Ron Dekker (NWO) in aanloop naar Open Access Week:

‘Datamanagementplan hoort bij goede wetenschap’

Maandag 24 oktober gaat de internationale Open Access Week van start. Volgens Ron Dekker (NWO) gaat het in Nederland in rap tempo de goede kant op.

Erica Renckens

“Die data zijn niet van jou,” vermaande Dekker de onderzoekers nog in 2011 in dit blad. Sindsdien zijn onderzoekers zich meer bewust geworden van het nut en de noodzaak om data toegankelijk te maken, aldus Ron Dekker, die sinds september is gedetacheerd bij de Europese Commissie als nationaal expert op het gebied van open science. “Soms door negatieve ervaringen als datafraude en -manipulatie, waarbij de wetenschappelijke integriteit op het spel staat. Maar er zijn gelukkig ook positieve ontwikkelingen, zoals het Open Science-beleid van NWO en de NWO-pilot met een datamanagementplan (DMP) in het onderzoeksvoorstel.”

Data zijn een ‘asset’

Deze pilot is onlangs succesvol afgerond en is sinds oktober ingevoerd bij nieuwe programma's. Elke aanvraag bij NWO bevat voortaan een

Ron Dekker *foto Monique van Zeijl*

datamanagementparagraaf. Na toekenning moet de onderzoeker zo snel mogelijk een DMP opstellen. “Dat is ook in het belang van de onderzoeker,” aldus Dekker. “Het vooraf vastleggen en verzamelen van metadata die de data beschrijven, bespaart achteraf veel tijd en zorgt voor een hogere kwaliteit metadata. Het hoort in feite bij goed wetenschap doen.” Het opstellen van een DMP kost tijd, maar het hergebruik van data dat dit mogelijk maakt, bespaart ook weer tijd en geld. “Bovendien, als je weet dat jouw data beschikbaar zullen zijn voor derden, dan komt dat ook de wetenschappelijke integriteit ten goede.” Er moet nog wel een systeem van datacitatie van de grond komen, beaamt Dekker. “Data zijn een ‘asset’ – onderzoekers hebben daar veel tijd en energie in gestoken, daar wil je op z’n minst de credits voor krijgen.”

Tijd voor actie

Dekker was projectleider Open Access bij het ministerie van OCW en speelde als adviseur van staatssecretaris Sander Dekker een belangrijke rol bij de voorbereiding van het Open Wetenschapsbeleid. In 2013 kondigde de staatssecretaris aan toe te willen werken naar 100% open toegankelijke publicaties in 2024. Onder zijn voorzitterschap vervroegde de Europese Raad voor Concurrentievermogen deze deadline zelfs naar 2020. Is dat niet wat ambitieus?
Volgens Dekker is het haalbaar: “De succesvolle onderhandelingen van de Nederlandse universiteiten laten zien dat grote uitgevers bereid zijn om én toegang tot hun tijdschriften te blijven verlenen én artikelen van Nederlandse onderzoekers direct via open access beschikbaar te maken. Dan gaat het snel met de percentages. Ook de Duitse Max Planck Geselschaft pleit voor een switch naar open access-modellen voor publicaties in 2020. Er wordt al bijna 25 jaar over open access gepraat – nu is de tijd om daadwerkelijk tot actie over te gaan.”

openaccessweek.org

Corpora SoNaR en CGN voortaan gezamenlijk te doorzoeken

Niet meer verdwalen in tekst- en spraakwoud

In OpenSoNaR+ zijn het tekstcorpus SoNaR en het spraakcorpus CGN tegelijk doorzoekbaar. Er is in het systeem nog plaats voor andere corpora. *Erica Renckens*

“Wetenschappers die kijken naar taalgebruik zijn meestal geïnteresseerd in tekst én spraak,” aldus Nelleke Oostdijk, taaltechnoloog aan de Radboud Universiteit. “Neem nou voegwoorden: in spraak kom je bijna alleen ‘en’, ‘want’ en ‘dus’ tegen, maar in teksten zie je veel meer variatie met ook archaïsche voegwoorden zoals ‘desalniettemin’.” Het is slechts een van de onderzoeksvragen die gesteld kunnen worden met OpenSoNaR+, een interface waarmee twee grote dataverzamelingen gelijktijdig doorzocht kunnen worden.
In het nieuwe systeem zijn momenteel twee corpora opgenomen. SoNaR bestaat uit ruim 540 miljoen woorden aan moderne Nederlandse teksten die automatisch zijn voorzien van tags en lemma’s. De teksten komen uit alle typen media en genre, die een gebalanceerde afspiegeling vormen van de hedendaagse schrijftaal. Het Corpus Gesproken Nederlands (CGN) bestaat uit 900 uur spraakopnames, gemaakt in een groot aantal verschillende situaties. De ruim 9 miljoen woorden zijn voorzien van transcripties, lemma’s en woordsoortinformatie.

