GEHOORD & BIJGEWOOND

NCDD zet zichzelf weer terug op de kaart

Heiko Tjalsma

Het was enige tijd stil rondom de Nationale Coalitie Digitale Duurzaamheid (NCDD), maar op 11 november 2013 zette de coalitie zich in het Eye Filminstituut met haar jaarcongres weer op de Nederlandse kaart. Tijdens dit congres presenteerde de NCDD haar werkplan 2013-2018, waren er presentaties uit de verschillende aandachtsgebieden van de NCDD (archieven, cultuur, audiovisuele en wetenschappelijke collecties) en werd een samenwerkingsovereenkomst met de Britse Digital Preservation Coalition ondertekend. In het werkplan initieert de NCDD een aantal samenwerkingsprojecten op het gebied van duurzame toegankelijkheid. Deze projecten moeten oplossingen bieden voor knelpunten van de NCDD-partners.

ncdd.nl

NCDD presenteert plannen tot 2018 *foto Jacqueline van der Kort*

Onderzoekers geven verlanglijst aan Nederlab

Peter Boot

Nederlab is een omgeving die het mogelijk maakt om alle gedigitaliseerde Nederlandstalige teksten van 800 tot heden wetenschappelijk te analyseren. Maar wat zijn de vragen die onderzoekers aan dat corpus gaan stellen, en welke hulpmiddelen hebben ze nodig om die vragen te beantwoorden? Daarover vroeg Nederlab advies aan onderzoekers tijdens een workshop die 4 december jl. werd georganiseerd door Nicoline van der Sijs (Meertens en Nederlab), Karina van Dalen-Oskam (Huygens ING en UvA) en Els Stronks (UU).

Van der Sijs introduceerde Nederlab. In de huidige demoversie zijn de teksten uit de Digitale Bibliotheek voor de Nederlandse letteren (DBNL) en een jaargang van de krantencollectie van de Koninklijke Bibliotheek (KB) aanwezig. Binnenkort worden de tienduizend boeken uit *Early Dutch Books Online* toegevoegd. Het gaat dus om heel diverse collecties, waarbij de kwaliteit van de teksten en metadata zeer verschillend zijn. Erik Tjong Kim Sang toonde de tools die in eerste instantie worden opgeleverd. Het betreft daarbij onder andere (taalkundige) preprocessing (zoals opwerken van de kwaliteit van de tekst, lemmatiseren, grammaticale analyse) en hulpmiddelen voor transformeren, tellen, analyseren en visualiseren. Er komt een toegang waarmee de onderzoekers via eigen scripts de teksten kunnen analyseren (waarbij vanwege rechtenproblemen de volledige tekst meestal niet toegankelijk zal zijn). Martin Reynaert toonde als inspiratie de OpenSonar zoekinterface dat de inhoud van het Sonar-corpus met hedendaags Nederlands toegankelijk maakt voor leken.

Sommige van de vragen van de historische en letterkundige onderzoekers zouden met de nu beschikbare tools al kunnen worden beantwoord (vragen naar frequenties van woordgebruik, van termen en de context waarin ze voorkomen, naar verschillen en overeenkomsten tussen tekstversies), maar in veel gevallen bleek er nog een behoorlijke afstand te bestaan tussen de wensen van de onderzoekers en de nu beschikbare tools. Vragen naar bijvoorbeeld de waarschijnlijke auteur van een anoniem overgeleverde tekst of naar de verspreiding van literaire teksten binnen en buiten het Nederlands taalgebied vereisen andere hulpmiddelen en in sommige gevallen ook een koppeling met nog niet in Nederlab aanwezige databronnen. Voorzien aan analytische tools zijn onder meer een *n-gram viewer* en een filter voor passages in vreemde talen, maar daarnaast hebben onderzoekers ook behoefte aan meer experimentele hulpmiddelen voor bijvoorbeeld het herkennen van intertekstualiteit (plaatsen waar de ene tekst de andere citeert), van metaforen, standpunten en onderwerpen. Bij visualisatietools werd vooral gedacht aan geografische visualisaties en visualisaties in de tijd. En omdat analyse staat of valt met kwalitatief goede tekst is ook het opwaarderen van optische tekenherkenning (OCR) en het uniformeren van historische spellingen uiterst belangrijk. Ten slotte, maar dat gaat de mogelijkheden van een enkel project te boven, is het essentieel dat ons verouderde auteursrecht zo wordt aangepast dat in elk geval voor onderzoeksdoeleinden het creëren en openstellen van digitale collecties van recente tekst mogelijk wordt.

