E-DATA & RESEARCH

Jaargang 13 | nummer 2

Nieuwsbrief over data en onderzoek in de alfa- en gamma-wetenschappen.

E-data & Research verschijnt drie keer per jaar en wordt mogelijk gemaakt door: CentERdata, CLARIAH, DANS, Huygens ING, de Koninklijke Bibliotheek en het Rijksmuseum.

INHOUD

E-data wordt gratis toegezonden aan relaties van de stakeholders. Ook een uitgave ontvangen? Mail de redactie: edata@dans.knaw.nl.

Scan deze QR-code met een smartphone om de website van E-data te bezoeken.
edata.nl

Fries-Nederlandse spraakherkenner ontwikkeld

Speuren in de archieven van Omrop Fryslân

Sinds afgelopen najaar zijn de radio-archieven van Omrop Fryslân digitaal doorzoekbaar, dankzij een spraakherkenner die zowel Fries als Nederlands herkent.

Erica Renckens

Ruim 3.000 uur aan oude analoge radio-opnames van Omrop Fryslân zijn online te doorzoeken. De Friese omroep riep hiervoor de hulp in van spraak- en taaltechnologen van de Radboud Universiteit. Zij ontwikkelden binnen het project FAME een Nederlands-Friese spraakherkenner.

Code switching

"In het Fries wordt veel geswitcht met het Nederlands", vertelt projectleider Henk van den Heuvel. "Dat is erg lastig voor een spraakherkenner." Om ervaring op te doen met dit zogenaamde 'code switching', bezocht één van de projectleden, spraaktechnoloog Emre Yilmaz onder andere Zuid-Afrika, waar het Afrikaans is doorspekt met Engelse woorden. "Je kunt het probleem grofweg op twee manieren benaderen: je maakt één herkenner voor beide talen of je stelt eerst van elk woord de taal vast en herkent dat vervolgens. Uit Emres onderzoek bleek de eerste aanpak in ons geval het beste te werken." De Friese spraakherkenner heeft nu een error rate van zo'n 23 procent. "Dat is een stuk lager dan we bij aanvang van het project verwachtten. Het lijkt misschien nog steeds hoog, maar het is laag genoeg om in het archief relevante fragmenten te kunnen vinden."

Van den Heuvel verwacht dat de spraakherkenner ook buiten het project gebruikt zal worden. "Omdat het systeem is ontwikkeld voor omroepmateriaal, is het vocabulaire vrij breed. Onlangs heeft iemand het getest voor een zorgrobot en dat leek veelbelovend uit te pakken."

Doorontwikkeling

De zoekmachine is opgeleverd in drie versies voor verschillende doelgroepen: het grote publiek, journalisten en onderzoekers. "De journalisten kunnen de zoekresultaten ook downloaden. In de versie voor de onderzoekers is daarnaast ook sprekerherkenning beschikbaar. Het systeem clustert de spraak per spreker en probeert deze vervolgens te linken aan een van de 336 stemprofielen. Die zijn van mensen die regelmatig in de opnames terugkomen, zoals presentatoren."

Hoewel het project inmiddels is afgerond, zal het zoeksysteem nog worden doorontwikkeld. Van den Heuvel: "Momenteel is alleen nog het analoge archief doorzoekbaar, met materiaal tot het jaar 2000. De private partners uit het project – Omrop Frylân, Tresoar en Gridline – zullen ook nog het latere materiaal doorzoekbaar maken, dat is van oorsprong al digitaal. De spraakherkenner zelf wordt daarvoor ook bij Tresoar ondergebracht."

Wie zelf wil zoeken in het materiaal kan terecht op: ***zoeken.fame.frl***

Radio-opnames op analoge banden en taperecorder van Omrop Fryslân. Inmiddels is het mogelijk de opnames digitaal terug te luisteren en te doorzoeken. *credits Omrop Fryslân*

Rijksmuseum sluit zich aan bij Linked.Art consortium

Erfgoeddata van LOD naar LOUD

In Linked.Art werkt een internationaal consortium aan de verbetering van de bruikbaarheid van Linked Open Data voor erfgoedobjecten. Onlangs sloot ook het Rijksmuseum zich aan.

Maarten Heerlien

Afgelopen maand is het Rijksmuseum als eerste Nederlandse museum toegetreden tot het internationale consortium van 23 erfgoedinstellingen en onderzoeksinstituten uit Noord-Amerika, Europa en Azië dat uitvoering geeft aan het Linked.Art-project. Het consortium wordt geleid door Robert Sanderson, semantisch architect bij de J. Paul Getty Trust. Voor het Rijksmuseum biedt Linked.Art mooie kansen, aldus Saskia Scheltjens, Hoofd Research Services bij het Rijksmuseum: "Deze samenwerking kadert binnen de Memorandum of Understanding tussen het Rijksmuseum en het Getty. Het geeft ons de gelegenheid om verder te bouwen aan onze expertise en deze kennis maximaal te delen met anderen op een internationaal niveau."

Bruikbaarheid vergroten

Doelstelling van het Linked.Art-project is om tot een concrete, op gebruikers gerichte toepassing te komen van CIDOC-CRM, het Conceptual Reference Model voor beschrijving en uitwisseling van informatie over erfgoedobjecten van de International Council of Museums (ICOM). Hoewel CIDOC-CRM al sinds 2006 een ISO-standaard is, blijft implementatie ervan door erfgoedinstellingen achter. Chris Dijkshoorn, databeheerder bij het Rijksmuseum en betrokken bij het ontwikkelteam van het project: "CIDOC-CRM is een expressief, maar zeer theoretisch model. Data is met het model op verschillende manieren vast te leggen, wat de concrete toepassing van CIDOC-CRM hindert. Linked.Art richt zich op het bereiken van consensus over de toepassing van het model door middel van een zogeheten applicatieprofiel. Daarbij worden alleen de elementen uit CIDOC-CRM gebruikt die voor specifieke use-cases relevant zijn. Uiteindelijk vergroot dat de bruikbaarheid van cultureel erfgoed data."

Linked Open Usable Data

Linked.Art werkt dit applicatieprofiel uit volgens de principes van LOUD, Linked Open Usable Data. LOUD, een model van Sanderson, is een aanvulling op Tim Berners-Lee's vijfsterrenmodel voor Linked Open Data. Berners-Lee's model is gericht op dataproviders en minder op datagebruikers, waardoor vijfsterren datasets vaak onderbenut blijven. LOUD richt zich op ontwikkelaars, de intermediairs tussen dataproviders en datagebruikers. Om de balans in datasets tussen bruikbaarheid en rijkheid te optimaliseren, formuleerde Sanderson vijf aanvullende sterren voor bruikbare Linked Open Data en evenzoveel ontwerpprincipes om daar concrete invulling aan te geven. Een belangrijke daarvan is de consequente uitdrukking van Linked Data in JSON-LD, een gebruiksvriendelijker alternatief voor RDF-XML.

De deelname van het Rijksmuseum aan Linked.Art, gecoördineerd door de afdeling Research Services, is een eerste stap in de intensivering van de samenwerking tussen deze afdeling en de afdeling Digital van de J. Paul Getty Trust. Saskia Scheltjens neemt namens het museum zitting in de projectstuurgroep. Linked.Art heeft een initiële doorlooptijd tot 2021.

linked.art

GEHOORD & BIJGEWOOND

Hackathon voor betere onderwijskansen

Seyit Höcük

Op zaterdag 8 december vond de hackathon “Hack for Future Talent” plaats op de campus van Tilburg University. Deze Hack Marathon duurde maar liefst 12 uur en 44 ‘hackers’, verdeeld over elf teams, deden mee. Na een sfeervolle opening met onder andere wethouders van Tilburg en Eindhoven beten de deelnemers zich urenlang vast op één van de drie thema’s: spookjongeren, talentoptimalisatie en gelijke onderwijskansen. De Hackathon was vrij toegankelijk voor iedereen met affiniteit voor onderwijs en data. Elk thema viel uiteen in meerdere ‘challenges’. Dat waren doelen die vervuld moesten worden om te winnen, zoals het leveren van onverwachte inzichten, het goed visualiseren van onderwijsdata en met datagedreven oplossingen komen. Aan het einde van de dag was er onder meer een geldprijs van 1.000 euro beschikbaar voor het team dat één van deze thema’s het beste volbracht. Team Thefantasticfour, een team van PABO University, heeft de Hackathon gewonnen met hun idee voor een gelijke start voor kinderen. Op basis van bestaande data hebben ze laten zien dat leerlingen met migratieachtergrond een lager eindadvies van hun school krijgen dan je zou verwachten op basis van hun Cito-score, terwijl leerlingen zonder migratieachtergrond juist een opmerkelijk hoog eindadvies krijgen. Daarnaast kwamen ze met ideeën voor (betere) dataverzameling en het meer toegankelijk maken van bestaande open data. Ook andere teams hebben het goed gedaan. Zo eindigde team Vantage AI, een team van vier data scientists, als tweede. Hun doel was om met machine learning, zoals dimensionaliteitsreductie en clustering, vergelijkbare scholen met elkaar te verbinden, zodat een effectieve uitwisseling van informatie en best practices mogelijk zouden worden. De hackathon werd afgesloten met een gezellige borrel. De betrokken gemeenten en het ministerie van OCW gaan nu bekijken of ze sommige ideeën kunnen vertalen naar concrete nieuwe projecten.

