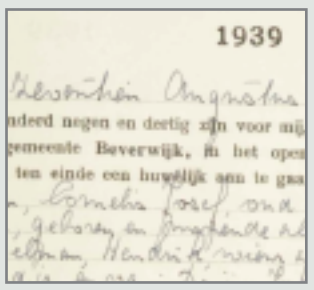
E-DATA & RESEARCH

Jaargang 18 | nummer 3

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

E-data & Research verschijnt drie keer per jaar en wordt mogelijk gemaakt door: CBS, Centerdata, CLARIAH, DANS, KNAW Humanities Cluster, ODISSEI en RKD.

INHOUD



E-data & Research liever (tijdelijk) op een ander adres ontvangen? Geef wijzigingen door via edata@dans.knaw.nl.

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

100.000 volksverhalen uit Nederland en Vlaanderen online

De Volksverhalenbank van de Lage Landen

Twee Nederlandstalige volksverhalen-databases zijn samengegaan en bevatten nu 100.000 volksverhalen uit Nederland en Vlaanderen.

Theo Meder

De Nederlandse Volksverhalenbank werd in 1994 opgezet door verhalenonderzoeker Theo Meder (Meertens Instituut, KNAW Humanities Cluster). Het betrof aanvankelijk een stand-alone database, maar in 2004 ging een nieuw ontwikkelde versie online. Stefaan Top, hoogleraar Volkskunde aan de KU Leuven van 2002 tot 2011, liet zich door deze database inspireren toen hij de Vlaamse Volksverhalenbank opzette. Hierdoor vertoonden de twee databases van het begin af aan veel gelijkenis: allebei leverden ze volksverhalen en metadata zoals plaats en tijd van vertellen, een korte samenvatting, trefwoorden, subgenre, regionale taal, schriftbron et cetera.

Volkskunde

Het initiatief voor de Vlaamse Volksverhalenbank kwam vanuit het Seminarie voor Volkskunde van de KU Leuven, en naast initiatiefnemer Stefaan Top werkten daar ook Katrien van Effelterre en Chris Corneillie aan mee. Toen Stefaan Top in 2006 met emeritaat ging, dreigde het vak volkskunde in Leuven te verdwijnen; Top is toen nog tot 2011 als bijzonder gasthoogleraar blijven doceren, maar dit heeft uiteindelijk niet kunnen verhinderen dat volkskunde toch werd opgeheven. Mede hierdoor was de Vlaamse volksverhalenbank al lange tijd niet onderhouden en aangevuld.

Sprookjes en sagen

Omdat de twee databases zo compatibel zijn, is het samenvoegen een betrekkelijk eenvoudig proces geweest. Wel is het materiaal anders van aard. De Nederlandse verhalenbank is qua tijdsspanne veel ruimer van opzet en loopt van de middeleeuwen tot heden. De Vlaamse verhalenbank bevat verzamelingen sagen uit de periode 1943-2003. Sagen zijn betrekkelijk korte 'geloofsverhalen' over het bovennatuurlijke. Ze gaan over spoken, weerwolven, kabouters, tovenaars, heksen, duivels, meerminnen, plaaggeesten en dergelijke. De Nederlandse Volksverhalenbank heeft meer subgenres opgenomen zoals: sprookjes, sagen, legenden, raadsels, moppen en broodjeaapverhalen. In de beginjaren zijn er ook Vlaamse sprookjes opgenomen, maar zodra de Vlamingen aan hun eigen database begonnen, is de invoer van Vlaams materiaal in Nederland gestaakt.

Geluidsopnames

Zowel de Nederlandse als de Vlaamse verhalenbank beschikken soms over (fragmenten van) geluidsopnames die in de twintigste eeuw gemaakt zijn. Die opnames zijn gedigitaliseerd naar MP3-formaat en zijn steeds aan de betreffende verhalen toegevoegd. De

'Deze database bevat straks sagen in het Nederlands, Fries, Vlaams, Duits, Deens en IJslands'

Vlaamse sagen zullen ook worden opgenomen in de internationale sagen-database 'Intelligent Search Engine for Belief Legends' (ISEBEL). Daarmee bevat deze database straks sagen in de volgende talen: Nederlands, Fries, Vlaams, Duits, Deens en IJslands. Het gaat daarbij in totaal om zo'n 140.000 sagen.

https://www.verhalenbank.nl/

Credits: Pixabay (vervaardigd met AI)

Project CLARIAH-Plus afgerond

CLARIAH verder als organisatie

In juni loopt het project CLARIAH-Plus af. De onderzoeksinfrastructuur voor de geesteswetenschappen wordt vanaf dan beheerd door CLARIAH-NL.

Erica Renckens

Deze zomer betekent het einde van een tijdperk aan CLARIAH-projecten. Ruim dertien jaar is er dan gewerkt aan de ontwikkeling van een digitale onderzoeksinfrastructuur voor de geesteswetenschappen. Het begon in 2009 met het project CLARIN-NL, waarna vanaf 2014 drie CLARIAH-projecten volgden: SEED, Core en Plus. "De ontwikkelfase van een deel van de infrastructuur wordt daarmee tijdelijk afgesloten, nu volgt de beheerfase", vertelt Dirk van Miert, PI van CLARIAH-Plus.

