and developers. ***www.openplanetsfoundation.org/comment/244***

9 februari • Utrecht

SURF-onderzoeksdag

Deze dag is speciaal bedoeld voor onderzoekers en hun ondersteuners. Op de SURF-onderzoeksdag laten onderzoekers zien hoe ICT-methoden en -hulpmiddelen wetenschappelijk onderzoek, en de presentatie daarvan, kunnen verrijken en vernieuwen. De nieuwste nationale en internationale ontwikkelingen zullen de revue passeren.

www.surf.nl/ozdag

16 februari • Den Haag

DANS-Geonovum Studiemiddag, 'Digitale duurzaamheid van geodata'

Er is sprake van een voortdurende groei van geo-informatie en ook het gebruik neemt aanzienlijk toe. Maar hoe zorgen we er nu voor dat deze informatie ook in de toekomst nog te raadplegen is? Reden voor Geonovum en DANS om een studiemiddag over duurzame toegang tot geodata te organiseren.

Tijdens deze interactieve bijeenkomst wordt gesproken over de wenselijkheid om geodata op de langere termijn toegankelijk te houden, en welke rol Geonovum en DANS daarbij kunnen spelen.

www.geonovum.nl/dossiers/kennissessies

15 maart • Utrecht

DANS-SURFfoundation Symposium Databeheer in de praktijk. Resultaten van het CARDS-project

Digitale onderzoeksgegevens krijgen een steeds grotere rol binnen wetenschappelijk onderzoek. Het CARDS-project ondersteunt onderzoekers bij het beheren van onderzoeksdata. Bij een aantal universiteiten in Nederland zijn pilots uitgevoerd waarin onderzoekers samen met de universiteitsbibliotheek het bewaren en beschikbaarstellen van data in de praktijk hebben gebracht. Tijdens het symposium worden de uitkomsten van de pilots gepresenteerd.

www.dans.knaw.nl/content/symposia

19 - 20 maart • Den Haag

Interedition. Scholarly Digital Editions, Tools and Infrastructure (Huygens ING)

Huygens ING hosts a symposium about Interedition, COST (European Cooperation in Science and Technology) Action IS0704. This event will also serve as a springboard for further work based on the principles of interoperability promoted by Interedition.

www.textualscholarship.nl/?p=10089

11 - 14 april • Glasgow

European Social Science History conference 2012

The IISH organizes the ESSHC once every two years. The conference does not have a central theme and welcomes papers about all periods and subjects. The main objective of the conference is to introduce historians who use the insights and techniques from the social sciences to social scientists that focus on the past in their research and vice versa.

www.iisg.nl/esshc/2012

21 - 27 mei • Istanbul

The eighth international conference on Language Resources and Evaluation (LREC)

LREC has become the major event on Language Resources (LRs) and Evaluation for Language Technologies (LT). The aim of LREC is to provide an overview of the state-of-the-art, explore new R&D directions and emerging trends, exchange information regarding LRs and their applications, evaluation methodologies and tools, activities, industrial uses and needs, requirements coming from the e-society, both with respect to policy issues and to technological and organisational ones.

www.lrec-conf.org/lrec2012

De opmars van de Virtual Research Environments

De wandelende werkvloer

Het is een sterke trend: samenwerken via het net. De universiteitsbibliotheken leveren er speciale werkplaatsen voor. *Peter Boot*

illustratie Rosa Vitalie

Steeds meer wetenschappers werken online samen in Virtual Research Environments (VRE's) – wetenschappelijke werkplaatsen die door universiteitsbibliotheken zijn opgezet en die onderzoekers de mogelijkheid geven om virtueel samen te werken. VRE's kunnen hulpmiddelen voor communicatie bieden (een agenda, projectbeschrijvingen, een weblog), voor samenwerking (een wiki, workspaces met documenten, beheer van apparatuur) en een repository met publicaties, toegang tot de bibliotheekinformatie en andere gegevensbronnen.

Jikke de Groot, projectmanager bij de Utrechtse UB, heeft al een flinke ervaring met dit soort omgevingen. "De fase van de grote aanpassingen zijn we voorbij. We zijn er in 2006 mee begonnen, we hebben ook al een groot gebruikersonderzoek achter de rug en we beheren inmiddels ruim 30 virtuele kenniscentra (VKC's) in alle wetenschapsgebieden."

Vijanden van de staat

Een voorbeeld van een VRE is de omgeving die de Leidse UB opzette voor het project van onderzoekster Beatrice de Graaf, 'Enemies of the State'. De Graaf: "Een belangrijk onderdeel is de database die we op reis als wandelend archief gebruiken, waar scans uit archieven en teksten in terechtkomen. Daarnaast hebben we een soort intranet gemaakt, waar mensen van buiten onze eigen groep congrespapers kunnen uploaden, en we hebben een publiekssite om de buitenwereld over ons onderzoek te vertellen. Op het moment staan die componenten nog los van elkaar, maar binnenkort gaan we de publiekssite dynamischer maken en vullen uit de database." Er zijn ook wel beperkingen, vindt De Graaf. Soms is het lastig bepaalde documenten terug te vinden. Een zoekmogelijkheid op concepten zou handig zijn. De Graaf: "Hoewel de noodzaak van heldere rubricering ook je wetenschappelijke ideeën kan aanscherpen."

De VRE's zoals de UB's die beschikbaar stellen, lijken weer net iets anders dan de VRE's waaraan het onlangs afgesloten Alfalab-project heeft gewerkt. In de versie van de UB's heeft ICT een faciliterende rol, waarbij de gebruikelijke wetenschappelijke processen efficiënter, opener en meer in teamverband worden uitgevoerd. Bij de Alfalab-achtige VRE's is sprake van meer fundamentele wetenschappelijke vernieuwing – die dan ook veel meer experimenteel van aard is. De omgevingen die Alfalab vorig jaar bij de afronding van het project presenteerde waren 'demonstrators', die digitale onderzoeksmogelijkheden toonden in labs voor onder andere tekst, gebruikersinterfaces, geografische data en levensloopgegevens. Het Gislab maakt het bijvoorbeeld mogelijk om historische kaarten te voorzien van geografische coördinaten en zo te gebruiken in combinatie met een database van veldnamen. Maar Peter Verhaar, projectleider bij de UB Leiden, ziet geen tegenstelling. "Elke VRE bevat tools gericht op het onderzoek én op ondersteuning van het onderzoeksproces." Verhaar leidde ook een project van SURF Foundation waarbij een VRE Starters Kit is gemaakt. De kit bevat informatie die onderzoekers en hun ondersteuners helpt bij het opzetten van een VRE. Daarbij is zowel aan organisatorische als aan technische aspecten gedacht.

www.uu.nl/university/library/nl/informatie/Pages/VKC.aspx
http://hum.leiden.edu/history/enemies-of-the-state/
www.surffoundation.nl/nl/projecten/Pages/SamenwerkingVREs.aspx

VERVOLG VAN PAGINA 1

Handwerk

ringe inspanning doorzoekbaar kan maken. Het project is gericht op de ontsluiting van het Archief van het Kabinet der Koningin (1814-1988). Er wordt gebruikgemaakt van software die is ontwikkeld aan de Rijksuniversiteit van Groningen, Monk geheten.