hackforfuturetalent.nl

Het team Thefantasticfour was de winnaar van Hack for Future Talent. *credits gemeente Tilburg*

HuC LIVE! slaat een brug

Thijs van der Veen

Het KNAW Humanities Cluster (HuC) presenteerde op 12 december in het Compagnietheater zijn onderzoek, producten en plannen voor de komende jaren aan een internationaal publiek bestaande uit onderzoekers, managers, beleidsmakers en partners tijdens HuC LIVE! Het thema ‘bridging the gap’ tussen de geestes- en de bètawetenschap was gebaseerd op een citaat van antropoloog Clifford Geertz: “I think the perception of a deep gulf between science and the humanities is false.” De verschillende sprekers toonden aan dat het HuC die brug inderdaad kan slaan. Zo betoogde Antal van den Bosch, hoogleraar taal- en spraaktechnologie en directeur van het Meertens Instituut, dat we toe zijn aan een ‘cultural AI’. Kunstmatige intelligentie heeft behoefte aan ethiek en een besef van genderkwesties, diversiteit en inclusiviteit, en meertaligheid. De geesteswetenschappen kunnen dat leveren. Adina Nerghes en Marijn Koolen presenteerden er hun onderzoek en Jauco Noordzij (hoofd Product Development) ging dieper in op de relatie tussen onderzoeker en software developers. En dan was er de presentatie van de mystery guest, Dee. Deze bot presenteerde de producten van de afdeling Digitale Infrastructuur en legde uit hoe de infrastructuur in elkaar zit. Als afsluiter legde Elli Bleeker via Kahoot pittige stellingen voor aan het publiek en aan directeur Digitale Infrastructuur Gertjan Filarski en teamleider Marieke van Erp van DHLab. Op YouTube zijn de keynote en de presentatie van Dee terug te zien op het kanaal van KNAW Humanities Cluster. Alle presentaties staan als blogpost op:

huc.knaw.nl/blog

Keynote spreker Antal van den Bosch, directeur Meertens Instituut. *credits Humanities Cluster KNAW*

DANS-workshop ‘RDM in the Time of the GDPR’

Widia Mahabier

Dinsdag 11 december organiseerde DANS de workshop ‘Research Data Management (RDM) in the Time of the GDPR (General Data Protection Regulation)’. De workshop ging in op de ethische en juridische aspecten van data delen en hoe er omgegaan moet worden met vertrouwelijke informatie, nu sinds 25 mei in Europa de GDPR en in Nederland de Nederlandse vertaling hiervan, de Algemene Verordening Gegevensbescherming (AVG) van kracht is. De workshop bestond uit een aantal lezingen afgewisseld met praktische onderdelen. De deelnemers, zowel datasupporters als onderzoekers, waren erg enthousiast en vonden de workshop interessant en leerzaam. Na een welkomstwoord door Peter Doorn (directeur DANS) hield Ricarda Braukmann (programmaleider sociale wetenschappen DANS) een inleiding in RDM en het gebruik van de door CESSDA (Consortium of European Social Sciences Data Archives) ontwikkelde online ‘Data Management Expert Guide’.
Hierna presenteerde Libby Bishop (Data Linking and Data Security GESIS, Leibniz Institute for the Social Sciences) de wettelijke en ethische aspecten van RDM in Europees perspectief. Zij gaf aan dat data nog steeds gedeeld kunnen worden, alhoewel het door de GDPR lastiger is geworden. Zo moet er nu vaker ‘Informed Consent’ gevraagd worden aan participanten in onderzoek. Anonimiseren van data, wat ertoe moet leiden dat de identiteit van participanten niet herleidbaar is, wordt door de grote hoeveelheid beschikbare data steeds moeilijker. Na de presentatie kregen de deelnemers de opdracht om in kleine groepen ‘Informed Consent’ formulieren uit te werken. De middagsessie werd gepresenteerd door Marlon Domingus (Data Protection Officer Erasmus Universiteit Rotterdam, EUR) en was gericht op de AVG in Nederland. Domingus gaf aan dat het voor veel onderzoekers onduidelijk is hoe zij de AVG moeten toepassen. Als hulpmiddel heeft de EUR een app ontwikkeld die over allerlei privacyvraagstukken uitleg geeft. De sessie werd gevolgd door een praktische opdracht over Data Protection Impact Assessments (DPIA) voor onderzoek. Tijdens het afsluitende panel focuste de discussie zich op het spanningsveld tussen open science en de AVG.

dans.knaw.nl

Jaarvergadering Europeana Network Association

Maarten Heerlien

Voor Europeana was het afgelopen jaar er een van transities. Na ruim 10 jaar gaf Jill Cousins begin 2018 het stokje van Executive Director door aan Harry Verwayen. Ook gingen het financieringsmodel van Europeana’s Digital Service Infrastructure en de organisatiestructuur van de Europeana Network Association (ENA) op de schop in het tweede lustrumjaar van Europa’s digitale platform voor cultureel erfgoed. Deze en andere onderwerpen passeerden de revue op 5 december in Wenen, tijdens de jaarvergadering van de ENA. Het thema luidde Building Communities, een verwijzing naar de introductie in de ENA van een meer community-gedreven organisatiemodel. Zes initiële communities, thematische groepen, werden gepresenteerd: Communicators, Copyright, Education, Impact, Research en Tech. De communities krijgen een open karakter en ENA-leden kunnen vrijblijvend participeren. Teruggeblikt op het afgelopen decennium werd er onder andere aan de hand van de door de Europese Commissie uitgevoerde evaluatie van Europeana. De voornaamste conclusie luidde dat het platform van grote toegevoegde waarde is voor de Europese Unie, erfgoedinstellingen en de Europese burger, maar dat er ruimte is voor groei op het vlak van datakwaliteit en capaciteitsopbouw. Ruimte voor groei bleek er ook buiten de grenzen van Europa. Tijdens de AGM zette Harry Verwayen zijn handtekening onder een Memorandum of Understanding met de Chinese Academie voor Sociale Wetenschappen, gericht op samenwerking op het vlak van gedeeld digitaal erfgoed. Bij sommigen riep dit vragen op gezien de (niet ongebruikelijke) censuur van culturele bronnen in China. Een schone taak dus voor de nieuwe communities om te bewaken dat Europeana’s morele kompas, later op de dag nog door Verwayen aangehaald, ook in de toekomst de juiste kant op blijft wijzen.

pro.europeana.eu

COLOFON **Uitgever**: E-data & Research. **Redactieadres:** Anna van Saksenlaan 51, 2593 HW Den Haag, 070-3494450, edata@dans.knaw.nl, www.edata.nl.
Hoofd-/eindredacteur: Heidi Berkhout. **Redactie**: Marika de Bruijne, Steven Claeyssens, Maarten Heerlien, Machteld Maris, Erica Renckens, Marion Wittenberg.
Redactiesecretariaat: Lucas Pasteuning. **Aan dit nummer werkten mee**: Patrick J.C. Aerts, Ricarda Braukmann, Felix van Gelderen, Seyit Höcük, Olaf Janssen, Frederik Kampstra, Widia Mahabier, Thijs van der Veen, Iris Vocking, Henk Wals, Maarten Zeinstra, Erik van Zummeren. **Opmaak**: Colette Sloots, Haarlem.
Productie: Amsterdam University Press. **Druk**: Ten Brink, Meppel. **Webmaster**: Sonja Duijkers **Oplage**: 6500 papier, 4800 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.

Van de Sompel, Chief Innovation Officer bij DANS:

‘Automatisch artefacten vinden, ophalen en duurzaam archiveren’

“Onderzoekers willen onderdeel uitmaken van hun sociale netwerk op het web, ook voor hun onderzoek.”
E-data interviewt Herbert van de Sompel, per 1 januari Chief Innovation Officer bij DANS.