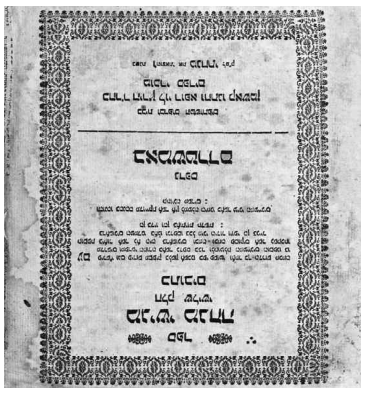
nederlab.nl

CLIN 2014: lexicologie en Big Data

Dirk Roorda

Op 17 januari jongstleden organiseerde het Instituut voor Nederlandse Lexicologie (INL) samen met de TST-Centrale de 24e editie van 'Computational Linguistics in the Netherlands' (CLIN) in Leiden. De bijeenkomst was bedoeld voor iedereen die zich bezighoudt met taal- en spraaktechnologie en haar toepassingen. CLIN telde maar liefst 15 sessies met 60 papers. Zo sprak collega Menzo Windhouwer (DANS/TLA) met Ineke Schuurman (KU-Leuven) over registers van (taal)wetenschappelijke termen waarmee onderzoeksmateriaal beschreven kan worden (ISOcat en Relcat) en presenteerden Martijn Naaijer (VU-Amsterdam) en ik een nieuw instrument om onderzoek te doen naar taalvariatie in het bijbels Hebreeuws. Andere sessies liepen uiteen van historische data via lexicologie en semantiek naar de industrie. Gezien de locatie was het bijna voor de hand liggend dat het onderwerp lexicologie extra werd belicht. Dit uitte zich onder andere in de keynote sprekers: lexicologen Patrick Hanks (University of Wolverhampton) en Dirk Geeraerts (Leuven University). De spannende vraag was 'hoe de lexicologie het beste in kan spelen op de mogelijkheden die Big Data bieden'. De corpora worden groter, raken beter geannoteerd en zijn beter doorzoekbaar geworden. Veel hedendaagse artikelen rapporteren hierover. Aan de andere kant is de lexicologie steeds beter gaan beseffen dat een oud ideaal, namelijk het beschrijven van discrete woordbetekenissen, aan richtinggevende kracht verloren heeft. Keynote speaker Hanks verwoordde het als volgt: "een woord heeft geen betekenis maar een betekenispotentieel, en dat potentieel wordt pas gerealiseerd door de patronen waarin het voorkomt". Geeraerts maakte duidelijk dat de informatie over woordgebruik, zoals we die uit de diverse corpora krijgen, meteen het scherpe onderscheid uitwist tussen vaste grond en trends in woordbetekenissen. CLIN24 werd georganiseerd door het INL en de TST-Centrale. Het INL is de plek voor iedereen die iets wil weten over woorden, hun spelling, vorm, betekenis of gebruik door de eeuwen heen. De TST-Centrale is het kennis- en distributiecentrum voor Nederlandstalige tekstverzamelingen, woordenlijsten, wetenschappelijke woordenboeken, spraakcorpora en taal- en spraaktechnologische software.

clin24.inl.nl



Titelblad Hebreeuwse bijbel *bron Rijksdienst Cultureel Erfgoed*

VERVOLG VAN PAGINA 1

Subsidie ZonMW

aan ZonMw ligt. We worden hierbij geholpen door de tijdgeest, bijvoorbeeld door de brief van Edith Schippers, minister van Volksgezondheid, Welzijn en Sport (VWS) dd. 23 oktober 2013 over een 'duurzaam informatiestelsel'.