Schat aan data

Het CGN had al een eigen zoekinterface, COREX, maar deze bleek niet om te kunnen gaan met de grote hoeveelheden data uit SoNaR. “Tijdens de ontwikkeling van SoNaR was er helaas onvoldoende budget voor een exploitatie-omgeving,” herinnert Oostdijk zich. “Als je geen technische achtergrond had, verdwaalde je in al die data. Daardoor maakte maar een heel beperkte groep gebruik van deze schat aan data.”
“In een CLARIN-NL demonstratieproject is toen OpenSonar ontwikkeld, waarin het tekstcorpus voor iedereen doorzoekbaar werd, maar die interface was weer niet bruikbaar om ook spraak mee te doorzoeken,” aldus de projectleidster van OpenSoNaR+, één van de laatste nog door CLARIN-NL gefinancierde projecten. “Met OpenSoNaR+ kan het nu eindelijk allemaal. Het systeem kan zelfs ook nog andere corpora inpassen, zoals het JASMIN-corpus met spraak van jongeren, anderstaligen en senioren.”

Snoepwinkel

Oostdijk is enthousiast over de mogelijkheden van de nieuwe zoekinterface: “Je kunt natuurlijk zoeken naar woorden, maar je kunt ook aanvullende eisen stellen, zoals ‘fiets’, maar alleen als werkwoord. Bij de resultaten zie je steeds een stukje context en kun je doorklikken om het hele fragment te lezen of te beluisteren. Het is een snoepwinkel voor onderzoekers.”
“We hopen dat het gebruik van de corpora zich nu meer zal gaan verspreiden. Communicatiewetenschappers wisten het CGN al wel te vinden, maar ook SoNaR is voor hen relevant. Dat corpus bevat ook teksten van sociale media, waar mensen vaak juist schrijven in spreektaal.”

opensonarplus.science.ru.nl

Oudezijds Achterburgwal 185, Amsterdam *foto Geert Jan van Rooij*

Nieuw: ‘KNAW Humanities Cluster’

Vanaf 1 oktober 2016 vormen het Meertens Instituut, Huygens ING en het Internationaal Instituut voor Sociale Geschiedenis (IISG) het KNAW Humanities Cluster. De instituten spelen hiermee in op internationale ontwikkelingen binnen het geesteswetenschappelijk onderzoek. Doel van de samenwerking is het versterken van vernieuwend, interdisciplinair, geesteswetenschappelijk onderzoek dat belangrijke wetenschappelijke en maatschappelijke thema’s aansnijdt. Het KNAW Humanities Cluster bouwt ook aan een digitale infrastructuur om het geesteswetenschappelijk onderzoek methodologisch te ondersteunen. Datasets, digitale collecties en daarmee verbonden diensten worden in het KNAW Humanities Cluster gezamenlijk gevormd, ontwikkeld en beheerd als onderdeel van de nationale infrastructuur voor de geesteswetenschappen.

knaw.nl

KORT

Nieuw dataplatform voor de sociale wetenschappen

Donderdag 27 oktober 2016 wordt in het Centraal Museum te Utrecht het Dataplatform voor de mens- en maatschappijwetenschappen gelanceerd. Binnen dit platform werken de Nederlandse universiteiten, NWO, CBS, DANS, Rijkskennisinstellingen, overheden en bedrijven samen om één netwerk met data ten behoeve van de mens- en maatschappijwetenschappen te creëren. Tijdens dit event wordt ook de naam van het platform, werktitel Nationale Data-infrastructuur voor de Sociale Wetenschappen (NDSW), bekendgemaakt. (MW)
nwo.nl/onderzoek-en-resultaten/programmas/NDSW

Nieuwe site NWO toont grootschalige voorzieningen

Afgelopen zomer heeft de Permanente Commissie voor Grootschalige Wetenschappelijke Infrastructuur van NWO een nieuwe website gepresenteerd: onderzoeksfaciliteiten.nl. In deze online database kunnen onderzoekers, beleidsmakers en andere geïnteresseerden onderzoeksfaciliteiten vinden en mogelijkheden tot samenwerking verkennen. Alle vormen van grootschalige onderzoeksfaciliteiten uit de volle breedte van de Nederlandse wetenschap komen aan bod: van wetenschappelijke meetapparatuur tot databases en ICT-infrastructuren. (NWO)
onderzoeksfaciliteiten.nl

Beeld en Geluid ontvangt keurmerk

Het Nederlands Instituut voor Beeld en Geluid heeft als eerste nationale audiovisuele archief het Data Seal of Approval (DSA) gekregen. Hiermee is het nationale omroeparchief gecertificeerd als ‘Trusted Digital Repository’ (TDR). Beeld en Geluid heeft volgens de beoordeling van het Data Seal of Approval aangetoond dat het de digitale duurzaamheid van de audiovisuele collecties kan waarborgen. Het instituut heeft met verschillende afdelingen meerdere jaren gewerkt aan het verkrijgen van deze certificering. Het ging hierbij zowel om het vastleggen van beleid en procedures voor digitale duurzaamheid van de audiovisuele collecties, als om het daadwerkelijk inrichten en aanpassen van de systemen, de processen en de werkwijzen.

beeldengeluid.nl