Marion Wittenberg

Herbert van de Sompel
Van de Sompel (1957) studeerde wiskunde en computerwetenschap aan de universiteit van Gent. Hij promoveerde in 2000 op een proefschrift over contextgevoelig en dynamisch linken van wetenschappelijke informatie.
Vanaf 2002 was hij leider van het onderzoeksteam ‘Digital Library Research and Prototyping’ aan het Los Alamos National Laboratory (New Mexico, USA). Hij is één van de grondleggers van veel gebruikte informatiestandaarden (onder meer OAI-PMH, OpenURL, OAI-ORE, Memento, NISO/OAI ResourceSync, Web Annotation) en nam deel aan invloedrijk onderzoek naar alternatieve metrieken voor wetenschappelijke publicaties (MESUR-project) en ‘reference rot’ in wetenschappelijke communicatie (Hiberlink project).
In 2017 ontving hij de prestigieuze Paul Evan Peters Award voor zijn bijdragen aan duurzame digitale infrastructuren die een diepgaande en blijvende impact hebben gehad op de wetenschappelijke communicatie.
Van de Sompel was in 2010/2011 en in 2013/2014 visiting fellow bij DANS.
hvdsomp.info/bio/

“Onderzoekers zetten overal op het web artefacten neer van hun onderzoeksactiviteiten in voor hen aantrekkelijke systemen”.
credits Bart van Vliet

INTERVIEW

Na bijna twintig jaar in de USA gewerkt te hebben, is Herbert van de Sompel, grondlegger van onder andere het OAI-PMH-protocol, op 1 januari aan de slag gegaan bij DANS. DANS stimuleert onderzoekers om hun digitale onderzoeksgegevens vindbaar, toegankelijk, interoperabel en herbruikbaar te maken. Welke plannen heeft van de Sompel bij DANS? “Ik wil eerst een goed overzicht krijgen van de ontwikkelingen en projecten bij DANS en daarna bekijken wat ik vanuit mijn expertise kan inbrengen. Er zijn een aantal concepten uit mijn recente werk bij het Los Alamos National Laboratory, onder andere het ‘Scholarly Orphans Project’, die hierbij mogelijk interessant kunnen zijn”.

Een spoor van artefacten

“Onderzoekers zetten overal op het web artefacten neer van hun onderzoeksactiviteiten in voor hen aantrekkelijke systemen: presentaties in SlideShare, code in GitHub, data in Figshare of Zenodo. Ze kiezen voor deze globale portals omdat deze zich binnen hun sociale netwerk bevinden en hun zichtbaarheid vergroten. Onderzoekers verplichten om al die bijdragen ook bij hun instelling te deponeren, is problematisch: de toegevoegde waarde voor de onderzoeker is niet meteen duidelijk. Het resultaat is dat instituten geen volledig zicht hebben op wat hun onderzoekers doen, ze zien het uiteindelijke paper, maar niet alle stappen die tot dat paper geleid hebben. En artefacten die op commerciële platforms gedeponeerd worden kunnen ook zomaar verdwijnen”. Het ‘Scholarly Orphans Project’ onderzoekt dit probleem en heeft een prototype ontwikkeld dat automatisch artefacten vindt, ze ophaalt en duurzaam archiveert. “In plaats van de onderzoeker verplicht te stellen hun artefacten bij hun instelling te deponeren, draait de instelling processen waarmee ze die volautomatisch binnenhalen.”

Myresearch.institute

Van de Sompel legt uit dat het project uitgaat van twee perspectieven. Ten eerste dat onderzoeksinstellingen de facto geïnteresseerd zijn in de artefacten van hun onderzoekers. Ten tweede dat de processen van ophalen en archiveren schaalbaar moeten zijn, het gaat om heel veel materiaal. De combinatie van beide perspectieven heeft geleid tot een prototype voor een fictief onderzoeksinstituut: myresearch.institute. “Voor een tiental onderzoekers, die geselecteerd werden omdat ze erg actief op het web zijn, werden de identiteiten die ze in de webportals gebruiken, verzameld. Die identiteiten worden gebruikt als sleutel voor de portal APIs om dagelijks te kijken of een onderzoeker iets nieuws heeft gedeponeerd. Bij GitHub kan dat al gauw om 50 tot 100 nieuwe bijdragen per onderzoeker per dag gaan. Metadata over die bijdragen wordt in een institutionele databank gestopt en de bijdragen zelf worden met web-archiveringstechnieken opgehaald en gearchiveerd”. Voor het ophalen van het materiaal heeft het ‘Scholarly Orphans Project’ een innovatieve procedure ontwikkeld, Memento Tracer. “Omdat steeds meer websites voor gebruikersinteractie van client-side JavaScript gebruikmaken, zijn ze moeilijk automatisch te archiveren. De enige techniek om dergelijke webpagina’s met hoge kwaliteit te archiveren, is om een gebruiker alle essentiële interacties te laten uitvoeren en de resultaten van die interacties weg te schrijven naar een webarchiveringsbestand. Perfect, maar niet schaalbaar. Bij de Memento-Tracer-aanpak doet een curator eenmalig een sessie van interacties met een bepaald type pagina, bijvoorbeeld een landingspagina van SlideShare. Een browser plugin legt deze interacties vast: wat is er achter de schermen gebeurd toen de curator aan het klikken was, welke JavaScript calls zijn aangeroepen. Dit levert een JSON-file met instructies op die als template wordt gebruikt voor een web-archivering crawler. Met deze methode kunnen alle pagina’s van hetzelfde type op industriële schaal en met hoge kwaliteit geharvest worden. Het is momenteel nog experimenteel, maar het concept is echt een doorbraak die we van tevoren niet verwacht hadden”.

‘Volg het online spoor van de onderzoeker’

Duurzaam archiveren

“Initieel waren enkel de eindresultaten van wetenschappelijk onderzoek (papers) beschikbaar op het web. In toenemende mate zijn ook artefacten die tijdens het wetenschappelijk proces gemaakt worden daar te vinden. Er zijn momenteel nog geen frameworks die dat alles archiveren”. Volgens van de Sompel is dit probleemdomein relevant voor DANS. “Het duurzaam bewaren van data is momenteel de missie van DANS, het archiveren van artefacten zou hier naadloos op aan kunnen sluiten. De onderzoeker geeft aan DANS een overzicht van zijn identiteiten bij de door hem gebruikte portals, DANS haalt alles op en archiveert het materiaal”. Dit kan zowel een dienst voor de onderzoeker zijn, als voor de instellingen. “De onderzoeker blijft doen wat hij of zij doet, maar al het materiaal wordt automatisch gearchiveerd. En DANS kan dit vervolgens terugleveren aan de instellingen, waardoor de instellingen een overzicht krijgen van al het materiaal wat hun onderzoekers gecreëerd hebben. Zo zorgen onderzoekers, instellingen en DANS samen voor persistentie van de wetenschappelijke record.”

Technologie en beleid

We vroegen van de Sompel wat hem aantrekt in zijn nieuwe functie bij DANS. De meerwaarde van DANS is in zijn ogen de combinatie van technologie en beleid: “DANS is een voortrekker op dit gebied. Technologisch kan er heel veel, maar het beleid moet de context bieden, het zorgt voor meer realiteitszin”.

myresearch.institute
tracer.mementoweb.org

Werken van in 1948 overleden makers nu vrij herbruikbaar

Nederlandse Publiek Domeindag 2019

Auteursrecht is tijdelijk. Elk jaar komen werken beschikbaar voor vrij hergebruik. Op 3 januari jongstleden werd gevierd dat 'de klas van 1948' tot het publieke domein toetrad.
Olaf Janssen, CC-BY-SA 4.0

Werken die per 1 januari 2019 tot het publieke domein behoren, met de klok mee: Zelfportret van NEMO-grondlegger Herman Heijenbrock *credits Wikimedia Commons.* ***Pot van aardewerk van keramist Chris Lanooy*** *credits Collectie Rijksmuseum.* ***Kinderboek van schrijfster Freddie Langeler*** *credits Digitale Bibliotheek voor de Nederlandse Letteren.* ***250 Polygoonjournaals uit 1948*** *credits Open Beelden*

De Nederlandse Auteurswet zegt dat 'het auteursrecht vervalt door verloop van 70 jaren, te rekenen vanaf de 1e januari volgende op het sterfjaar van de maker'. Elke jaarwisseling loopt dus de auteursrechtelijke bescherming af op teksten, beelden, muziek, films etc. van auteurs, schilders, componisten, cineasten en andere creatieve makers die dan 70 jaar geleden overleden zijn. Hun werk komt na de jaarwisseling in het zogenaamde publieke domein terecht. Dit betekent dat iedereen - natuurlijke én rechtspersonen – zonder beperking, toestemming, bronvermelding, licentie of vergoeding ermee mag doen wat hij/zij wil, mét of zonder winstbejag. Daarom wordt 1 januari ook wel Publiek Domeindag genoemd.

De klas van 1948

Voor de editie van 2019 komen werken van 'de klas van 1948', dat wil zeggen makers die in 1948 zijn overleden, in aanmerking voor onbeperkte digitalisering, herdruk, bewerking, verspreiding en alle andersoortig hergebruik. Hierbij is het overigens niet zo dat alle werken automatisch naar het publieke domein overgaan, dat hangt van de specifieke verschijningsvorm af. Wanneer bijvoorbeeld in een boek waarvan de tekstschrijver in 1948 is overleden ook illustraties staan, dan is dat boek alleen auteursrechtvrij als ook die illustrator in 1948 (of eerder) is overleden. De tekst valt dus wel in het publieke domein, het boek niet per se.