Herkenning

Monk werkt op basis van patroonherkenning, maar ook hier wordt van de diensten van vrijwilligers gebruikgemaakt. "Vrijwilligers kunnen via de website helpen om Monk te trainen", aldus Van Schie. "In een spelvorm geven zij aan waar in de scans de woorden staan en wat er precies staat. Deze resultaten neemt Monk mee in zijn berekeningen, waardoor de kwaliteit van de herkenning verbetert." Monk heeft minimaal vijf voorbeelden van hetzelfde woord in hetzelfde handschrift nodig om het een zesde keer enigszins te kunnen herkennen. "Vanaf twintig voorbeelden gaat het heel goed en bij vijftig bijna perfect." Familienamen en geografische namen komen over het algemeen niet in die aantallen voor in een handschriftencollectie. Dat betekent dat volledige herkenning nog niet haalbaar is, maar dat is ook niet wat het project nastreeft.

"Scratch4All wil met Monk een instrument maken dat onderzoekers en erfgoedinstellingen kunnen gebruiken", licht Van Schie toe. "Net als bij ingescande kranten zal niet elk zoekwoord resultaten opleveren, maar wel zoveel dat je gericht en efficiënt verder kunt zoeken."

De resultaten van Scratch4All zullen in de loop van 2012 ook toegankelijk worden voor andere erfgoedinstellingen.

Eén punt voor elke ingevoerde tekst

Vrijwilligers doen hun werk voor VeleHanden.nl vanuit hun eigen huis. Via de website kunnen ze zich inschrijven, waarna ze een korte training volgen. De vrijwilligers krijgen vervolgens willekeurige scans voorgeschoteld om te transcriberen. Elke scan wordt door twee vrijwilligers overgetypt, waarna een controleur checkt of alles goed is overgenomen. Elke ingevoerde scan levert de vrijwilliger één punt op. Voor vijfentwintig punten kan hij een scan naar keuze uit de Militieregisters downloaden. Mensen die geïnteresseerd zijn in genealogie worden zo gestimuleerd om mee te werken aan de ontsluiting van de registers.

foto Leo van Velzen

Zoeken door de hele Loe de Jong

Peter Boot

In december jl. plaatste het NIOD (Instituut voor Oorlogs-, Holocaust- en Genocidestudies) een digitale versie online van Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog. De dertig banden van het geschiedwerk van Loe de Jong kunnen als pdf-bestanden worden gedownload.

De servers stonden roodgloeiend, meldt David Barnouw, NIOD's persvoorlichter. De eerste dagen moest in allerijl extra capaciteit worden bijgeschakeld. In december is de NIOD site meer bezocht dan in de elf voorafgaande maanden. Er komen veel enthousiaste reacties binnen, zowel van historici als van het grote publiek.

Alle delen downloaden

Op diverse weblogs en op Twitter werd na de publicatie ook wel kritiek geuit. Om in alle delen te zoeken, moet je bijvoorbeeld alle delen downloaden. Barnouw: "Het is waar dat mensen die al vaker met digitale uitgaven hebben gewerkt ook wel met kritiek of suggesties komen, maar we hebben dit ook niet in de eerste plaats voor de voorlopers gemaakt. Een uitgebreidere versie is bovendien in de maak."

Edwin Klijn, teamleider Diensten bij het NIOD, vertelt over de plannen voor verdere ontsluiting. Van CLARIN is onlangs subsidie gekregen voor een samenwerkingsproject ('Verrijkt Koninkrijk') met de Universiteit van Amsterdam en de Vrije Universiteit waarin de delen doorzoekbaar worden gemaakt. Bovendien zal de inhoud ter beschikking komen in de vorm van Linked Open Data (waarmee de computer tot op zekere hoogte de inhoud kan begrijpen en koppelen aan andere gegevensverzamelingen). Klijn: "We gaan deze ontsluiting testen in een onderzoek naar De Jong's interpretatie van de verzuiling tijdens de oorlog."

De CLARIN-projecten hebben een doorlooptijd van een jaar. Begin 2013 kunnen we dus meer verwachten. Klijn: "Maar er zijn nu al een paar verbeteringen aangebracht. Je kunt nu alle delen in één keer ophalen, doordat we het aanbieden als een torrent. Binnenkort komen we nog met een bijgewerkte versie van de PDF's, waarin een aantal errata is gecorrigeerd. En de gegevens zijn beschikbaar onder heel vrije licentievoorwaarden: wie verwijst naar het NIOD heeft het recht om zelf met een andere publicatievorm te komen!"

www.niod.knaw.nl/koninkrijk

Tot Nut van den Javaan

Het Nederlandse middenveld was breed in de negentiende eeuw. Op de website van het Huygens ING staan duizenden verenigingen uit die tijd. *Milja van Tielhof*

In 1830 werd in Utrecht de Maatschappij voor Moederlijke Liefdadigheid opgericht. De leden hielpen behoeftige kraamvrouwen met een groot gezin. De kraamvrouw kon rekenen op bonnen voor vier bossen stro, zeven broden, wat kruidenierswaren en zeep, een deken, een stel kleren voor de baby of een mand met kleertjes en luiers. Tussen november en mei kregen ze ook wat brandstof. Er waren wel voorwaarden: het gezin moest in Utrecht wonen en er moesten al drie kinderen zijn. Verder moest de zwangere vrouw zich drie maanden voor de bevalling melden, zodat de vereniging het gedrag van man en vrouw kon onderzoeken. De hulp werd ook verstrekt wanneer de vrouw beviel van een levenloos kind, maar dan verviel het stel kleren of de luiermand.

De Utrechtse Maatschappij is een van de organisaties in de database *Verenigingen voor armenzorg en armoedepreventie in Nederland in de negentiende eeuw*, een van de drie databases van verenigingen die in 2011 door het Huygens ING zijn gepubliceerd op de website Historici.nl. De tweede betreft *Sportbonden, sportclubs en sportperiodieken tot 1940*, en geeft een overzicht van de snelle ontwikkeling van gymnastiek, hockey, korfbal, schaken, tennis en voetbal. De derde database heet *Erkende verenigingen, 1855-1903* en bevat alle verenigingen die bij Koninklijk Besluit erkend waren als rechtspersoon. Daaronder vallen schoolverenigingen, sociëteiten, zangclubs, woningbouwcorporaties en politieke organisaties zoals het Anti-dagbladzegelverbond en de Maatschappij tot Nut van den Javaan. Alle drie de databanken bevatten vele duizenden verenigingen.

Sportvereniging HFC in 1887

Bloeiend verenigingsleven

In de loop van 2012 komen er ook nog een databank met rooms-katholieke religieuze broederschappen in de negentiende eeuw, een databank met sociale voorzorgfondsen (1827-1880) en een databank met genootschappen van patriotten en prinsgezinden (1780-1795). Per organisatie worden de oprichtingsdatum, plaats van vestiging en doelstelling gegeven plus, afhankelijk van het project, andere bijzonderheden zoals de levensbeschouwing van de leden.

Verenigingen waren in de negentiende en een groot deel van de twintigste eeuw actief op een opvallend breed terrein. De sterke toename van de vrije tijd creëerde ruimte voor toneelgezelschappen, voetbalclubs en ijsverenigingen. Tegelijkertijd ontstond er door industrialisering en door de enorme groei van de grote steden behoefte aan allerlei voorzieningen. De overheid bleef zich echter lang terughoudend opstellen. Particulieren sprongen in dat gat en realiseerden in verenigingsverband voorzieningen zoals scholen, banken voor microkrediet, bibliotheken, instellingen voor ziekenzorg, tehuizen voor ongehuwde moeders enzovoort. Het was tot nu toe niet eenvoudig beter zicht te krijgen op dat bloeiende verenigingsleven. Met deze databases wordt dat een stuk gemakkelijker.

www.Historici.nl/Onderzoek/Projecten/Armenzorgverenigingen
www.Historici.nl/Onderzoek/Projecten/Sportverenigingen
www.Historici.nl/Onderzoek/Projecten/ErkendeVerenigingen

En weer ligt het auteursrecht dwars

Bewaren van bestanden gaat heel goed als je een oude computer nabootst op een nieuwe. Maar het mag niet van Europa. *Inge Angevaare*

De Europese Commissie financiert veel onderzoek naar de kwetsbaarheid van digitale gegevens en naar oplossingen voor dat probleem. Eén van deze projecten is KEEP, Keeping Emulation Platforms Portable. Dit project wil handzame applicaties ontwikkelen, binnen de grenzen van de wet. Dat bleek niet mogelijk, want diezelfde Europese Commissie en de lidstaten hebben wetgeving uitgevaardigd die emulatie als techniek vrijwel ondoenlijk maakt.

