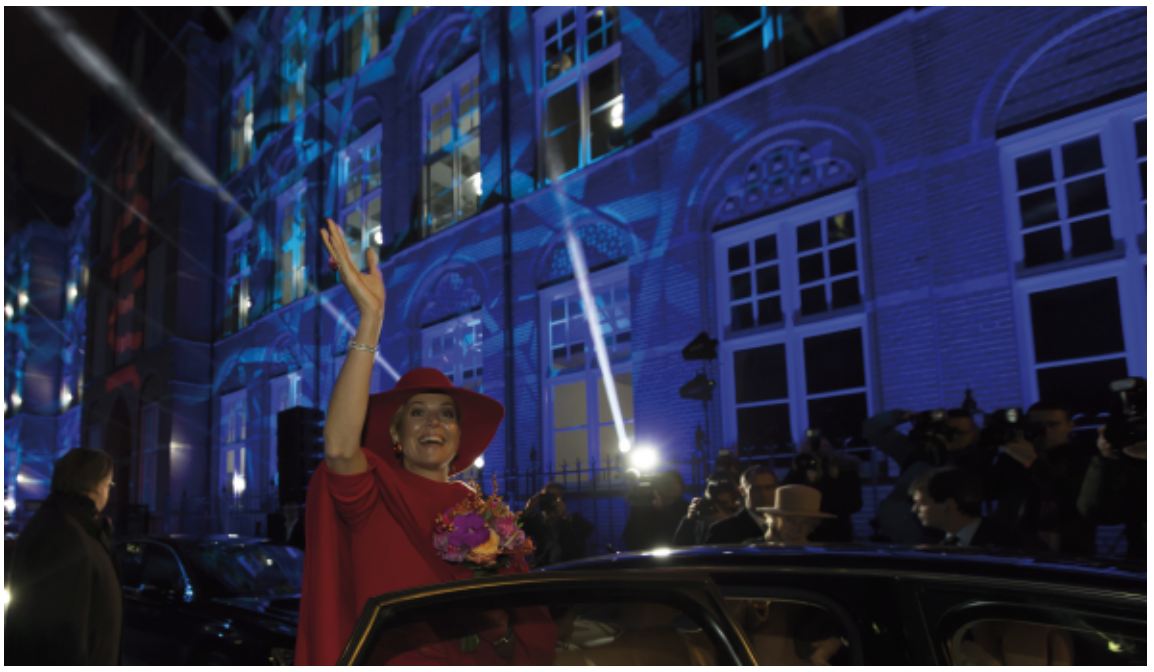
JADS brengt onderwijs, onderzoek en bedrijfsleven bijeen

Data science centrum JADS opent met lichtspektakel

Tijdens een schouwspel van lasers opende Hare Majesteit Koningin Máxima op 1 december de Jheronimus Academy of Data Science (JADS): het grootste data science centrum in Nederland.

Marika de Bruijne



Hare Majesteit Koningin Máxima vertrekt na de openingsceremonie. Op de achtergrond JADS, gevestigd in het voormalig klooster Mariënburg in 's-Hertogenbosch foto Marjo van de Peppel

JADS, een samenwerking tussen Tilburg University, Technische Universiteit Eindhoven, de provincie Noord-Brabant en de gemeente 's-Hertogenbosch, biedt als eerste in Nederland een universitaire bachelorstudie Data Science en de masteropleiding Data Science Entrepreneurship. Directeur Arjan van den Born: "JADS verschilt op drie kenmerken van de andere data science centra in Nederland. Eén verschil is de omvang. Wij hebben bijna tweeduizend studenten en zijn daarmee veel groter dan andere data centra. Het tweede verschil is dat we de opleidingen van nul af aan hebben opgebouwd. Meestal zijn het econometrie- of informatica-opleidingen die opeens data science heten. Het laatste verschil is dat we de business- en IT-werelden bij elkaar brengen, zo werkt JADS samen met het bedrijfsleven. We gebruiken altijd datasets; studenten leren de data in de praktijk toe te passen. Vaak zijn het datasets van bedrijven waar we analyses op mogen doen. Ook voeren we onderzoek naar data science in opdracht uit. Zo combineren we denken en doen."

Impact op maatschappij

"De wetenschap verandert. Tien jaar geleden was een dataset van 10.000 observaties groot, nu is dat pas het geval met 1,2 miljoen observaties. De t-waarde zegt dan niets meer, maar de basisvraag - hoe patronen toe te kennen zijn aan data - blijft. In plaats van traditionele statistiek beantwoord je de vraag met kunstmatige intelligentie, algoritmes en visualisaties", legt Van den Born uit. In de komende jaren verwacht hij dat technologieën zoals machine learning (bijvoorbeeld patroonherkenning in video en tekst), collaborative assistance (meer taken als algoritme kunnen uitvoeren) en internet-of-things (voorwerpen verbonden met het internet) zich verder zullen ontwikkelen. Van den Born: "Deze technologieën zullen op veel domeinen een verschil maken, van marketing tot onderwijs, zorg en logistiek. Om bijvoorbeeld ziekten als schizofrenie beter te herkennen bij mensen. Om data te gebruiken voor personalisatie. Niet alleen opdat Netflix je de juiste film kan laten zien, maar ook opdat de basisschool beter kan zien wat een kind nodig heeft om talenten optimaal te gebruiken."