Gevarieerd programma

Net bekomen van de jaarwisseling kwamen op 3 januari zo'n 75 mensen in Den Haag bijeen voor de tweede feestdag van het jaar: Publiek Domeindag 2019. De invulling van dit evenement was tweeledig: enerzijds stonden sprekers van Open Nederland, de Koninklijke Bibliotheek, Beeld en Geluid en Wikimedia Nederland stil bij de betekenis, kansen en weerbarstigheid van het publieke domein, anderzijds werd aandacht besteed aan het leven en werk van een aantal in 1948 overleden makers. Zo passeerden de levensverhalen van schrijfster Cissy van Marxveldt (van de Joop ter Heul-serie), schrijfster-illustratrice Freddie Langeler, filmregisseur Haro van Peski, keramist Chris Lanooy, schrijver Emiel Fleerackers en de schilder Herman Heijenbrock, grondlegger van wetenschapsmuseum NEMO, de revue. Na het succes van 2018 was dit de tweede keer dat Publiek Domeindag in Nederland werd georganiseerd. Een initiatief van Creative Commons Nederland en de KB, met ondersteuning van Wikimedia Nederland en het Nationaal Archief. Internationaal wordt Publiek Domeindag al langer georganiseerd, zo waren er in voorgaande jaren vieringen in onder andere Polen, Duitsland en Israël.
Alle lezingen zijn gepubliceerd als video op:
publiekdomeindag.nl

Olaf Janssen is Wikimedia- en open-data-coördinator bij de Koninklijke Bibliotheek.
Heb je vragen? Mail ze naar olaf.janssen@kb.nl

OPROEP

Doe mee met Creative Commons Nederland

Onderzoeksdata vrijgeven onder een Creative Commons-zero publiek domein verklaring (CC0) betekent dat je toestaat dat anderen kunnen voortbouwen op de vrijgegeven data. Open Nederland, de onafhankelijke vereniging die de activiteiten van Creative Commons ondersteunt en hier voorlichting over geeft, ziet deze vorm van het delen van data als de basis voor een gezond en effectief academisch werkveld. Maarten Zeinstra, voorzitter Open Nederland: "Creative Commons helpt om data en content breed zichtbaar en toegankelijk te maken. Onze licenties geven aan wat gedaan mag worden met auteursrechtelijk beschermd materiaal. We zijn sinds 2004 actief, eerst als samenwerking tussen Kennisland, het Instituut voor Informatierecht en Waag en nu als vereniging. Ook voor onderzoekers is het mogelijk om lid te worden van ons netwerk. Als lid maak je kennis met geïnteresseerden in het publiceren van data onder open licenties. Het lidmaatschap is gratis, aanmelden kan via onze site".

opennederland.nl

SINDS KORT BESCHIKBAAR

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

CentERdata

• Lokale democratie in beeld

Opvattingen van burgers over de lokale politiek, betrokkenheid bij gemeentepolitiek en uitspraken over hoe een goede lokale democratie eruit zou moeten zien. Dit zijn enkele onderwerpen uit het Lokaal Kiezersonderzoek (LKO). Dit onderzoek is in maart 2016 afgenomen in het LISS panel in opdracht van Tom van der Meer, hoogleraar politicologie aan de Universiteit van Amsterdam. Op basis van de data is het rapport 'Democratie dichterbij: Lokaal Kiezersonderzoek 2016' gepubliceerd door Kiezersonderzoek Nederland (SKON). Uit de resultaten blijkt onder andere dat Nederlanders zich minder verbonden voelen met hun gemeente dan met Nederland, maar tegelijkertijd relatief veel vertrouwen hebben in politieke instituties op lokaal niveau. Het databestand Local Voters Survey 2016 is beschikbaar via LISS Data Archive.

lissdata.nl

Ook sinds kort beschikbaar:
Studies LISS panel

• CentERdata, juni - juli 2018, Economic Situation: Income - Wave 11
• Pieters, R.; Giesen, R. van, september 2015 - december 2015 - maart 2016 - juni 2016 - september 2016, Tilburg Consumer Outlook Monitor
• CentERdata, mei - juni 2018, Personality - Wave 10
• CentERdata, april - mei 2018, Work and Schooling - Wave 11
• Soest, A. van; Bonekamp, J., december 2016 - Saving and Spending in Retirement

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

DANS

• Nieuw in EASY: Leidse Wezenzorg

Via EASY is de dataset Leidse wezenzorg 1690-1841 beschikbaar gesteld. De dataset bevat de data van onderzoek naar de weeshuiszorg in Leiden in de periode 1690-1841; een tijd waarin grote veranderingen optraden in de organisatie van de zorg voor minnekinderen. De studie besteedt aandacht aan het uitbesteden van de zorg voor peuters en zuigelingen. Ook komt het profiel van de min aan bod: haar burgerlijke staat, haar ouderdom en haar maatschappelijke positie. De dataset is in 1980 gecreëerd door de in 2007 overleden Leidse sociaal-historicus Dirk-Jaap Noordam.
DOI: 10.17026/dans-zyx-w2vz

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

• Berkhout, dr D.J. (UvA) (2018): Pyttersen's Almanak. DANS.
DOI: 10.17026/dans-z73-s9k7
• Cavallo, Dr C. (UvA) (2018): Velsen 1: data retrieval of analyses on the faunal remains from the Roman harbor (15-30AD). DANS. DOI: 10.17026/dans-zat-586g
• Dijk, Dr. S. van (Huygens ING) (2018): NEWW Women Writers. DANS.
DOI: 10.17026/dans-x4u-2vha
• Doorenbosch, Dr M. (Faculty of Archaeology Leiden University) (2013): Ancestral Heaths. Reconstructing the barrow landscape in the central and southern Netherlands. DANS.
DOI: 10.17026/dans-xy4-by6m
• Enckevort, H. van; Harmsen, C. (Gemeente Nijmegen) (2015): In de periferie van de canabae legionis. Archeologisch onderzoek in de Frans Halsstraat en de Daalseweg, Archeologische Berichten Nijmegen - Briefrapport 192. DANS.
DOI: 10.17026/dans-zh9-gmug
• Jordanov, drs. M.S.; Hoof, drs. B.I. van (Raap Archeologisch Adviesbureau B.V.) (2013): Plangebied Veerstalblok 17 te Gouderak, gemeente Ouderkerk; archeologisch onderzoek: een opgraving. DANS. DOI: 10.17026/dans-za4-nac5
• Klandermans, prof. dr. P.G. (VU); Van Stekelenburg, dr. J.; Gaidyte, Dr. T. (2014): Caught in the act of protest: CCC-project. DANS. DOI: 10.17026/dans-zwj-gkeu
• Leije, MA J. van der (Archeologisch Onderzoek Leiden) (2018): Archeologisch onderzoek van boerderij Veldheim. DANS.
DOI: 10.17026/dans-xpj-2hhz
• Ringenier, drs. H. (Archeologie Deventer) (2018): Inventariserend proefsleuvenonderzoek en opgraving Douweler Leide Zuid. DANS.
DOI: 10.17026/dans-xtc-5xg8
• Smole, drs. L. (Gemeente Arnhem) (2018): Musis Sacrum, Archeologische begeleiding van de sloop van de theaterzaal t.b.v. nieuwbouw. DANS.
DOI: 10.17026/dans-zg2-aznd

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

PAN, BBMRI-omics en OpenINTEL winnaars Dataprijs 2018

Toegang tot en delen van data beloond met prijs

“Alle inzendingen zijn inspirerende eindproducten van onderzoeken.” De lovende woorden van juryvoorzitter Stan Gielen tijdens de start van de prijsuitreiking van de Nederlandse Dataprijs beloofden al veel goeds.

Heidi Berkhout

De winnaars van de Dataprijs 2018 van links naar rechts: Bas Heijmans, Roland van Rijswijk-Deij en Stijn Heeren. *foto Bart van Vliet*

Eind november werd voor de vijfde keer de Nederlandse Dataprijs uitgereikt door Research Data Netherlands (RDNL). De prijs (een Dataprijs 2018-sculptuur en € 5.000 om de dataset toegankelijk(er) te maken) geeft waardering aan onderzoekers die extra bijdragen aan de wetenschap door onderzoeksdata beschikbaar te stellen voor nieuw of aanvullend onderzoek. Uit 47 inzendingen en 9 nominaties werden uiteindelijk drie winnaars gekozen. E-data zet ze op een rij.

PAN: uniek online platform

Portable Antiquities of the Netherlands (PAN), een uniek online platform dat archeologische vondsten van burgers beschikbaar stelt voor wetenschappelijk onderzoek en publieke interesse, is winnaar in de categorie humaniora en sociale wetenschappen. De jury oordeelde als volgt: “PAN bouwt bruggen tussen amateurs en de professionele archeologie en is een schoolvoorbeeld van de manier waarop citizen science de wetenschapspraktijk kan veranderen en verbeteren. De data zijn op een voorbeeldige manier gecureerd en via linked open data beschreven en ontsloten. Vele duizenden vondsten van amateurarcheologen worden op deze manier voor de wetenschap beschikbaar gesteld en dat leidt ook al daadwerkelijk tot belangrijke nieuwe wetenschappelijke inzichten over de (pre)historie van onze voorouders.” De prijs werd overhandigd aan Dr. Stijn Heeren. Heeren: “Het geeft een enorme boost om te merken dat wij in de ontwerpfase van PAN de juiste keuzes hebben gemaakt. Heel stimulerend om erkenning te krijgen voor de dataset van PAN, en daarmee ook een extra compliment aan de hobby-archeologen van Nederland, die de gegevens over hun vondsten aan ons hebben geleverd. We gaan het prijzengeld besteden aan extra functionaliteit op de publieke website (dus vóór de inlog) zodat het brede publiek nog beter de PAN-dataset kan inzien en bevragen.”

