SINDS KORT BESCHIKBAAR

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

CentERdata

• Denkend aan Nederland

Wat zijn de meest typerende kenmerken voor Nederland? En welke factoren dragen bij aan het gevoel van verbondenheid met Nederland? Dit heeft het Sociaal en Cultureel Planbureau (SCP) onderzocht in 2019. Het doel was om de Nederlandse identiteit in beeld te brengen. Daartoe is gebruikgemaakt van het LISS panel. Vanwege de omvang van het onderzoek werden de vragen in twee metingen afgenomen van juli tot en met september 2019. Mede op basis van deze data is het 'Sociaal en Cultureel Rapport 2019 - Denkend aan Nederland' gepubliceerd. De data zijn beschikbaar via LISS Data Archive.

Credits: SCP

lissdata.nl

Ook sinds kort beschikbaar:

Studies LISS panel

- Abidi, L.; Nilsen, P., april 2017, Implementation of alcohol prevention in healthcare in the Netherlands
- Kok, L., april 2017, Pension designs and continued working after retirement
- Portegijs, W., juni 2018, Emancipatiemonitor 2018
- CentERdata, oktober-november 2018, Social Integration and Leisure - Wave 11
- CentERdata, juni-juli 2019, Economic Situation: Income - Wave 12
- CentERdata, juli-augustus 2019, Economic Situation: Housing - Wave 12

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

DANS

• Nieuw in EASY:

Maritieme opgravingsdossiers

Sinds de inpoldering van de Wieringermeer zijn honderden scheepswrakken aangetroffen, vergaan op de voormalige Zuiderzee. De documentatie hiervan varieert van eenvoudige meldingen tot volledig uitgevoerde archeologische opgravingen. Het gaat om waardevolle en unieke brondocumentatie voor (scheeps)archeologisch onderzoek. Stichting Batavialand beheert zowel de archeologische objecten als het bijbehorende archief van de maritieme rijkscollectie namens de Rijksdienst voor het Cultureel Erfgoed. Onlangs heeft de Stichting de papieren (scheeps)archeologische opgravingsdocumentatie gedigitaliseerd. Deze bijzondere collectie wordt nu toegankelijk gemaakt via EASY. DOI: 10.17026/dans-x6z-3dnp.

Credits: DOI: 10.17026/dans-2z5-jmy2

Ook sinds kort beschikbaar:

De volgende datasets zijn open access beschikbaar via het online archiveringssysteem EASY van DANS:

- Berkel, dr. R. van (Utrecht University) (2020): Versterking methodisch werken via HRM.
DANS. DOI: 10.17026/dans-x3w-7q4b.
- Farace, dr. D. (GreyNet International) (2020): Grey Literature Resources generate and drive Awareness to the Circular Economy.
DANS. DOI: 10.17026/dans-zhz-kg3z.
- Frankena, dr. K. (Wageningen University) (2020): ROMAN, Few-Foods-Diet and ADHD in Practice.
DANS. DOI: 10.17026/dans-xn4-6pjh.
- Gregory, K.M. (Data Archive and Networked Services) (2018): Data Discovery and Reuse Practices in Research.
DANS. DOI: 10.17026/dans-xsw-kkeq.
- Heine, F.A. (Tilburg University) (2020): Using Moral Foundations in Government Communication to reduce Vaccine Hesitancy.
DANS. DOI: 10.17026/dans-xuv-vyzk.
- Jordanov, drs. M.S. (RAAP) (2020): Kasteelpark IJsselstein, gemeente IJsselstein, een archeologische opgraving.
DANS. DOI: 10.17026/dans-z33-gtvv.
- Leemans, L.H. (Radboud University) (2020): A mutualism between unattached coralline algae and seagrasses prevents overgrazing by sea turtles. Ecosystems.
DANS. DOI: 10.17026/dans-25p-82rx.
- Lutkie, T. (2019): De pot en de ketel: Nederlandse dagbladen en hun oordeel over communisme en fascisme, 1918 - 1939.
DANS. DOI: 10.17026/dans-zeq-tnzx.
- Moretta, dr. T.M. (Department of General Psychology, University of Padova) (2019): Data from problematic and non-problematic Facebook users who performed a Go/Nogo task with Facebook-related, pleasant, unpleasant, and neutral pictures and a self assessment manikin (SAM).
DANS. DOI: 10.17026/dans-zqm-d9zh.
- Nollen, drs. J.H. (Gemeente Breda) (2020): Breda Kasteelplein (AO).
DANS. DOI: 10.17026/dans-zxr-3xtd.
- Scholtens, J. (Commissariaat voor de Media) (2019): Representatie van mannen en vrouwen in Nederlandse non-fictie televisieprogramma's 2019.
DANS. DOI: 10.17026/dans-27s-4q6g.
- Sociaal en Cultureel Planbureau (SCP) (2018): Vrouwen in besluitvorming 2018 - VIB2018.
DANS. DOI: 10.17026/dans-26j-7rw8.
- Westen, dr. C.J. van (University of Twente) (2020): Landslide inventory of the 2018 monsoon rainfall in Kerala, India.
DANS. DOI: 10.17026/dans-x6c-y7x2.