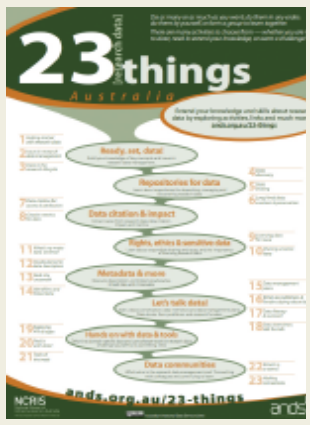
"Het gebruiken van big data moet niet doorslaan, maar verantwoord blijven. Anders gaat de maatschappij tegenkracht leveren." JADS heeft zich mede hierom aangesloten bij het consortium Responsible Data Science (RDS). "Samen willen we ervoor zorgen dat iedereen eerlijk, accuraat, vertrouwelijk en transparant omgaat met data."

jads.nl

GELEZEN

23 (research data) Things by Australian National Data Service (ANDS)
Heidi Berkhout

ANDS en zijn partners hebben hun kennis over onderzoeksdata gebundeld in 23 (research data) Things. Deze gratis online training helpt bibliothecarissen en datamanagers om (het potentieel van) onderzoeksdata steeds beter te begrijpen. De basisgedachte is dat iedereen op elk moment 23 dingen kan doen op het gebied van onderzoeksdata. 23 Things biedt een scala aan leermogelijkheden met activiteiten op drie niveaus: 'Aan de slag', 'Meer informatie' en 'Daag me uit'. De stappen zijn in willekeurige volgorde te doorlopen. Alleen als je nieuw bent in de wereld van onderzoeksdata, adviseren de auteurs om bij stap 1 te starten. Enkele voorbeelden: de stappen 4, 5 en 6 gaan over data repositories, de stappen 7 en 8 over datacitatie. Met het doorlopen van de verschillende stappen, bouw je steeds meer kennis op over (het omgaan met) onderzoeksdata.
ands.org.au/23-things



credits website ANDS

COLUMN

Datasets uit Delpher

Onlangs bezocht ik een congres van de Koninklijke Bibliotheek (KB) in Den Haag over 'Historische kranten als *big data*'. De KB kwam er met een verrassing, namelijk dat ruim 370.000 rechtenvrije kranten tot 1876 als dataset kunnen worden gedownload.

Voorheen kon je, als onderzoeker, een bepaalde titel of een reeks kranten uit een bepaalde periode opvragen, maar daar moest het nodige papierwerk voor worden ingevuld. Nu is dat een stuk makkelijker geworden: op de pagina delpher.nl/data/kranten staan 22 zipbestanden die voor iedereen te downloaden zijn. Eén zip-bestand met 17de-eeuwse kranten, tien zip-bestanden met 18de-eeuwse kranten en elf zip-bestanden met 19de-eeuwse kranten.

Er staat op die pagina ook een klein proefbestand en dat heb ik daags na het congres gedownload. Op het congres werd gemeld dat je 'complete sets kranten' kon downloaden, maar dat bleek niet helemaal te kloppen. In de zip-bestanden zit de OCR-laag van de kranten: de tekstlaag achter de afbeeldingen van de krantenpagina's. Als je de scans of pdf's wilt downloaden, moet je zelf aan de slag met de meegeleverde metadata.

Laat ik om te beginnen zeggen dat ik blij ben met deze stap. Ik ken veel mensen die intensief van Delpher gebruikmaken, maar de meesten van hen zijn vooral bezig met knippen en plakken uit deze onmisbare bron. De mogelijkheid om grotere sets te kunnen downloaden stond bij menigeen op het verlanglijstje, dus het is mooi dat daar nu een begin mee is gemaakt.

Toch voorspel ik dat de meeste onderzoekers gewoon in Delpher zullen blijven zoeken en niet in de bulkpartijen die nu worden aangeboden. Dat komt door de kwaliteit van de OCR. Letters op afbeeldingen van boeken en kranten worden automatisch 'herkend' door een tool en de kwaliteit van de OCR in Delpher – vooral van de oudere kranten – laat veel te wensen over.

Dat is geen nieuws. Sterker nog: twee jaar geleden, op het eerste KB-congres over dit onderwerp,

foto Leo van Velzen

kwam dit uitgebreid ter sprake en nu weer. Voor sommigen was het even schrikken dat er in de afgelopen twee jaar in Delpher nauwelijks iets is gedaan aan OCR-verbetering, maar het staat hoog op de agenda dus dat geeft goede hoop.

Zoals gezegd bevatten die zipsets de OCR-tekst plus metadata. In het proefbestand opende ik de eerste OCR-tekst uit 1876, want hoe jonger de tekst, hoe groter de kans op relatief goede OCR-tekst. Er stond: 'goj Â°J correspondent der KÃ¶ln. Zeitung is in het bezit IL1*,', van bot ontwerp, dat door de Russische regoe- Wl .^Â° cÂ°nferentie zal worden ingediend on dat de >U(v<,0l'mingen schetst, dio zij voor Bulgarije wenscht YvÂ°erd te zien.'

Ik ben onvoldoende technisch onderlegd om te snappen hoe je via de meegeleverde metadata de oorspronkelijke pagina kunt vinden. En dus zocht ik maar in de gewone zoekregel bij Delpher op 'bot ontwerp, dat door de Russische'. Dat leverde één artikel op, uit het Algemeen Handelsblad van 1876. Het stuk gaat onder meer over de verdeling van grondgebied in Bulgarije om de vrede te bewaren tussen Christenen en Mohammedanen. De OCR-tool heeft hiervan gemaakt: 'verdeeling tussehen christenen **ld 1 mmeaaneni A-an'.

Ik ga al die Delpher-sets zeker downloaden, maar wellicht wacht ik nog even op de volgende grote stap: de verbetering van de OCR.

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