BBMRI-Omics: unieke samenwerking

Winnaar in de categorie medische en levenswetenschappen is BBMRI-Omics, een unieke samenwerking van alle academische centra in Nederland gericht op moleculaire big data voor het ontdekken van ziektemechanismen en biomarkers. Uit het juryrapport: “De dataset geeft een nieuwe dimensie aan het gebruik van data in de medische wetenschap. Het team heeft al veel ingezet op kennisoverdracht en is van plan dit in de toekomst nog meer te doen. De data worden continue verbeterd en aangevuld. Sinds 2014 zijn de ruim 60 publicaties al meer dan 2.500 keer geciteerd, waaronder artikelen in toptijdschriften als Nature, Nature Genetics, Science, Genome Biology en Nature Communications. Het is een unieke samenwerking tussen de Nederlandse onderzoekscentra met biobanken. BBMRI-Omics heeft aangegeven het prijzengeld in te gaan zetten voor een training om nog meer onderzoekers optimaal gebruik te laten maken van BBMRI-Omics. Een hartstikke goed initiatief dat de jury van harte toejuicht!” De prijs werd overhandigd aan Bas Heijmans. Heijmans: “We zijn vereerd met deze fantastische erkenning voor het werk van honderden onderzoekers uit heel Nederland. De € 5.000 komt op het goede moment. BBMRI-Omics heeft kortgeleden een grote update gehad en daarom willen we trainingen geven aan jonge onderzoekers zodat ze nog meer uit de gegevens kunnen halen.”

OpenINTEL: buitengewoon originele dataset

In de categorie exacte en technische wetenschappen werd OpenINTEL Active DNS Measurements winnaar. De jury was onder de indruk van de rijkdom van deze dataset waarmee meer dan 60% van alle domeinen op de wereld in kaart gebracht wordt en noemde het een buitengewoon originele dataset dat zich ook door het live karakter onderscheidt van andere inzendingen. Roland van Rijswijk-Deij: “Het was echt een hele mooie verrassing om de Dataprijs te winnen! Het klinkt misschien cliché, maar de andere genomineerde projecten vond ik ook heel interessant. In Twente werd door de rest van ons team enorm meegeleefd met de uitreiking, we hebben de hele dag berichtjes uitgewisseld. Toen de prijs bekend was, stroomden de felicitaties binnen, tot aan de rector toe, daar zijn we best trots op! Het prijzengeld kunnen we goed gebruiken.
Zo helpt een nieuw systeem ons bij het uitbreiden van onze capaciteit om open data beschikbaar te stellen. Daarnaast willen we onze open data verrijken met afgeleide data, zoals data over in welke landen domeinnamen worden gehost, welke partijen mail afhandelen voor domeinen en hoe goed domeinnaamhouders hun e-mailafhandeling beveiligen. Onze site visualiseert dit al middels grafieken, maar om de nodige scripts te schrijven en te onderhouden, kan een student assistent ons helpen.”

researchdata.nl

Celebrating Data! What’s next?

De Dataprijs werd dit jaar uitgereikt tijdens het event ‘Celebrating Data! What’s next?’. Onderzoekers en dataprofessionals kregen deze dag aangeboden door RDNL, het Landelijk Coördinatiepunt Research Data Management, Werkgroep Research Data van het samenwerkingsverband van universiteitsbibliotheken en de Koninklijke Bibliotheek, Netherlands Federation of UMCs Data4Life-Sciences en het Nationaal Platform Open Science.

AGENDA

11 februari • Brussel
NPSO lezingendag
Met als thema ‘Uitdagingen bij datakoppelingen voor statistisch en surveyonderzoek’.
npso.net/evenementen

14 februari • Den Haag
Open dag over open science
‘Love to share data’. Voor de 2e keer organiseert DANS een open dag over open science.
dans.knaw.nl/love-to-share-data

18 - 21 maart • Hilversum
Dutch Digital Conference
De eerste editie van Dutch Digital Conference heeft als thema: ‘Smart applications in the area of artificial intelligence’.
nederlanddigitaal.nl

19 - 20 maart • Berlijn
Open Science Conference
Een uniek forum voor onderzoekers, bibliothecarissen, uitvoerders, beleidsmakers en andere belanghebbenden.
open-science-conference.eu

16 mei • Keulen
SHARE User Workshop
Ontmoetingsplaats voor gebruikers van SHARE data.
share-project.org/press-news/news.html

27 - 31 mei • Sydney
IASSIST Annual Conference
De jaarlijkse conferentie van de International Association for Social Science Information Services and Technologies
iassist2019.org

28 - 31 mei • Fisole
QQML 2019
Met extra aandacht voor: ‘Libraries and Information services: New technologies, innovative processes and the Information Professional’.
qqml.org/event/qqml2019

KORT

Onderzoek NPO-programma’s met 888zoeker

Vrijwel alle NPO-programma’s hebben ondertiteling voor doven en slechthorenden. De 888zoeker is een niche zoekmachine die deze ondertiteling indexeert en doorzoekbaar maakt op woordniveau. De tool is in eerste instantie voor journalisten gemaakt, maar kan ook door academici gebruikt worden om ontwikkelingen in de media te onderzoeken. Dat kan onder meer met een n-gram gedaan worden, maar op aanvraag is ook een API beschikbaar en er wordt momenteel gewerkt aan een koppeling met LIWC. De tool is gemaakt door Erik van Zummeren (vanuit Research Assistant) en is gedeeltelijk gefinancierd door het Stimuleringsfonds voor de Journalistiek. (EvZ)
888zoeker.nl

KORT

FAIRsFAIR van start

Met een financiering van 10 miljoen euro en ruim 20 partners start op 1 maart het driejarig Europese project FAIRsFAIR. De European Open Science Cloud (EOSC) moet het voor onderzoekers eenvoudiger maken om data te delen en te combineren. Het FAIRsFAIR-project gaat EOSC helpen bij het opstellen van FAIR-principes voor deelname aan de EOSC. Bovendien gaat het project alle kennis over FAIR bundelen en via leertrajecten toegankelijk maken. De zes kernpartners op een rij: Data Archiving and Networked Services (coordinator), CSC - IT Center for Science Ltd., het Digital Curation Centre en de Science and Technology Facilities Council, Trust-IT en de European University Association. (EF)
dans.knaw.nl

Beleef het verleden met Time Machine

Time Machine is een buitengewoon ambitieus plan van Frédérique Kaplan, directeur van het Digital Humanities Laboratory van de École Politechnique Fédérale de Lausanne. Via een FET Flagship-aanvraag, een prestigieus Europees programma, kan een bedrag van een half miljard euro worden toegekend. Een Time Machine maakt het mogelijk om virtueel terug te reizen in de tijd, doorgaans in een stedelijke omgeving. Enorme hoeveelheden informatie uit gedigitaliseerde documenten worden gehecht aan coördinaten op kaarten en momenten in de tijd. Gebouwen en straten worden driedimensionaal gerepresenteerd, waardoor het bijvoorbeeld mogelijk wordt in een bepaalde periode door de stad te bewegen en te zien wat er in elk gebouw te doen was.
Een mooi voorbeeld wordt geleverd door de (in aanbouw zijnde) Venice Time Machine. Tijdens de Time Machine Conference eind vorig jaar was Nederland sterk vertegenwoordigd, met vier projecten: Amsterdam (Julia Noordegraaf), Sittard-Geleen (Peer Boselie), Leiden (Hans Mol) en Utrecht (Toine Pieters). Naast inzicht in methodologie en techniek, werd ook aandacht besteed aan de toekomst van het beleven van het verleden. Het wachten is nu op de mogelijke toekenning van een subsidie waarmee een breed consortium (waaronder het KNAW Humanities Cluster, UvA, UU, TUDelft, Nederlands Instituut voor Beeld en Geluid en Picturae) de volgende stap kan zetten in dit fascinerende project.
(Henk Wals)
vtm.epfl.ch

Digitale verrijking kunsthistorische publicaties

Het leven van Kunstenaars 2.0

Met het initiatief ArtLives beogen het Huygens ING, Rijksmuseum en het RKD vroegmoderne kunstenaarsbiografieën digitaal te ontsluiten en te verrijken.