Emulatie is één van de technieken die wordt gebruikt om digitale gegevens voor de lange termijn bruikbaar te houden. Bij de meer bekende migratietechniek worden digitale objecten steeds aangepast aan nieuwe hardware/softwareomgevingen. Maar hierbij gaan altijd gegevens verloren en bovendien is deze methode niet geschikt voor samengestelde, complexe objecten als bijvoorbeeld computerspellen, software, databases en websites – en juist die komen in de wetenschap veel voor.

Bij emulatie laat men de digitale codes voor wat ze zijn, maar bouwt men software die nieuwe computers kan laten functioneren als oude computers. Op een dergelijk platform kan het digitale object draaien in de 'oorspronkelijke' omgeving. Daarvoor is dan wel het hele complex aan software nodig dat bij de productie van het digitale object ook aanwezig was: het operating system, de applicatiesoftware, en alle mogelijke plug-ins. Al die software moet gekopieerd worden om in de nieuwe omgeving te kunnen draaien.

David Anderson *foto Inge Angevaare*

Grote uitgevers als Microsoft blijken in de praktijk niet of nauwelijks bereid om software en broncodes af te staan

Moeilijk doen

Maar het kopiëren van software mag niet volgens Europese richtlijnen. Dat ontdekte David Anderson van Portsmouth University die voor het KEEP-project onderzoek deed naar wat wel en niet is toegestaan. Hij ontdekte een ingewikkeld stelsel van nationale en Europese wetgeving, zo ingewikkeld dat hij zijn eindrapport slechts een *Layman's guide* durfde te noemen. Maar het algemene beeld was helder: erfgoedinstellingen en data-archieven hebben juridisch gezien geen mogelijkheden om software te kopiëren naar emulatoren om zo digitale bestanden bruikbaar te houden. Ook van elektronische beveiligingen als wachtwoorden moet men afblijven.

Het lijkt misschien vreemd dat er zo moeilijk wordt gedaan over het maken van kopieën van software – hackers doen het aan de lopende band. Maar publieke instellingen willen binnen de wet opereren, en het kopiëren van software naar emulatoren is juridisch gezien iets heel anders dan het gebruiken van de software volgens de licentie. Die emulatoren komen open source beschikbaar, en zo komen er veel kopieën op de markt.

De auteursrechtwetten kennen wel enkele kleine uitzonderingen op het gebod 'niet kopiëren' (voor privégebruik, bijvoorbeeld), maar die uitzonderingen bieden geen soelaas voor de emulatietechniek. Tijdens een workshop in Den Haag vertelde Anderson E-data dat het Europese KEEP-project zelfs zijn eigen resultaten niet kan opleveren zonder EU-regels te overtreden.

Grote commerciële uitgevers van software als Microsoft blijken in de praktijk niet of nauwelijks bereid om software en broncodes af te staan. Zij beschermen hun recht om de software – ook al is die oud – te vermarkten. Een soort deponeringsplicht voor duurzaamheidsprojecten zou daar een eind aan kunnen maken, maar die bestaat niet voor software.

Het is de hele erfgoedsector duidelijk dat de wettelijke regels ten aanzien van auteursrecht die in het analoge tijdperk zijn ontwikkeld in het digitale tijdperk dringend aanpassing behoeven. Maar dat proces verloopt uiterst traag. De uitgevers van software hebben een sterke lobby in Brussel om hun belangen te beschermen.

www.keep-project.eu

Eten en je wassen op de smartphone

Onderzoek naar de besteding van tijd gaat meestal met dagboekjes. Maar een smartphone is veel handiger. *Eric Balster*

"Je smartphone heb je altijd bij je en dat biedt ook mogelijkheden om respondenten te herinneren aan het onderzoek", zegt Henk Fernee, onderzoeker bij het Sociaal en Cultureel Planbureau (SCP). "Hoe korter de tijd tussen de activiteit en de registratie, des te betrouwbaarder de data". Zoals in de vorige E-data al werd gemeld, onderzoekt het SCP of smartphones de tot dusver in het Tijdsbestedingsonderzoek (TBO) gebruikte dagboekjes kunnen vervangen. Een ander voordeel is dat smartphones extra gegevens (paradata) kunnen vastleggen: wanneer er iets is ingevuld, of de respondent nog iets verbeterd of toegevoegd heeft, en of het sturen van herinneringen effect heeft. Het is zelfs mogelijk om - met toestemming van de deelnemers - GPS-data van de smartphone uit te lezen en informatie te krijgen over verplaatsingen. SCP werkt in dit experiment samen met CentERdata, dat de smartphone applicatie (de *App*) ontwikkelt. Aan een eerste test namen 100 personen deel; de helft met eigen smartphones, de andere helft met een in bruikleen gegeven exemplaar. De onderzoekers van SCP en CentERdata testen eerst uit of smartphones geschikt zijn voor het TBO. Een belangrijke vraag is natuurlijk of de smartphonedata wel vergelijkbaar zijn met de dagboekdata. De dagboekaanpak is daarom vrijwel één op één vertaald naar de smartphone. Een andere vraag is of onervaren deelnemers die een smartphone in bruikleen hebben, meer moeite hebben met invullen dan ervaren gebruikers op hun eigen smartphone. Want voor een representatief onderzoek moeten ook mensen zonder eigen smartphone mee kunnen doen. Tot slot vallen ook GPS-data onder de test. Fernee: "Natuurlijk speelt hier een privacy issue, maar dat is in dit onderzoek beter afgedekt dan bij menig commerciële App."

Nathalie Sonck is net als Fernee vanuit het SCP betrokken bij het experiment. "Ik ben vooral geïnteresseerd in het gebruik van *social media*; de smartphone biedt goede mogelijkheden om daarop verder te experimenteren", zegt Sonck. In een van de volgende experimenten zullen de onderzoekers kijken of ze ook het bel- en sms gedrag kunnen uitlezen. Helemaal interessant is als de deelnemers ook de ervaringen of gevoelens die ze bij hun bezigheden hebben, kunnen intikken op de smartphone. Dergelijke studies worden bijvoorbeeld al in Groot Brittannië gedaan onder de naam 'Mappiness'-onderzoek of 'Experience sampling'. Sonck: "Maar alles op zijn tijd: eerst de geplande experimenten afmaken!"

www.scp.nl/Organisatie/Onderzoeksgroepen/Tijd_Media_Cultuur/Tijdsbesteding

Invulscherm van de smartphone

Bas Savenije, directeur van de Koninklijke Bibliotheek:

“We gaan binnenkort niet het boek, maar de digitale versie opslaan”

Heeft het papieren boek nog toekomst? Alleen als het gedigitaliseerd is, vindt ’s lands bibliothecaris. Een gesprek over e-books, copyrights en open access. *Warna Oosterbaan*

Bas Savenije, directeur van de KB *foto Ilya van Marle*

“In 2025 ongeveer”. Bas Savenije hoeft er niet lang over na te denken. Ik had hem gevraagd wanneer de KB zijn ambitieuze doel bereikt. Alle boeken, kranten en tijdschriften die sinds 1470 zijn verschenen worden gedigitaliseerd. Dat staat te lezen in de recente beleidsnota *De toekomst van de KB is digitaal*.