Margreet Bloemers *foto Jeltje Waagenaar*

In deze brief geeft ze aan dat standaardisatie en het beschikbaar stellen van zorggegevens essentieel is voor een zogenaamd 'duurzaam informatiestelsel' met actuele en betrouwbare gegevens over de volksgezondheid en zorg. ZonMw gaat de naleving van de gestelde eisen controleren. We willen geen controle-instituut worden, maar het is bijvoorbeeld mogelijk om een deel van het onderzoeksbudget achter te houden totdat aan alle datamanagement-eisen is voldaan. Of we belonen onderzoekers voor goed datamanagement. Hier wordt nog volop over nagedacht. We leren graag van die wetenschapsgebieden waar het delen van data al langer gemeengoed is."

zonmw.nl/ttd

KORT

Hoe lang moeten ruwe data bewaard?

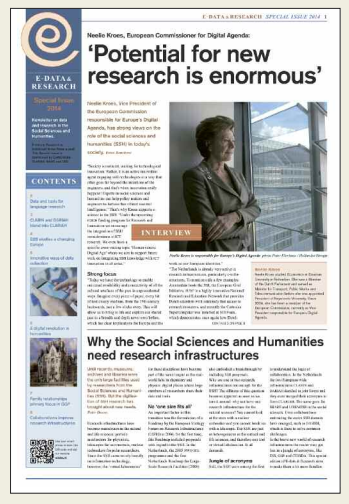
Bij onderzoekers leven veel vragen over de bewaartermijn van ruwe onderzoeksdata. DANS zette de regels op een rij. Volgens de Nederlandse Gedragscode Wetenschapsbeoefening, onderdeel III, Controleerbaarheid,moet de bewaartermijn van ruwe onderzoeksdata minimaal 5 jaar zijn. Van een maximumbewaartermijn spreekt deze code niet. Voor medische data met patiëntgegevens geldt wel een maximumbewaartermijn. Deze data moeten na 15 jaar vernietigd worden. Als medische data geanonimiseerd zijn, is er geen sprake van een vernietigingstermijn. Meer daarover staat in het rapport van ZonMw, Inventory Patients Registries in the Netherlands. DANS houdt zich aan de hierboven beschreven bewaartermijnen. Binnen andere instituten of organisaties kunnen andere regelingen gelden. (*CvZ*)

dans.knaw.nl

Special Issue E-data & Research

Onlangs heeft E-data & Research een speciale editie uitgebracht, geheel in het teken van onderzoeksinfrastructuren in de sociale en geesteswetenschappen. In deze special staat onder andere een interview met Neelie Kroes, Eurocommissaris voor de digitale agenda. Abonnees van E-data & Research hebben deze special automatisch ontvangen. Wilt u deze special ook ontvangen? Meld u dan aan als abonnee van E-data & Research door een e-mail te sturen naar de redactie (edata@dans.knaw.nl) of kijk op de website van E-data. (*ER*)

edata.nl



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. **Redactie**: Inge Angevaare, Heidi Berkhout (hoofd-/eindredacteur), Ronald van der Bie, Peter Boot, Marika de Bruijne, René van Horik, Rutger Nugteren, Erica Renckens. **Redactiesecretariaat**: Lucas Pasteuning. **Aan dit nummer werkten mee**: Marja Berendsen, Valentijn Gilissen, Evert van Ginkel, Marjan Grootveld, Lex Heerma van Voss, Hella Hollander, Ineke Huysman, Mike Kestemont, Marijn Koolen, Jacqueline van der Kort, Marja de Keuning, Katalin Márton, Milo van der Pol, Dirk Roorda, Ewoud Sanders, Jetske van der Schaaf, Nicoline van der Sijs, Mari Smits, Heiko Tjalsma, Erik Tjong Kim Sang, Jos Uljee, Jeltje Waagenaar, Carolien van Zuilekom. **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. Wilt u een artikel uit E-data overnemen? Neem dan contact op met de hoofdredacteur. Overname van artikelen geschied altijd onder vermelding van de bron E-data & Research en de naam van de auteur van het artikel.