Maarten Heerlien

Kunstenaarslevens vormen een beknopt maar bepalend genre binnen de kunstgeschiedenis in de Lage Landen. Het eerste in zijn soort, het in 1604 door Carel van Mander geschreven Schildersboeck, bevat uitvoerige biografieën over in die tijd bekende Nederlandse en Vlaamse kunstenaars. In de periode tot 1840 schreven verschillende auteurs vervolgen op Van Mander, met aanvullende, soms overlappende maar soms ook aangepaste informatie over kunstenaars: hun opleidingen, hun belangrijkste werken, de locatie daarvan en soms de roddels die over hen de ronde deden.

Rijke informatie

Jenny Reynaerts, senior conservator 18de en 19de-eeuwse schilderkunst bij het Rijksmuseum, neemt onder de noemer ArtLives het voortouw om deze nog altijd relevante contemporaine kunsthistorische bronnen digitaal te ontsluiten. “In geen ander land was er zo’n sterke traditie binnen het genre van Kunstenaarslevens als in Nederland. Deze boeken bevatten rijke informatie op het gebied van de Nederlandse kunstgeschiedenis en zijn bovendien van grote invloed geweest op de canon van de kunst.”
Het ArtLives-initiatief komt voort uit een in 2017 uitgevoerde haalbaarheidsstudie, gefinancierd door het Mondriaanfonds en het Rijksmuseum Fonds. Het beoogde project start met een pilot aan de hand van Arnold Houbraken’s Groote schouburgh der Nederlantsche konstschilders en schilderessen (1718-1721), het bekendste Kunstenaarsleven naast dat van Van Mander, met beschrijvingen over beroemde kunstenaars uit de Gouden Eeuw.

Titelprent uit de tweede editie van Houbraken’s Groote schouburgh der Nederlantsche konstschilders en schilderessen uit 1753.
Afbeelding: Rijksmuseum Research Library, PDM 1.0

Integraal digitaliseren

In de geplande pilot voorziet ArtLives een vergelijkbare aanpak als The Mondrian Letters, een project van het Huygens ING en RKD - Nederlands Instituut voor Kunstgeschiedenis. Houbraken wordt integraal gedigitaliseerd en de tekst wordt omgezet naar XML. Vervolgens worden verschillende *named entities* gelabeld om de tekst doorzoekbaar te maken en te analyseren op persoons- en instellingsnamen, locaties, kunstwerken en kunsttermen. Ook wordt de tekst semi-automatisch verrijkt met afbeeldingen en gestructureerde data uit RKDartists&, RKDimages en het Biografisch Portaal van het Huygens ING.
In ArtLives vormt de in de Kunstenaarslevens gebruikte kunstterminologie een bijzonder thema. Reynaerts: “In deze boeken wordt het eerste kunsthistorische en artistieke jargon geformuleerd. We willen dat jargon door de tijd heen volgen en veranderingen in vorm en betekenis analyseren. Dat is relevant voor de duiding van andere contemporaine bronnen. Dit onderdeel, ArtSpeak, vormt de laatste fase van het project, aangezien een dergelijke analyse pas kan als het totale corpus digitaal is verrijkt.”

Open content

Indien de pilot succesvol verloopt, wordt de verrijkte editie van Houbraken als open content online gepubliceerd. De resultaten van het project zullen dan worden geborgd in de digitale infrastructuren van de projectpartners.

Hoe houd je software levend, hoe maak je levende software?

Snapshot van software wijsheid

De ultieme software sustainability kwesties: hoe maak je levende software en hoe maak je software levend? Een snapshot van de huidige situatie. ***Patrick J.C. Aerts***

Software bestaat sinds de jaren vijftig. Eerst was er alleen assembler, sinds 1954 kwam Fortran, de waarschijnlijk eerste hogere programmeertaal, en in de jaren zestig volgden aanzienlijk meer hogere programmeertalen. Vervolgens werd software meer en meer geschreven voor specifieke machines, met hun eigen bedrijfssystemen en processoren met hun eigen instructiesets. In de loop van de tijd werd er geüniformeerd (denk aan Unix, Linux, 8086-instructies), maar trad ook diversiteit op (denk aan FPGA’s, GPU’s). Software uit voorbije periodes, waaraan nog steeds of opnieuw behoefte bestaat, maar welke moeilijk te restaureren is. De vraag is, hoe je software zo kunt schrijven dat deze makkelijk te onderhouden is en dus op duurzaamheid geschreven?
Het schrijven van software was traditioneel gericht op een direct en concreet resultaat: de computer binnen een omgeving een opdracht laten uitvoeren. Duurzaamheid stond niet op het netvlies. Maar tegenwoordig kunnen zelfs al tijdens de duur van een onderzoek de omgeving en het onderzoeksteam veranderen. Hoe houd je de software en de kennis over de software levend? En wat is er nodig om dit tegen minimale kosten te kunnen doen?

Duidelijke signalen

In de afgelopen twee á drie jaar werden over dit onderwerp steeds vaker conferenties en workshops georganiseerd. Tijdens de internationale conference Supercomputing & Communications (SC’18) in november in Dallas, werden duidelijke signalen gegeven:

- Het softwareprobleem is niet typisch voor de wetenschap: kijk ook naar andere domeinen (kunst, archief, bibliotheken, game-industrie);
- Maak gebruik van communities met een specifiek en gemeenschappelijk belang;
- Zet samenwerkingsverbanden op, bijvoorbeeld met het Software Sustainability Institute (UK) en/of de Research Data Alliance;
- Wees voorzichtig met containerisatie (zoals Docker). Het plaatsen van software met omgeving in een container transporteert de software handig tussen en over clouds, maar het houdt weinig rekening met veranderingen in omgeving;
- Breng het onderwerp in bij opleidingen waar les wordt gegeven in programmeren.

FAIR-software

DANS en het eScience Centrum werken samen aan de beste route voor onderzoekers die iets met software willen gaan doen en/of de vraag hebben hoe ze hun software FAIR kunnen maken. In 2019 wordt opnieuw een Software Sustainability workshop georganiseerd. Stuur een mail naar info@dans.knaw.nl om op de hoogte te worden gehouden.

Patrick Aerts is in dienst van NWO, werkt bij het Netherlands eScience Center (Strategic Alliances) en bij DANS (Senior research fellow) en is voorzitter van PLAN-E, het Platform of National eScience Centers in Europe.

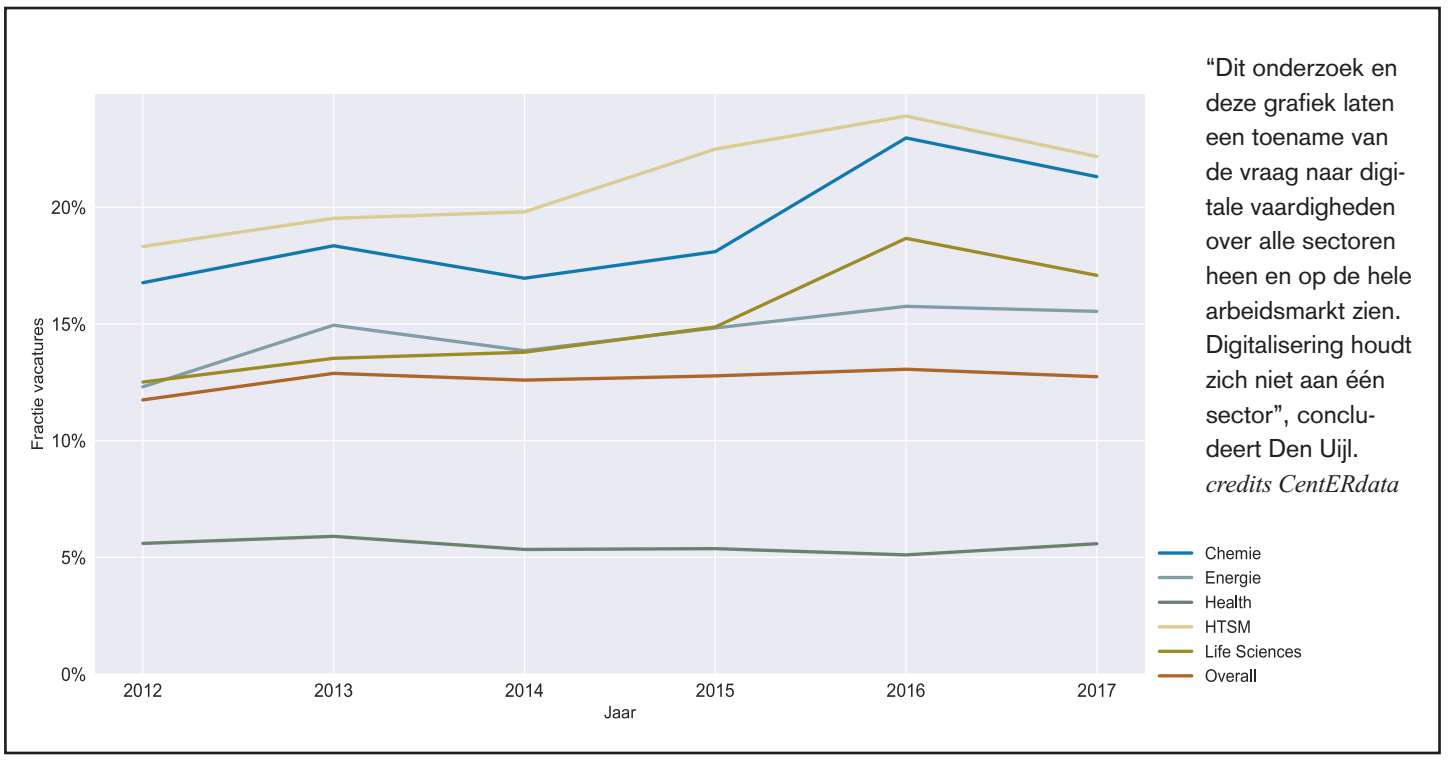
dans.knaw.nl

Rapport over effect digitalisering op arbeidsmarkt

Arbeidsmarktonderzoek ICT

Begin 2019 verschijnt het rapport van het 'Arbeidsmarkt-onderzoek ICT' over het effect van digitalisering op de arbeidsmarkt.
Marcia den Uijl, senior data scientist bij CentERdata, vertelt.