Waarom moet dat eigenlijk, dat digitaliseren?

“Wat gedigitaliseerd is, wordt meer en beter gebruikt. Daar komt bij dat de tendens is: alles wat niet digitaal is, bestaat niet. Dat wordt sterker naarmate er meer mensen in de digitale wereld zijn opgegroeid.
Gebruik is onze missie en daarom is digitaliseren zo belangrijk. Bewaren is ook belangrijk, maar je bewaart ergens voor. Voor de toekomst? Tsja. Dat zeg je nu en over tien jaar ook. Je bewaart voor nu!”

Waar hangt het vanaf of dit megaproject gaat lukken?

“Van de financiering, van mogelijkheden tot samenwerking, van kwesties rond het auteursrecht, van het tempo waarin materiaal digitaal beschikbaar komt. Onze inspanningen zijn er nu op gericht om alles op te slaan wat op een goeie manier digitaal beschikbaar komt. Het is zonde van de moeite om boeken die worden uitgegeven weer te digitaliseren als er een digitale versie beschikbaar is. We zijn al aan het praten met de uitgeversbond om de aanlevering te versimpelen. En bij de kranten slaan we de digitale versie ook al op. Dat is oneindig veel gemakkelijker dan achteraf inscannen. We mogen ze meestal niet op het net zetten, maar voor gebruik in de bibliotheek, *on site*, is het voor ons en voor de bezoekers een uitkomst.”

Dus binnenkort slaat u van een boek niet meer de papieren versie op, maar wel de digitale?

“Ja, en daar willen we zo snel mogelijk naar toe. Ik denk dat de ontwikkeling gaat verlopen via de *e-books*. Nu is het nog zo dat het papieren boek het primaat heeft in de hoofden van de mensen en dat de digitale versie daarvan een afgeleide is. Maar als je een e-book bekijkt, moet je vaststellen dat het al een equivalent is. Als die e-books gaan aanslaan, zul je nog wel gedrukte boeken houden, maar dat zal geleidelijk een vorm van *printing on demand* worden: als je het wilt, krijg je een gedrukt exemplaar. Maar dan wordt het natuurlijk hartstikke leuk om te kunnen zeggen: doe mij maar een hard kaft, of grote letters, of geïllustreerd en in kleur, of juist niet en in dundruk. Gevolg: het gedrukte exemplaar wordt een afgeleide versie en de digitale versie krijgt het primaat. En voor ons wordt het al snel ondoenlijk om al die gedrukte varianten te bewaren. Voor de mensen die geïnteresseerd zijn in de geschiedenis van het boek en in druktechnieken zullen we overigens nog wel papieren boeken blijven opslaan.”

INTERVIEW

Maar beschikbaar maken, via het net bijvoorbeeld, wordt nog steeds bemoeilijkt door de auteursrechten.

“Zeker. Het zou verschrikkelijk helpen als zich al een gangbaar model voor de e-books zou hebben uitgekristalliseerd. Een model waarin je kunt kiezen tussen bijvoorbeeld downloaden of dat je er on-line bij kunt en daar wat voor betaalt - op een Spotify-achtige manier - dat is er nog niet. Als je het wel zou hebben, zou je je kunnen voorstellen dat je op die manier ook kijkt naar de oude boeken waar nog wel keurig rechten op zitten, en die we al gedigitaliseerd hebben. Want dat lijkt dan wel veel op elkaar en dan kun je wellicht in dezelfde lijn relatief eenvoudig afspraken maken over ouder materiaal. Dat zal trouwens niet gratis beschikbaar komen, verwacht ik.”

Drs. J.S.M. Savenije

Drs. J.S.M. Savenije (Den Haag, 6 augustus 1947) studeerde wijsbegeerte (afstudeerrichting: logica) aan de Universiteit Utrecht. Hij is Algemeen directeur van de Koninklijke Bibliotheek sinds 1 juni 2009. Daarvoor was hij onder meer 15 jaar lang Bibliothecaris van de Universiteit Utrecht.

Er zijn wel veel klachten over de lopende digitaliseringsprojecten in Nederland. Het is te onoverzichtelijk, het verloopt ongecoördineerd en de scans vertonen veel fouten. Moet de KB hier niet regelend optreden?

“Het goede nieuws is natuurlijk dat er al zoveel is dat er over gemopperd kan worden. Iedereen is tien jaar geleden enthousiast aan de gang gegaan, vaak op zijn eigen manier. Maar inderdaad, het wordt wel tijd dat het gaat convergeren. We hebben daarom samen met de universiteitsbibliotheken van Amsterdam en Leiden een plan gemaakt om tot een standaard te komen en vast te stellen waar de prioriteiten liggen. En wat de scans betreft: we zijn voortdurend aan het werk om de ocr-technieken te verbeteren, maar perfect zal het nooit worden. Het is een hulpmiddel.”

U heeft zich altijd een groot voorstander betoond van open access, het beginsel dat wetenschappelijke artikelen vrij beschikbaar moeten zijn op internet. Hoe staat het daarmee?

“Ik zie dat de bekendheid van het fenomeen toeneemt, dat het aantal repositories (verzamelingen van publicaties op de website van een universiteit – *wo*) toeneemt en dat er steeds meer artikelen in komen. Ook komen er steeds meer tijdschriften die open access zijn. Maar bij de onderzoekers leeft het veel minder en bij de bestuurders van universiteiten gaat het ook nog maar langzaam. Daar staat tegenover dat NWO een paar fantastische stappen heeft gedaan, ze propageren open acces luid en duidelijk en steunen open acces tijdschriften.”

Toch zit er nog niet veel schot in.

“Nee. Zolang onderzoekers afhankelijk zijn van een beoordelingssysteem waarin ze beoordeeld worden op publicaties in tijdschriften met een hoge impact factor, en zolang tijdschriften met een hoge impact factor meestal niet in het publieke domein zitten, komt er weinig beweging in. Bovendien zie je dat de gevestigde wetenschappelijke uitgevers strenger aan het worden zijn, ze zijn toch wel bang voor die repositories. Want stel: je zoekt in een wereldwijde catalogus als WorldCat naar een artikel, en je ziet dat het een Elsevier-tijdschrift is. Maar op de site van Elsevier krijg je *access denied*, of betaal 30 dollar. Als dan op de site van WorldCat ook een knop is waarmee je van datzelfde artikel de finale auteursversie uit een repository kunt halen, dan zou dat ertoe kunnen leiden dat de universiteiten hun abonnementen gaan opzeggen. En daarom zie je nu dat de uitgevers steeds minder geneigd zijn toestemming te geven voor opname in een repository.”

Waar moet je het dan van hebben?

“Uiteindelijk van de financiers en de universiteitsbestuurders. Ze kunnen best wat meer druk op de uitgevers zetten. Die business gaat nog steeds heel goed en de uitgevers zouden best eens wat meer lef kunnen tonen bij het verkennen van nieuwe modellen. De wetenschappers blijven toch wel kiezen voor hun toptijdschriften.”

OPROEP

Wie wint de Nederlandse Dataprijs 2012?

Bent of kent u een onderzoeker of onderzoeksgroep die een bijzondere bijdrage levert aan de wetenschap, juist door de noeste arbeid op het gebied van data-archivering? Nomineer hem of haar voor de Nederlandse Dataprijs!