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

KORT

Nieuwe directeur DANS omarmt open science

Sinds 1 april is Henk Wals directeur van DANS. Zijn visie is helder: "Waar het om draait, is de beweging richting *open science*. Naarmate onderzoeksdata en -resultaten sneller en beter gedeeld worden, versnelt de kenniscirculatie en boekt de wetenschap in een hoger tempo resultaten. In Nederland houden ruim honderd organisaties zich bezig met data, opslag, infrastructuren, etc. Hoe ordenen we dat landschap, welke afspraken maken we over het verbinden van onderzoeksgegevens en hoe voorkomen we duplicatie van services? Gelukkig zijn er initiatieven als het Nationaal Platform Open Science en de European Open Science Cloud. Samen bewegen we richting een netwerkorganisatie met goede taakverdeling, coördinatie en afspraken. DANS heeft alles in zich om een nuttig knooppunt te vormen in een netwerk van Nederlandse en Europese instellingen die bijdragen aan de data-infrastructuur. Wij zijn bereid ons aan zo'n rol te committeren. En KNAW en NWO steunen deze gedachte, is mij verzekerd." (HB)

dans.knaw.nl

Mooie resultaten met Optical Character Recognition

Teamwork verbetert OCR gotische druk

OCR is een interessante tool met vele toepassingen. Of het ook werkt voor Nederlandse gotische druk, werd tijdens de workshop *ICT with Industry* onderzocht. *Rutger van Koert*

Optical Character Recognition (OCR) staat voor optische tekenherkenning: een methode waarbij een computer door middel van patroonherkenning tekens uit een afbeelding haalt. OCR werkt over het algemeen vrij goed op modern materiaal. Helaas gaat de kwaliteit van de herkenning achteruit naarmate het materiaal ouder is. Ook bij 'vreemde' fonts, vlekken en vervuiling verslechtert de kwaliteit. Genoeg motivatie om tijdens de jaarlijkse, door het ICT Research Platform Nederland (IPN) georganiseerde workshop *ICT with Industry* afgelopen februari aan deze wetenschappelijke uitdaging te werken.

Vier subproblemen

Door het team werden vier subproblemen gedefinieerd: preprocessing inclusief voorbewerken van de scans, segmentatie van de scans op woord- of zinsniveau, herkenning (de daadwerkelijke OCR) en postprocessing, het automatisch corrigeren van fouten van de herkenning. Samen met Mirjam Cuper (KB) zorgde ik voor scans, transcripties en rekenkracht voor de machinelearning, Jerry Guo (TU Delft) probeerde diverse algoritmes uit. Visueel was de verbetering goed zichtbaar, maar de resulterende OCR-output verbeterde nauwelijks. Voor de segmentatie, het tweede subprobleem, gebruikten we ARU-net. Samen met Xue Wang (CS, Leiden University) trainde ik het systeem op het detecteren van spaties en woorden met hulp van data van de ALTO-xml van een commerciële OCR-provider. We controleerden de resultaten weer visueel, op sommige punten was er zelfs een verbetering ten opzichte van de commerciële provider. Vervolgens werd Monk door Lambert Schomaker (AI/ML RuG) ingezet om data te labelen en ging Mahya Ameryan (AI, RuG) woorden herkennen met machinelearning. 88% van de woorden bleek correct te zijn herkend, een mooie score! Als laatste namen Koen Dercksen (Radboud Universiteit) en Konstantin Todorov (ILLC, UvA) het nabewerken op zich door gebruik te maken van BERT, gefinetuned op het tekstcorpus van de Meertens Kranten (1662-1795) en aansluitend een LSTM encoder-decoder netwerk. Met het softwareplan van Adriënne Mendrik (e-Science Center) kunnen we resultaten kwantificeren en meten wat daadwerkelijk de beste opties zijn voor specifieke onderdelen.

Titel-scan van het boek 'Gelegentheyt van s Hertogen-Bosch' door Pieter Bor, geschreven in 1630. De binnenzijde van dit boek bevat teksten in gotisch schrift, OCR-technieken maken het onderzoekers makkelijker om de teksten te gebruiken. *Credits: KB*

Flinke verbetering

Via *ICT with Industry* hebben we, naast een leuke week met slimme mensen uit de wetenschap en het bedrijfsleven, mooie resultaten bereikt. Samen concluderen we trots dat het mogelijk is om de herkenning van Nederlands gotisch drukwerk flink te verbeteren. Het KNAW Humanities Cluster en de KB gaan kijken hoe deze pijplijn voor vroegmoderne druk verder kan worden ontwikkeld.

ict-research.nl/ict-with-industry

Rutger van Koert is Lead Engineer Team Images bij het KNAW Humanities Cluster.