Marika de Bruijne



"Dit onderzoek en deze grafiek laten een toename van de vraag naar digitale vaardigheden over alle sectoren heen en op de hele arbeidsmarkt zien. Digitalisering houdt zich niet aan één sector", concludeert Den Uijl.
credits CentERdata

"De krimpende beroepen door digitalisering? Dat zijn met name secretaresses, boekhoudkundig medewerkers en callcenter-medewerkers outbound", aldus Den Uijl. "Arbeidsmarktramingen van het Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA) geven aan dat meer beroepen zullen krimpen, maar bij deze beroepen is het te verwachten dat men op zoek zal moeten naar ander werk." En dat al op de korte termijn. "De prognoses zijn gemaakt tot 2022", aldus Den Uijl. "Het goede nieuws is dat we konden aantonen dat er optimale overstapberoepen bestaan voor meer dan 90% van de getroffen personen."

Acht miljoen vacatures

In het Arbeidsmarktonderzoek ICT werden acht miljoen vacatures van de afgelopen vijf jaar bestudeerd, getrokken uit Jobfeed, de grote online databank van Textkernel. Deze ongestructureerde data is gecombineerd met de gestructureerde prognoses van het ROA. Een mooie uitdaging voor data science. Met behulp van Natural Language Processing (NLP) werd gekeken in hoeverre digitale vaardigheden nodig zijn voor verschillende beroepen en hoe vergelijkbaar verschillende beroepen met elkaar zijn. "We zien dat er beroepen zijn waar tekorten aan mensen voorkomen. Met de vergelijkbaarheidsanalyses kan een werkgever beoordelen uit welke andere beroepsgroepen nieuwe arbeidskrachten kunnen worden gehaald", vertelt Den Uijl.

Nieuwe technieken

"Door nieuwe tekstanalyse-technieken te combineren met arbeidsmarktramingen komen stabiele loopbaanpaden in zicht, welke in lijn zijn met de overstapberoepen die door het UWV zijn geïdentificeerd. Maar de hoeveelheid data maakt dat je completer kunt zijn. We hebben 371 beroepen in een matrix kunnen vergelijken. Omdat online vacatures minder representatief kunnen zijn - bedrijven vullen open posities ook via andere kanalen - blijven de 'oude manieren' van onderzoek doen, zoals surveys, echter ook van belang." Toch is Den Uijl enthousiast over de gebruikte data science analyses. "De agnostische benadering van data science benadrukt dat je kijkt naar wat mogelijk is, in plaats van wat al gebeurd is."
Het onderzoek is uitgevoerd in opdracht van een breed consortium met ICT-vertegenwoordigers (CA-ICT, Nederland ICT en CIO Platform Nederland), de vertegenwoordigers van vijf Nederlandse topsectoren en het ministerie van SZW. Het rapport verschijnt in 2019 en zal beschikbaar zijn via de website van CentERdata.

centerdata.nl

Slim en efficiënt zoeken met AGNES

Graven in archeologische onderzoeksrapporten

Archeologen in Nederland produceren zo'n 4.000 opgravingsrapporten per jaar. Alex Brandsen onderzoekt hoe deze schat aan informatie beter ontgonnen kan worden.

Steven Claeyssens

Nederlandse archeologen hebben op dit moment zo'n 60.000 rapporten digitaal beschikbaar, bij DANS en in andere e-depots. Al deze rapporten samen bevatten een gigantische hoeveelheid archeologische informatie, maar het is heel moeilijk om hierin alle relevante informatie over een bepaalde plaats of periode terug te vinden. De huidige systemen doorzoeken namelijk alleen de metadata van de rapporten. Deze metadata beschrijven bijvoorbeeld dat een rapport de Middeleeuwen behandelt, maar vermelden niet dat er ook enkele artefacten uit de Bronstijd zijn gevonden, terwijl deze objecten belangrijk zouden kunnen zijn voor een onderzoek over de Bronstijd. Daarom is het nodig om alle tekst goed doorzoekbaar te maken. Alex Brandsen nam deze taak op zich als promovendus aan de Universiteit Leiden.

Taal begrijpen

Na een archeologie-bachelor in Leiden, een master Archeological Information Sytems in York en ervaring als web developer in Leeds, startte Brandsen in 2017 met zijn promotieonderzoek. Hij wil de Nederlandse archeologische rapportproductie veel dieper ontsluiten: "Dat kan met full text-zoeken, zoals in Google, maar ook dan kunnen zich problemen voordoen. Bij de zoekterm Middeleeuwen vindt een full text-zoekactie bijvoorbeeld niet 'Middeleeuwse' en zeker niet '1000 na Christus'. Deze synonymie is een veelvoorkomend fenomeen in rapporten. Ook het omgekeerde probleem komt voor, namelijk wanneer één woord verschillende betekenissen heeft. Om al deze complicaties het hoofd te bieden, moet een zoeksysteem taal tot op zekere hoogte 'begrijpen' en ook specifiek archeologische concepten kunnen herkennen."

JONG TALENT

AGNES

"In mijn project pas ik text mining (en specifiek Named Entity Recognition) toe om automatisch relevante archeologische concepten te herkennen in tekst. Hiervoor gebruik ik machine learning, een vorm van kunstmatige intelligentie die op basis van voorbeelden uit handmatig geannoteerde teksten nieuwe woorden automatisch kan classificeren. In het verleden is daar mee geëxperimenteerd, een bruikbaar systeem heeft het helaas nog niet opgeleverd. Het doel van mijn project is om een webapplicatie te bouwen: AGNES (Archaeological Grey literature Named Entity Search). Met AGNES zoeken archeologen op een slimme en efficiënte manier door die stapels Nederlandse opgravingsrapporten, waardoor sneller en beter onderzoek te verrichten is in de Nederlandse archeologie."
Een aantal versies van AGNES staan al online en kunnen door iedereen na registratie gebruikt worden.

agnessearch.nl

Brandsen past textmining toe voor het automatisch herkennen van relevante archeologische concepten in teksten

Gratis online training helpt onderzoekers bij datamanagement

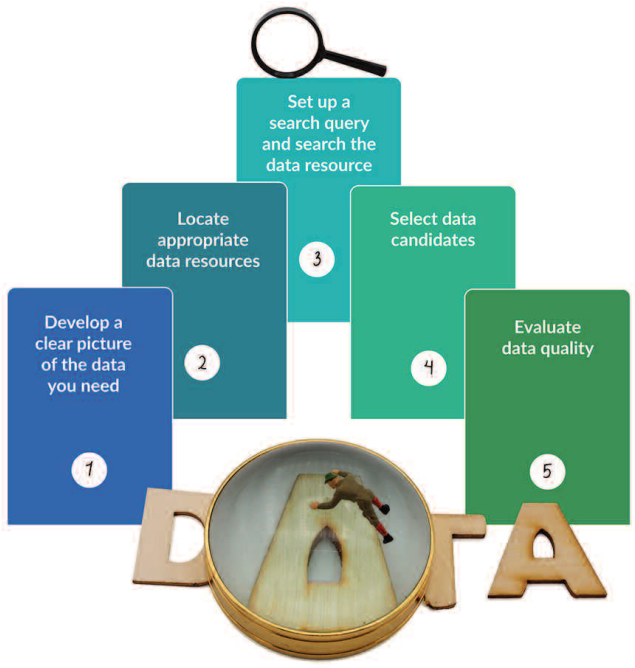
Leren hoe je data kunt vinden

CESSDA ERIC, het consortium van Europese sociaal-wetenschappelijke data-archieven, heeft aan haar gratis online training over datamanagement een nieuw hoofdstuk toegevoegd: data discovery.
Ricarda Braukmann

Vanaf 2017 telt de Data Management Expert Guide zes hoofdstukken over het management en hergebruik van sociaalwetenschappelijke data. Het bevat praktische tips voor het hele onderzoeksproces: over de planning van een onderzoeksproject, het organiseren van de dataverzameling, het verwerken van gegevens alsook het archiveren en publiceren van de onderzoeksdata (zo FAIR mogelijk).