De Nederlandse Dataprijs geeft waardering aan de vaak slecht zichtbare onderzoekers die de niet altijd dankbare arbeid verrichten om data bij elkaar te brengen, te documenteren en toegankelijk te maken.
Er is een prijs voor de **humaniora en sociale wetenschappen** en er is een prijs voor de **exacte en technische wetenschappen**.
Uit de voorgedragen datasets worden voor beide prijzen drie genomineerden gekozen.
De prijsuitreiking vindt in het najaar plaats.
Denkt u kans te maken op de prijs, of kent u een collega-onderzoeker die met de prijs in het zonnetje kan worden gezet? Op **www.dans.knaw.nl** staat meer informatie over de prijs, de procedure en de jury’s. U kunt ook contact opnemen met Heiko Tjalsma (DANS), secretaris van de Dataprijs voor de humaniora en sociale wetenschappen via heiko.tjalsma@dans.knaw.nl of met Jeroen Rombouts (3TU.Datacentrum), secretaris van de dataprijs voor de exacte en technische wetenschappen via J.P.Rombouts@tudelft.nl.
Vergeet niet om **voor 1 juli** uw nominatie door te geven!

Wat is het verschil tussen een goede en een slechte roman? Misschien dat de computer er ooit achter komt. In de alfa- en gamma-wetenschappen valt er digitaal nog veel te ontdekken.
Warna Oosterbaan

Prof. Sally Wyatt leidt de e-humanities groep van de KNAW

De humanities gaan nu ook *e*

FOCUS

Het voordeel van die *e* in e-humanities is dat hij zoveel kan betekenen: *electronic, enhanced, experimental*. Al die termen komen aan bod als je met Sally Wyatt praat over dit vakgebied. Wyatt (1959), van oorsprong Canadese, woont sinds een jaar of tien in de lage landen en spreekt vlekkeloos Nederlands. Ze is van opleiding econoom en een expert op het gebied van de maatschappelijke gevolgen van technologische veranderingen. Onder haar leiding is een jaar geleden het e-humanities programma van de KNAW van start gegaan.

Maar wat *zijn* e-humanities?

We spreken Wyatt in het opvallende gebouw aan de rand van Amsterdam waar ook het Meertens-instituut is gehuisvest. Ooit was dit modernistische beton het hoofdkantoor van Coca-Cola. Nu staat in de hal vlak bij de entree als een kolossaal relikwie een houten bureau met talloze laatjes, deurtjes en aflegplanken. Het is *Het* bureau, waaraan P.J. Meertens, de vroegere directeur van het Instituut voor Dialectologie, Volks- en Naamkunde zijn werk deed, en dat vereeuwigd werd in de befaamde romancyclus van J.J. Voskuil.

Patronen

De e-humanities behoren tot een heel andere wereld. “In het algemeen bedoelen we daarmee het gebruik van digitale *tools* en technieken in de alfa- en gammawetenschappen”, zegt Wyatt. “En eigenlijk zijn er twee richtingen. Je kunt denken aan het gebruik van digitale technieken om data te analyseren, om patronen zichtbaar te maken in grote hoeveelheden onderzoeksgegevens. Een andere richting is e- humanities als een onderwerp binnen de wetenschapscultuur: hoe maken onderzoekers gebruik van digitale *tools*? Hoe beïnvloeden internet, e-mail en de tekstverwerker de wetenschapsbeoefening?”

Wyatt en haar staf hebben van de KNAW vijf jaar de tijd gekregen. “Ons doel is de samenwerking tussen verschillende instituten en universiteiten te bevorderen, en om van elkaar te leren.”

Want hoe gewoon de computer in het leven van de modale historicus, socioloog, taalkundige of musicoloog inmiddels ook is, dat je er meer mee kunt dan een artikel schrijven of e-mailen is bij hen veel minder bekend. Dat literatuurwetenschappers er de authenticiteit van een manuscript – heeft Shakespeare dit toneelstuk werkelijk geschreven? – mee kunnen vaststellen en dat archeologen met behulp van digitale technieken een aannemelijke simulatie van een prehistorisch dorp kunnen maken is inmiddels al bewezen. Maar dat je verborgen structuren en patronen uit historische bronnen kunt lichten, dat je onverwachte samenhangen en ontwikkelingen in liedteksten kunt zien en dat je wellicht ooit kunt aantonen dat het verschil tussen een goede en slechte roman niet alléén een kwestie van slechte kritiek is – dat zijn óók beloften van de e-humanities.

Volksliedjes

Voorlopig ligt binnen het project van de KNAW het accent op drie projecten waarbinnen steeds drie onderzoekers zijn aangesteld. Het eerste heet *Tunes & Tales*. In samenwerking met drie universiteiten, het Meertens Instituut en de Fryske Akademy wordt in een groot corpus van volksliedjes en volksverhalen gezocht naar de ontwikkeling van de vormen en de thema’s in de tijd. Zo is misschien een model te maken van de wijze waarop een orale traditie zich ontwikkelt. En, ook niet onbelangrijk, zo kan een begin worden gemaakt met de automatische classificering van het corpus – en dat is weer erg handig om de collectie toegankelijk te maken voor etnomusicologisch onderzoek.

Al even ambitieus is het project *The Riddle of Literary Quality*, waaraan de Universiteit van Amsterdam, de Fryske Akademy en het Huygens-ING meedoen. “Kan een computer zien of een roman goed of slecht is? Natuurlijk zijn sociale en culturele factoren erg belangrijk om de ontvangst van een literaire tekst te verklaren”, zegt Wyatt. “Maar we nemen aan dat formele aspecten ook een rol spelen. Het gebruik van moeilijke woorden, een gecompliceerde grammatica, het gebruik van adjectieven, etcetera. Dat kan een computer betrekkelijk gemakkelijk analyseren.”

Het derde project is ook een poging in een groot aantal bronnen structuur te zien. In CEDA_R proberen het Internationaal Instituut voor Sociale Geschiedenis in Amsterdam, data-instelling DANS en de Amsterdamse Vrije Universiteit met behulp van historische economische gegevens uit verschillende landen verbanden te zien tussen macro-economische veranderingen, de levens van individuele burgers, politieke systemen, arbeidsmarkten en welvaart. Gegevens uit Nederlandse volkstellingen zullen worden gebruikt om de fundamenten te leggen voor een ‘semantisch web’ dat een antwoord kan geven op de vraag hoe al die factoren elkaar beïnvloeden.

Voor de projecten vormen Sally Wyatt en haar medewerkers het centrale punt. In de ruime kamers waar ooit de Coca-Colabestuurders hun frisdrankstrategie bepaalden, komen nu twee keer per week de onderzoekers bij elkaar om te vergaderen, te brainstormen en gegevens uit te wisselen. Want hoe *e* de wetenschap inmiddels ook is, zonder persoonlijke contacten kun je het nog steeds moeilijk stellen.

http://ehumanities.nl

CLARIN krijgt ERIC-status

Taaltechnologieproject CLARIN is erkend als European Research Infrastructure Consortium (ERIC), een rechtspersoon voor onderzoeksinfrastructuren. EC-voorzitter Barroso zal de CLARIN ERIC binnenkort ondertekenen. Nederland en acht andere landen en organisaties zijn vooralsnog de stichtende leden van de CLARIN ERIC.

