Christina Elsenga / c.m.elsenga@rug.nl
Maarten Goldberg / m.goldberg@rug.nl
Ineke Visser / i.w.visser@rug.nl

Datasets beschikbaar via RUG onderzoeksdatabase

Sinds 6 december 2017 zijn datasets van RUG- en UMCG-onderzoekers zichtbaar via de RUG research portal. Net als bij publicaties kunnen onderzoekers nu een beschrijving van hun dataset(s) toevoegen aan Pure en deze daarmee tonen op de portal en hun persoonlijke pagina. Dat het tonen van datasets via de RUG onderzoeksdatabase een mooie stap vooruit is, blijkt uit het verhaal van RUG-onderzoeker Martijn Wieling.

Naar aanleiding van een presentatie van Wieling bij een *Research Meeting* van de Letterenfaculteit afgelopen november, zijn we benieuwd naar zijn ervaringen met het delen van data. Wanneer we hem hiernaar vragen, vertelt Wieling over de aanleiding van zijn zoektocht naar data: 'Omdat een van mijn masterstudenten graag een onderzoek wilde reproduceren, was ik benieuwd naar de stand van zaken binnen de computationele taalkunde wat betreft het delen van data en vooral van software'. Wieling vertelt dat binnen dit vakgebied veel software wordt ontwikkeld door onderzoekers zelf. Hij legt uit dat met deze software vaak al bestaande datasets worden onderzocht, zoals datasets afkomstig van het *Linguistic Data Consortium*. 'Voor het reproduceren van onderzoek binnen de computationele taalkunde is de ontwikkelde software doorgaans van groter belang dan de geanalyseerde data', aldus Wieling.

Beschikbaarheid van data

Wieling onderzocht samen met twee collega's de beschikbaarheid van data binnen de computationele taalkunde door de gebruikte datasets en software op te vragen van alle artikelen die in 2011 en 2016 gepresenteerd (en gepubliceerd) zijn bij het belangrijkste congres in het vakgebied, de *Annual Meeting of the Association for Computational Linguistics* (ACL). Veel datasets en software waren niet online te vinden, dus mailden ze de auteurs. Helaas kwamen lang niet alle aangeschreven auteurs met hun dataset en/of software over de brug. Soms omdat ze dit niet wilden, maar soms ook omdat ze dit niet (meer) konden. In veel gevallen kregen Wieling en zijn collega's zelfs geen enkele reactie. Wel bleek bij zowel de datasets als de software dat deze van artikelen uit 2016 beter te vinden of achterhalen waren dan van artikelen uit 2011.

Uiteindelijk is het bij het merendeel van de artikelen wel gelukt de datasets te verkrijgen. Wat betreft de software bleek dat een stuk lastiger, maar uiteindelijk lukte het om bij meer dan de helft van de artikelen uit 2016 de software te verkrijgen. Over de precieze bevindingen schrijven Wieling en zijn collega's een artikel dat ingediend gaat worden bij een van de ACL-con-

- Linguistic Data Consortium: **www.ldc.upenn.edu**
- Website Martijn Wieling: **www.martijnwieling.nl**
- Wielings strip over zijn onderzoek naar dialecten: **www.martijnwieling.nl/strip**
- DataverseNL: **www.rug.nl/research/search/research-data-office/facilities/storage/dataverse-storage**

gressen. Een duidelijke bevinding van dit artikel zal zijn dat het niet eenvoudig is om gebruikte datasets en software te verkrijgen.

'Het is misschien niet zo vreemd dat veel software niet beschikbaar is. Het zijn vaak net persoonlijke aantekeningen met slordigheden en die maak je ook niet snel openbaar', redeneert Wieling. Volgens hem zijn er veel redenen te bedenken waarom wetenschappers niet altijd hun data en software willen delen, ook al is er geen sprake van commerciële belangen of privacyaspecten. 'Doorgaans wordt alleen opgeschoonde data gedeeld, het materiaal dat je verhaal ondersteunt', merkt hij op. 'Bovendien is het beschrijven van ruwe data, zodanig dat anderen er wat mee kunnen, een enorme klus. Voor wie doe je dat dan?'

Wielings eigen datasets

Heeft Wieling zelf al een of meerdere datasets in Pure beschreven? En waar publiceert en archiveert hij zijn data zelf? Op dit moment heeft hij zijn datasets opgeslagen in de *Mind Research Repository* en op zijn eigen website. Hij stelt al zijn materiaal beschikbaar voor hergebruik, zonder verdere voorwaarden. 'Iedereen mag mijn datasets hergebruiken, het lijkt me niet nodig om dat er nog eens apart bij te zetten' [*zie kader Licenties*]. Wat betreft Wieling zou het zeker aantrekkelijk zijn als de RUG een faciliteit biedt voor het delen en opslaan van onderzoeksdata [*zie kader DataverseNL*].

Is ondersteuning van de universiteit wat betreft het delen van data gewenst? De RUG stelt al een plek beschikbaar waar onderzoekers data kunnen delen (DataverseNL), maar Wieling denkt dat onderzoekers pas daadwerkelijk data gaan delen wanneer je het ze makkelijk maakt. Een goede werkwijze zou volgens Wieling kunnen zijn dat de RUG een plek beschikbaar stelt waar de onderzoeker zijn of haar data met een beschrijving uploadt, waarna de universiteit de rest regelt. 'Faciliteer de wetenschapper hierin zoveel mogelijk', zegt hij.

Dr. Martijn Wieling is universitair docent bij de afdeling Informatiekunde van de Faculteit der Letteren. Daarnaast is Martijn sinds 2015 lid (en vanaf eind maart 2018 vicevoorzitter) van *De Jonge Akademie*. Hij doet onderzoek naar kwalitatieve taalvariatie en –verandering. In 2017 schreef hij een strip over zijn onderzoek naar dialecten.

DataverseNL

In deze landelijke repository van DANS, lokaal beheerd door het Research Data Office, kunnen onderzoekers hun data sinds 2015 bewaren en desgewenst publiceren. DataverseNL (DVNL) is bij uitstek geschikt om de bij een artikel behorende data te publiceren. Meer en meer tijdschriften (zoals PLOS) stellen deze vereiste. Datasets in DVNL krijgen een 'persistent identifier' (een zogenaamde *Handle*) waardoor er gemakkelijk vanuit het artikel naar kan worden verwezen.

Onderzoekers kunnen ook hun dataset uploaden in Pure en aangeven of deze open access of onder voorwaarden ter beschikking wordt gesteld. Het Research Data Office van de RUG archiveert de geüploade dataset dan in DVNL en neemt bij vragen contact op met de onderzoekers.

Onderzoekers die vragen hebben over DVNL of advies willen over het archiveren van hun dataset(s) kunnen contact opnemen met het RDO via **researchdata@rug.nl**

Toestemming geven voor hergebruik (licenties)

Als je data beschikbaar stelt voor hergebruik, is het aan te raden dit er expliciet bij vermelden. De 'hergebruiker' moet er zeker van kunnen zijn en desgevraagd kunnen aantonen dat het hergebruik toegestaan is en onder welke voorwaarden.

De meest gebruikte voorwaarden zijn:

- **Attribution (CC-BY)**
  Licensees may copy, distribute, display and perform the work and make derivative works and remixes based on it only if they give the author or licensor the credits (attribution) in the manner specified by these.
- **Zero CC0**
  Besides licenses, Creative Commons also offers through CC0 a way to release material worldwide into the public domain. CC0 is a legal tool for waiving as many rights as legally possible.