Data life cycle

Aan de bestaande training is onlangs een hoofdstuk over het vinden van data (data discovery) toegevoegd, waardoor de training nu alle stappen van de data life cycle omvat. Dit nieuwe hoofdstuk biedt onderzoekers tips en trucs over het vinden van bestaande data, data die zij kunnen hergebruiken om nieuwe onderzoeksvragen te beantwoorden.



Bij het vinden van data zijn vijf stappen belangrijk: verkrijg een duidelijk beeld van de benodigde data, bedenk welke bronnen interessant kunnen zijn, zoek actief binnen dataverzamelingen, selecteer interessante datasets en als laatste stap: evalueer de kwaliteit en toepasbaarheid van de gevonden data.
credits Verbeeldingskr8t / CESSDA ERIC

Vijf stappen

Het hoofdstuk beschrijft vijf stappen in het vinden van data. De eerste stap gaat over de uitdaging om een duidelijk beeld te krijgen van de soort data die men wil vinden. Het hoofdstuk presenteert een aantal vragen om goed te definiëren naar welke data gezocht wordt. Vervolgens wordt een overzicht van mogelijke bronnen waar data gevonden kunnen worden, gegeven. Naast de CESSDA ERIC data-archieven die sociaalwetenschappelijke data vindbaar en toegankelijk maken, worden ook andere Europese en internationale databronnen toegelicht. Stap drie in het vinden van data is het actief zoeken binnen een archief of dataverzameling. Het hoofdstuk geeft een aantal tips voor het formuleren van effectieve zoekopdrachten waarmee bruikbare data gevonden kunnen worden. Als laatste wordt aandacht besteed aan het selecteren van datasets en aan het evalueren van de kwaliteit en toepasbaarheid van de data. Wat zijn vaak voorkomende toegangscategorieën? Waar mag ik de data voor gebruiken? Zijn ze beschikbaar in het juiste formaat? Zijn er kosten verbonden aan hergebruik? Hoe refereer ik aan de data? Dit zijn allemaal vragen waar het hoofdstuk onderzoekers mee op weg helpt om uiteindelijk de juiste dataset te kunnen selecteren.

De Data Management Expert Guide is online gratis beschikbaar.

cessda.eu/DMEG

Dr Ricarda Braukmann is programmaleider sociale wetenschappen bij DANS. DANS heeft als Nederlandse Service Provider van CESSDA ERIC bijgedragen aan de ontwikkeling van de Data Management Expert Guide.

GELEZEN

ESFRI Roadmap

Maarten Heerlien

In het najaar van 2018 publiceerde ESFRI, het European Strategic Forum on Research Infrastructures, een nieuw strategisch rapport, met daarin de landschapsanalyse van elk van de zes ESFRI-kennisdomeinen. Voor het SSH-domein (Social & Cultural Innovation) ziet het ESFRI-forum kansen in de verdere ontwikkeling van big-data-analyse in taaltechnologie. Godsdienstwetenschappen en doorontwikkeling van digitale diensten voor open science worden aangewezen als strategisch belangrijk voor SSH.
Aan de Roadmap 2018 zijn zes projecten toegevoegd, wat het totaal aan lopende ESFRI-projecten op 18 brengt. Voor twee van deze nieuwkomers fungeert Nederland als lead country. European Holocaust Research Infrastructure beoogt een onderzoeksinfrastructuur te ontwikkelen voor eenduidige toegang tot en analyse van geografisch verspreide bronnen over de Holocaust en wordt gecoördineerd door het NIOD. Distributed System of Scientific Collections richt zich op virtuele integratie en ontsluiting van Europese natuurhistorische collecties. Coördinatie van DiSSCo ligt bij Naturalis Biodiversity Center.
roadmap2018.esfri.eu

COLUMN

Een beleefde revolutie: differential privacy

Reden waarom ik van de statistiek houd, nummer 433: statistici zijn van die heerlijk ingetogen mensen. De hoogste lof die je als statisticus kan ontvangen is dat je 'voorzichtig' bent. En de meest negatieve reactie waar ik getuige van ben geweest, van een statisticus op een tochniet-zo-heel-zinnig plan van onze groep: een bedachtzame pauze - gevolgd door 'kán je doen...'. De drie puntjes waren hoorbaar, maar vergevingsgezind.
Dus als je in een vakblad leest, 'de bezorgdheid is reëel en het gevaar is ook reëel', dan let je op. Nu is dit citaat al uit 1972. Maar wel van Ivan Fellegi, een statisticus die zijn tijd ver vooruit was. Fellegi was een Hongaarse immigrant die na de opstand noodgedwongen naar Canada vluchtte, om daar allerlei briljante artikelen te schrijven over onderwerpen die vandaag de dag opeens zeer actueel zijn, zoals het koppelen van verschillende databestanden met onzekerheid. Hij werd ook de 'Hoofdstatisticus' van Canada, een titel die ik persoonlijk veel mooier vind dan 'dichter/theoloog/ramenlapper des vaderlands'. Wat dit vooral betekende is dat hij zich bemoeide met de officiële statistiek. En het gevaar waar hij zich druk om maakte? Privacy.

In de jaren '70 bestond de bezorgdheid over privacy slechts bij een paar helderziende individuen, die toen al inzagen dat de opkomst van computers en grote databestanden een nieuwe tijd inluidde. Sla nu maar eens een krant open zonder dat allerlei privacy horror stories je bespringen. Zelfs de politiek is wakker geworden, dus dan weet je zeker dat het allang uit de hand is gelopen.

Waarschijnlijk het gevaar, maar zeker de bezorgdheid, zijn nu uitgegroeid tot zo'n groot probleem, dat de Census Bureau (het Amerikaanse CBS) een drastische beslissing heeft genomen. Vanaf nu worden de resultaten van de volkstelling uitsluitend gepubliceerd met behulp van een statistische databeschermingstechniek genaamd "differential privacy".

Wat is dat nou weer? Welnu. Zelfs als je duidelijke 'identificatoren' - variabelen zoals naam, adres, postcode - uit een bestand verwijdert, blijkt het toch vaak mogelijk om personen te herleiden. Dit kan bijvoorbeeld door zo'n 'opgeschoond' databestand te koppelen aan andere bestanden die her en der te vinden zijn. Een beroemd voorbeeld is de heridentificatie van een aantal mensen uit een dataset met miljoenen Amerikanen die Netflix online beschikbaar had gesteld voor onderzoeksdoeleinden. Het CBS beschermt ons al sinds jaar en dag tegen dit soort praktijken, en speelt internationaal zelfs een leidende rol in het ontwikkelen van het soort databeschermingstechnieken waar Fellegi in 1972 over schreef.

Maar een groep informatici, aangevoerd door Cynthia Dwork van Harvard, was toch ontevreden. Ze bedachten een strenge, formele definitie van privacy en een set methoden om die te waarborgen: differential privacy. Het idee is simpel: stel, er moet een 'uitkomst' gepubliceerd worden. Dat kan een tabel zijn of een correlatie, maar ook een volledige dataset. Deze data worden niet lukraak op het internet geplempt, maar moeten eerst een verstoring ondergaan, bijvoorbeeld door er willekeurige ruis bij op te tellen. Als je uit deze verstoorde uitkomst niet met voldoende zekerheid kan bepalen hoe de oorspronkelijke dataset er uit zag, dan is er ook bijna geen kans op het herleiden van individuen. Je kunt zelfs niet goed bepalen óf een bepaalde persoon wel of niet in de oorspronkelijke dataset zat, ook al weet je verder letterlijk alles over die persoon.

Differential privacy is een fascinerend, maar controversieel, begrip. Open data wordt een fluitje van een cent, als je eraan kunt voldoen. Het nadeel is natuurlijk dat je door de verstoringen ook minder kunt met de data: er moet een balans gevonden worden tussen de bruikbaarheid en de bescherming van de data. Daarover wordt nu dan ook (voor statistische begrippen) fel gedebatteerd in Amerika. Is John Abowd, het hoofd van de Census, wel 'voorzichtig'?

De discussie komt ook naar ons land. Gevaar en bezorgdheid zijn er al. Differential privacy dient zich binnenkort vast ook aan in de Europese officiële statistiek, en in software voor onderzoeksdatabeheer zoals iRods, Dataverse, of Figshare. In het slechtste geval moet de sociale wetenschap op de schop: iets lastigere data-analyses, grotere steekproeven, meer preregistratie, en nieuwe onderzoeksontwerpen. In het beste geval zijn er binnenkort geen excuses meer om onderzoeksgegevens over mensen niet open te delen. "Goed te doen…"

Daniel Oberski

Daniel geeft de volgende column graag aan Pearl Dykstra.

Daniel Oberski
Daniel Oberski is universitair hoofddocent in methodologie van data science en statistiek aan de Universiteit Utrecht. Hij promoveerde in Tilburg en Barcelona en was visiting professor in Maryland. In 2014 ontving hij een Veni-subsidie voor het ontwikkelen van methoden die meetfouten in administratieve registerdata opsporen en corrigeren.