Binnenkort starten ook de projecten die gehonoreerd zijn binnen de derde oproep van CLARIN-NL. Geïnteresseerden mochten voorstellen indienen om bestaande data en toepassingen uit de geesteswetenschappen naar CLARIN-standaarden om te zetten, specifieke onderzoekers werden uitgenodigd een voorstel in te dienen om de spreiding over de verschillende geesteswetenschappen te garanderen. Vierentwintig onderzoekers hebben een voorstel ingediend. In totaal zijn dertien projecten gehonoreerd. (ER)

Wurdboek in Taalbank INL

Het volledige *Wurdboek fan de Fryske Taal* (WFT) is opgenomen in de Geïntegreerde Taalbank van het Instituut voor Nederlandse Lexicologie (INL). De benodigde aanpassingen en integratie in de Taalbank zijn financieel mogelijk gemaakt door CLARIN-NL. De Fryske Akademy begon in 1938 aan het Wurdboek, in 2011 is het vijfentwintigste en laatste deel gepresenteerd. Het WFT bevat zo’n 115.000 lemma’s met hun betekenis, uitspraak, dialectische varianten, vervoegingen, zegswijzen en etymologie. De nieuwste woorden in het overzichtswerk stammen uit 1975 – het woord kompjûter staat er nog net in. (ER)

http://gtb.inl.nl

SINDS KORT BESCHIKBAAR

Het overzicht toont een aantal databestanden die recent voor onderzoekers beschikbaar zijn gekomen bij CBS, CentERdata, Huygens ING (Huygens Instituut en Instituut voor Nederlandse Geschiedenis) en DANS.

Centraal Bureau voor de Statistiek

Een volledig overzicht van de CBS-bestanden staat op www.cbs.nl/microdata

- Waarde onroerende zaken, 2009-2010
- Productiestatistieken: Autohandel, 2009; Bouwnijverheid, 2009; Commerciële diensten, 2009; Delfstoffenwinning 2009; Detailhandel, 2009; Energie, 2009; Groothandel, 2009; Industrie, 2009; Transport, 2009
- Algemene Nabestaandenwet, 2001-2006
- Bijstandsuitkeringenstatistiek, Registraties, 2001-2006
- Gemeentelijke Basisadministratie Personen, 1995-2011
- Mobiliteitsonderzoek Nederland, 2009
- Integrale veiligheidsmonitor, 2010
- ICT huishoudens en personen, 2008
- Onderzoek verplaatsingsgedrag in Nederland, 2010

CentERdata LISS Data Archive

De LISS-bestanden zijn kosteloos beschikbaar via www.lissdata.nl/dataarchive

Studies LISS panel

- Validating the Dutch SF-6D and EQ-5D Using Pairwise Comparisons and Best-Worst Scaling (Jonker, M., Donkers, B., Bekker-Grob, E. De), april 2011
- Religion and Ethnicity - Wave 4 (CentERdata), januari 2011/februari 2011
- Personality - Wave 4 (CentERdata), mei 2011/juni 2011
- The public’s opinion on the control of terrorism: Attitudes and willingness-to-pay - Wave 2 (Wilsem, J. v., Woude, M. v.d.), maart 2011
- The Impact of Style and Rhetoric on the Perception of Right-Wing Populist Leaders (Bos, L.), februari 2011
- Social Integration and Leisure - Wave 4 (CentERdata), februari 2011/maart 2011
- Civic Participation, (Ingen, E. J. v.), mei 2011
- Alcohol Use and Coping with Stress - Wave 1, (Crutzen, R.), januari 2011
- Alcohol Use and Coping with Stress - Wave 2, (Crutzen, R.), april 2011
- Telephone use and regional elections, (CentERdata), maart 2011

Studies Immigrant panel

- Religion and Ethnicity - Wave 1 (CentERdata), januari 2011
- Personality - Wave 1 (CentERdata), mei 2011
- Action Control Scale (ACS-90) (Chasiotis, A., Bender, M., Vijver, F. van de), juni 2011

Huygens ING

De bestanden staan op www.historici.nl

- Weensche gezantschapsberichten van 1670 tot 1720 (2 delen)
- Holland bestuurd. Teksten over het bestuur van het graafschap Holland 1299-1567
- Bronontsluiting voor historisch onderzoek
- Rijmkroniek van Holland (366-1305)
- Waalse kerken 1601-1697
- Overzicht van de door bronnenpublicatie aan te vullen leemten der Nederlandse geschiedkennis
- Dagboekaantekeningen van vice-admiraal F. Pinke, commandant zeemacht in Nederlands-Indië 1914-1916, 1 deel
- Handel op de Oostzee 1122-1499, 6 delen
- De Vroedschap van Amsterdam, 1578-1795, 2 delen
- Relazioni Veneziane. Venetiaansche berichten over de Vereenigde Nederlanden van 1600-1795
- Dagboek Egbert Alting 1533-1594
- De Nederlandse kerk in Londen 1569-1585

DANS EASY

De bestanden zijn kosteloos beschikbaar via http://easy.dans.knaw.nl

Sociale Wetenschappen

- Data - Voer voor psychologen, archivering, beschikbaarstelling en hergebruik van onderzoeksdata in de psychologie 2010 (C. Voorbrood en H. van Luijn - DANS)
- Onderzoek Verplaatsingen in Nederland 2010 – OVIN (CBS – Rijkswaterstaat)

Ruimtelijke Wetenschappen

- TOP10NL 2010 (Kadaster)
- Bestand Bodemgebruik 2008 (CBS – Kadaster)

Online databank StatLine wordt dagelijks zevenduizend keer bezocht

De toppers van de statistiek

Het Centraal Bureau voor de Statistiek (CBS) heeft een dataschat van bijna 50 gigabyte op het web gedeponeerd. StatLine heet die schat, en zevenduizend keer per dag wordt er in gegraven. Het gaat snel, gemakkelijk en het is gratis. *Ronald van der Bie*

StatLine is de *online* databank van het CBS. Hij biedt statistische informatie in de vorm van tabellen en grafieken over vele maatschappelijke en economische onderwerpen. StatLine is een echte *visithit*. Het CBS registreerde in de maand november van vorig jaar 225 duizend bezoeken aan de databank. Tijdens die bezoeken werden er bijna 6 miljoen pagina's bekeken.

De tabellen van StatLine vormen een belangrijk deel van de ruim drieduizend publicaties die het CBS jaarlijks uitbrengt: persberichten, conjunctuurberichten, boeken, elektronische publicaties, visualisaties en de Statlinetabellen dus. De StatLinedata zijn gevat in bijna 2500 Nederlandstalige en ruim 300 Engelstalige tabellen en gerubriceerd naar thema. De databank is interactief: de gebruiker kan de gewenste gegevens zelf bij elkaar zoeken en een tabel op maat samenstellen. StatLine is niet uniek; ook in andere landen van de Europese Unie ontsluiten de statistische bureaus hun databanken op een dergelijke manier. Maar Statline valt wel op vanwege zijn omvang, toegankelijkheid en gebruiksgemak.

Sinds 1996

Dat elektronisch ontsluiten van zijn data doet het bureau sinds 1996. StatLine is gratis toegankelijk via de website van het CBS (www.cbs.nl/statline) en sinds juni 2010 kun je er ook met een iPhone App terecht. Het aantal bezoeken neemt jaarlijks toe. In 2008 werden 1,24 miljoen bezoeken geteld, in 2009 waren het er al 1,85 miljoen en in 2010 2,01 miljoen. In november 2011 stond de teller al op 2,2 miljoen bezoeken. In de zomermaanden en in december vertoont het bezoek een dip. Maart is traditioneel de topmaand, in 2011 steeg die maand het aantal bezoeken tot boven de 230 duizend. Dat zijn 7.500 bezoeken per dag. Het drukst is het op maandag. *Rush hour* is tussen 11 en 12 uur: StatLinetijd op kantoor. Het gemiddelde bezoek duurt ongeveer 7 minuten.

Ruim de helft van de StatLinebezoekers komt via de CBS-site binnen. Via Google.nl en Wikipedia.nl komt nog eens 10 procent bij het CBS. Van een derde van de bezoekers is de route naar de CBS-site niet bekend. Zij zijn waarschijnlijk de vaste bezoekers die op hun computer een *short cut* (url-code) hebben staan die ze gebruiken om direct naar de juiste StatLinepagina te kunnen gaan.

StatLine trekt vooral een professioneel publiek, blijkt uit een panelonderzoek. Bijna iedereen komt er voor het werk of studie. Bijna 60 procent van de gebruikers komt er minstens eenmaal per week data halen (*downloaden*) of raadplegen.

Sommige tabellen worden duizenden keren bezocht en gehaald, andere tabellen worden weinig ingezien of blijven ongelezen. De tien populairste tabellen zijn bij elkaar ruim 130 duizend keer bezocht (derde kwartaal 2011): 17 procent van alle bezoeken. De twintig meest bezochte tabellen trokken 180 duizend bezoeken, bijna een kwart van alle bezoeken, de tabel top-100 was goed voor bijna de helft van alle bezoeken.

Het meest gezocht zijn maandcijfers over prijzen (consumentenprijsindex, huizenprijzen), bevolking en huishouden (kerncijfers) en de cao-lonen (maandcijfers). Dat het bureau zuinig moet zijn op zijn data bewijst het aantal bezoeken aan niet-actuele (tijdelijk) stopgezette statistieken. Die trokken in 2010 bijvoorbeeld nog altijd 416 duizend bezoeken, bijna 12 procent van alle *visits*.

De infoservice van het CBS helpt bezoekers die hun cijfer niet kunnen vinden of achtergrondinformatie willen hebben. De *helpdesk* wordt per maand zo'n tweeduizend keer gebeld. Eén op de vijf bellers wil meer weten over het inflatiecijfer (consumentenprijsindex) of over huren.

Hebbedingetje

De StatLine-applicatie voor de iPhone lijkt vooralsnog een hebbedingetje voor een select gezelschap. Het aantal iPhone-bezoeken aan StatLine lag in de periode januari-oktober 2011 tussen 3,5 en 3,9 duizend per maand, zo'n 120 en 130 bezoeken per dag, en dit aantal neemt niet toe. De grootste groep app-gebruikers bezoekt de CBS-site niet vóór 21.00 uur, vaker nog een uur later. Afgaande op het aantal bezoeken in oktober 2011 zoeken de iPhoners vooral prijzeninformatie. Van de twintig meest bezochte tabellen bevatten zes tabellen informatie over prijzen: inflatie, huizenprijzen en tarieven van gas en elektriciteit.

Het CBS verspreidt zijn output ook via verschillende *social media*. Deze activiteiten passen in de *online* strategie van het bureau, dat daarmee het bereik van zijn cijfers verder hoopt te vergroten. Op YouTube (*youtube.com/statistiekcbs*) heeft het bureau een aantal korte filmpjes geplaatst met uitleg over statistische begrippen, statistieken en over het gebruik van StatLine. Ook op Twitter is het CBS actief. De ruim 4,3 duizend volgers van het bureau ontvangen nieuwsberichten en statistische weetjes.

Het CBS heeft grootse plannen met het ontsluiten van data via een webservice waarbij het voor de dataopbouw gebruik gaat maken van Statistical Data and Metadata Exchange (SDMX). Dat maakt het bijvoorbeeld straks voor iedere gebruiker buiten het CBS mogelijk om een eigen applicatie, visualisatie of website te vullen met CBS-data en deze *up to date* te houden, zonder tussenkomst van mensenhanden.

http://statline.cbs.nl/statweb/

Beeld Steamwork Graphics

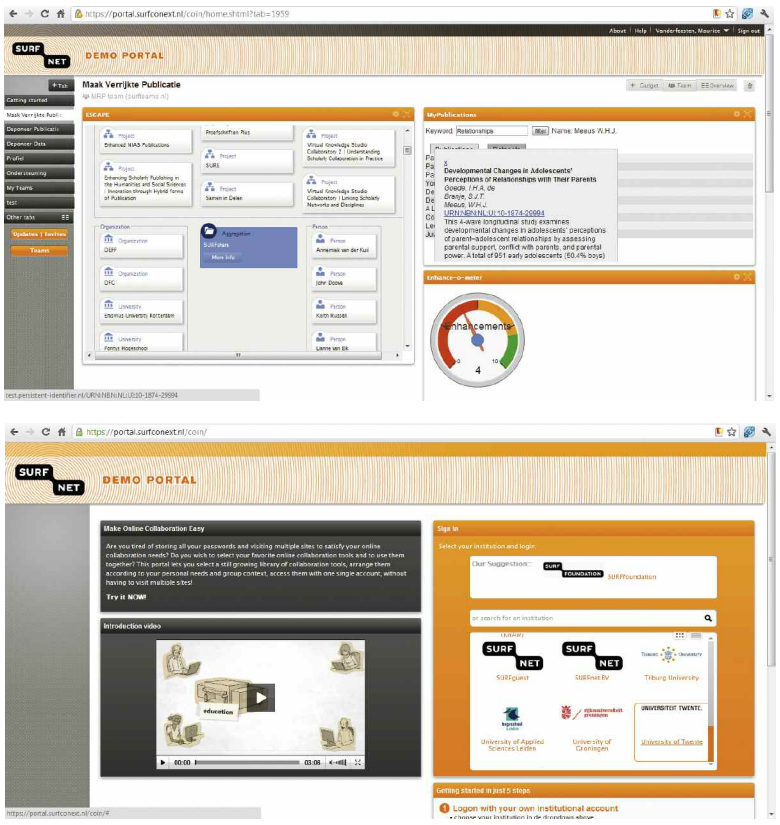
Mijn portaal voor mijn verantwoord onderzoek

Je kunt er je data delen, je publicaties verrijken en je onderzoeksresultaten doorsluizen naar anderen. Dat alles en nog veel meer kan in My research portal van SURF.
Inge Angevaare

“We willen onderzoekers enthousiast maken om meer te doen met hun publicaties en onderzoeksdata”, zegt Martin Feijen van SURF-foundation over ‘My research portal’, een applicatie die SURF heeft ontwikkeld in het kader van het SURF-share project. My research portal is een modulair product waarin instellingen allerhande hulpprogramma’s voor hun onderzoekers gemakkelijk toegankelijk maken. Zo kunnen wetenschappers hun publicaties en datasets bij elkaar brengen, er zijn tools om verrijkte publicaties te maken, tools om data te archiveren bij DANS of in een andere repository, tools om gegevens over onderzoek door te sluizen naar NARCIS, etcetera. Een deel van die hulpprogramma’s is ontwikkeld binnen SURF-share zelf; in My research portal worden ze handzaam bij elkaar gezet. Maar instellingen kunnen ook hun eigen applicaties toevoegen.

Veel eerder

E-data vroeg de Utrechtse socioloog Richard Zijdeman wat hij van het platform vindt. Zijdeman is één van de twaalf wetenschappers die betrokken is bij het testen van de portal.



Screenshot van ‘My research portal’ *bron SURF*

“In de eerste plaats is deelname aan dit project een goede gelegenheid om mijn datasets openbaar te maken”, zegt Zijdeman. “Dat had ik natuurlijk al veel eerder moeten doen, maar het komt er vaak niet van. De druk om te publiceren is groot. Goed zorgen voor je datasets en de benodigde metadata bijleveren, schiet er vaak bij in.” Dat het deponeren van datasets belangrijk is, heeft Zijdeman niettemin zelf ondervonden. Hij stuitte op onderzoek van een collega uit de jaren zeventig dat nu prachtig vergelijkingsmateriaal had opgeleverd voor zijn eigen werk. “Maar de datasets waren er niet meer. Dat was erg jammer.”
Zijdeman is enthousiast over de nieuwe mogelijkheden die verrijkte publicaties bieden. “Je kunt nu ook materiaal publiceren dat net niet belangrijk genoeg was om het artikel te halen, maar wat voor collega’s wel degelijk interessant is. En je kunt veel links leggen naar andere artikelen, ook in minder bekende tijdschriften.”

Steeds bijgewerkt

Het duurzaam toegankelijk houden van die datasets is nog wel een hele uitdaging, zo bleek recentelijk uit onderzoek dat de Koninklijke Bibliotheek en DANS samen uitvoerden. De verrijkte publicaties uit het SURF-share project bleken nauwelijks bestand tegen de tijd. Het ontbrak vooral aan goede metadata en aan ‘persistent identifiers’ – web-links die steeds worden bijgewerkt als digitale objecten verhuisd worden naar andere weblocaties.
Voor de wetenschapper is het invullen van de metadatavelden en het zoeken naar persistent identifiers allemaal extra werk waar momenteel weinig tegenover staat. Niettemin zegt Zijdeman: “Als je doel echt is om wetenschap te bedrijven, dan neem je die verantwoordelijkheid.” Om er meteen bij te vertellen dat hij het zelf ook niet altijd doet. “Maar als verrijkte publicaties de norm worden”, voegt Zijdeman eraan toe, “dan worden wetenschappelijke credits ook afhankelijk van goede datasets, en dat zou een enorme stimulans zijn voor onderzoekers om hun data goed toegankelijk te maken voor anderen.” Zijdeman wil de portal ook gebruiken in de cyclus van zijn onderzoek. “Door de link met wetenschapsportal NARCIS krijg je overzicht over wat je zelf hebt gedaan, wat anderen aan het doen zijn, en waar de hiaten zitten die nog bestudeerd moeten worden.”
Op dit moment is My research portal nog een demo-versie. SURF zelf zal het product niet gaan onderhouden als dienst, want de opdracht van SURF gaat niet verder dan aanjagen en ontwikkelen. SURF hoopt dat de instellingen het product zullen omarmen en er hun eigen versies van zullen gaan aanbieden. Een landelijke dienst vanuit bijvoorbeeld DANS behoort ook tot de mogelijkheden.

Richard Zijdeman *foto Inge Angevaare*

GELEZEN

Research Infrastructures in the Digital Humanities. Science policy briefing 42, september 2011
De European Science Foundation (ESF) bracht onlangs een rapport uit over research infrastructures voor de digitale geesteswetenschappen. Het rapport onderzoekt de stand van zaken voor wat betreft digitale infrastructurele voorzieningen, en kwam tot stand na een uitgebreide consultatieronde onder Europese experts. Een belangrijke vraag die aan de orde komt, is wie verantwoordelijk is voor zulke voorzieningen. Vaak worden ze met behulp van subsidie opgezet door een (groep van) onderzoekers, maar is behoud en beheer na de subsidieperiode niet goed geregeld.
In een reeks van case studies van prominente projecten worden deze en vergelijkbare uitdagingen belicht. Als een belangrijk voordeel van digitale infrastructuren ziet het rapport dat ze interdisciplinariteit bevorderen. Een belangrijke uitdaging - en daarin verschilt de infrastructuur voor de geesteswetenschappen van die voor de ‘harde’wetenschappen - vormen de culturele en taalkundige diversiteit die voor de geesteswetenschappen essentieel zijn.
(PB)

COLUMN

Ik zou best iets voor de KB willen doen

Voor mij ligt een boek getiteld Receptenschat, onmisbare raadgever voor het huisgezin. In alfabetische volgorde bevat dit naslagwerkje, dat in 1900 is verschenen, honderden adviezen over de gezondheid en over praktische zaken in en om het huis. Hoewel dit ooit een populair naslagwerk was, is het slechts in twee openbare collecties bewaard gebleven: in Amsterdam en Tilburg.
Ik vind dat een boek als dit zeker ook in de collectie van de Koninklijke Bibliotheek (KB) thuishoort en ik zou het zo willen schenken, maar ik wil het zelf ook makkelijk kunnen blijven raadplegen.
Is daar een oplossing voor? Natuurlijk – een heel voor de hand liggende zelfs. Ik schenk mijn papieren exemplaar aan de KB en in ruil daarvoor krijg ik een nette scan retour.
In het vorige nummer van E-data stond een mooi stuk van Peter Boot over de ideale digitale bibliotheek. Boot somt daarin allerlei eisen op waaraan zo’n bibliotheek zou moeten voldoen. De ideale digitale bibliotheek bevat bijvoorbeeld van alles wat in Nederland is gedrukt of in het Nederlands is geschreven een volledige en foutloze tekst, plus een afbeelding van de oorspronkelijke pagina’s.
Ik ben het volledig met Boot eens, maar ik wil er nog iets aan toevoegen: de ideale bibliotheek moet echt interactief zijn. Sommige bibliotheken hebben wel een digitaal loket, maar daar kun je als particulier alleen iets bestellen of afhalen. Scans van een boek of artikel bijvoorbeeld.
Dat is fijn, maar het is wel eenrichtingsverkeer. Jaarlijks kom ik tientallen boeken tegen die helemaal niet of nauwelijks bewaard zijn gebleven in openbare collecties. Sommige van die boeken wil ik nog even houden, maar ik ben graag bereid er een (professionele) scan van af te staan. Of ik ruil het boek voor een scan.

Kan dat al ergens? Bij mijn weten niet. Ik heb dit de afgelopen jaren hier en daar weleens aangekaart en krijg dan te horen dat de infrastructuur ontbreekt om dergelijke digitale schenkingen te verwerken. De KB beschikt wel over een E-Depot, maar daarin kunnen alleen professionele uitgevers hun spullen kwijt. Voor scans van bijvoorbeeld duizend streekromans of bijzondere oude kinderboeken is geen plaats.
Tweerichtingsverkeer hoeft natuurlijk niet beperkt te blijven tot de uitwisseling van scans. Wie intensief gebruikmaakt van digitale bibliografieën als Picarta, komt daarin doorlopend foutjes tegen. Bestaat er een eenvoudige mogelijkheid om zo’n fout te corrigeren? Nee, nog steeds niet.
Kun je op websites van bibliotheken makkelijk informatie toevoegen in bijvoorbeeld beschrijvingen van objecten? Hier en daar kan dit, maar een en ander staat nog in de kinderschoenen.
Er staan bij Nederlandse instellingen miljoenen pagina’s uit historische kranten, tijdschriften en boeken online. De oude teksten die gelezen zijn met ocr (optische tekenherkenning) wemelen vaak van de fouten. Bestaan er al mogelijkheden voor het publiek om hier, al dan niet na registratie, verbeteringen in aan te brengen?
Niet of nauwelijks.
Ik weet heel goed dat een en ander technisch en redactioneel een uitdaging vormt, maar ik vind het een enorme gemiste kans om geen gebruik te maken van de kennis en tijd van het publiek. Een ideale digitale bibliotheek heeft diverse loketten, waar je niet alleen iets kunt halen, maar ook iets kunt brengen. Bibliotheken die niet echt interactief worden, missen de boot.

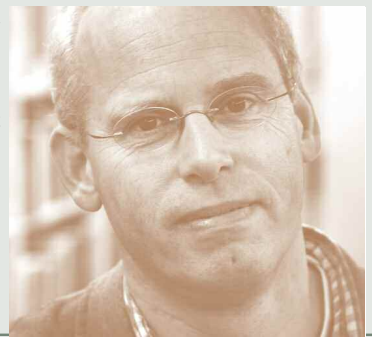


foto Leo van Velzen

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