Voortbestaan

De ontwikkelfase van CLARIAH-Plus werd officieel afgesloten tijdens de CLARIAH-conferentie op 13 juni. "Toen hebben we de portal INEO gepresenteerd, waarin je tools, data, standaarden, workflows en educatief materiaal kunt vinden. Alles wat in CLARIAH zelf is ontwikkeld en nog heel veel extra."

Ook is er uitgebreid stilgestaan bij de toekomst van CLARIAH. "Begin dit jaar is SSHOC-NL van start gegaan, het project waarin CLARIAH samenwerkt met ODISSEI, de on-

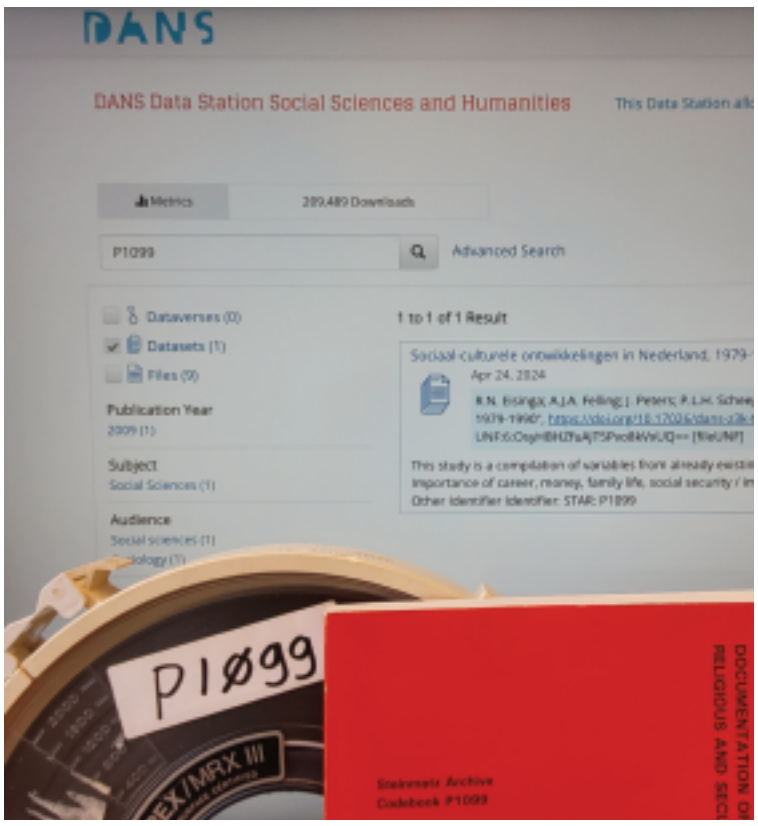
Vervolg op pagina 2

Het DANS Data Station SSH bouwt voort op een rijke traditie

Terugblik op 60 jaar Open Science

Op het moment van publicatie van dit magazine is het een jaar geleden dat het DANS Data Station Social Sciences and Humanities (SSH) werd gelanceerd. Dit Data Station bouwt voort op een respectabele traditie van 60 jaar data-archivering in de sociale wetenschappen in Nederland.

Marion Wittenberg



Van tape naar online. Steinmetzarchief databestanden nog steeds beschikbaar. *Credits: Marion Wittenberg*

In november van dit jaar is het 60 jaar geleden dat de Steinmetz Stichting werd opgericht. In 1972 werd de stichting onderdeel van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) en kreeg het de naam Steinmetz-archief. De missie van dit archief was '… het verzamelen en voor sekundaire analyse toegankelijk maken van grondmateriaal van sociaal-wetenschappelijk onderzoek in de ruimste zin van het woord'. In de beginjaren kwamen de bestanden binnen op ponskaarten, vergezeld van papieren codeboeken. Elk databestand kreeg een uniek archiefnummer (P-nummer) en werd zorgvuldig gedocumenteerd. Een deel van deze metadata werd gepubliceerd in een gedrukte catalogus. De data werden, samen met 'machineleesbare' codeboeken, verspreid op computertapes die per post werden verzonden.

Hergebruik

Het Steinmetz-archief diende niet alleen als een opslagplaats voor gegevens; de databestanden werden actief hergebruikt voor methodologisch onderzoek, heranalyse en cross-nationaal onderzoek. Een aantal van deze databestanden vormde de basis voor het onderzoek 'Culturele Veranderingen' van het Sociaal Cultureel Planbureau (SCP). De oorspronkelijke metingen van dit longitudinale onderzoek, gestart in 1975 en sindsdien (twee)jaarlijks uitgevoerd, waren gebaseerd op 200 onderzoeksvragen afkomstig uit 17 oudere databestanden van het Steinmetz-archief. Door de vragen exact te herhalen, konden onderzoekers veranderingen in de opvattingen van de Nederlandse bevolking in de loop der tijd meten. Het onderzoek 'Culturele Veranderingen' wordt nog steeds voortgezet en is nu toegankelijk via het DANS Data Station SSH.

Internationaal

Vanaf het begin stond internationale samenwerking centraal in de ontwikkeling van het Steinmetz-archief, dat samen met dataarchieven uit Engeland, Duitsland en Noorwegen, een essentiële rol speelde bij de oprichting van het Consortium of European Social Science Data Archives (CESSDA). Gezamenlijk werkten zij aan het ontwikkelen van standaarden voor de archivering en documentatie van data. Via CESSDA kregen deze archieven toegang tot datacollecties uit andere landen, waardoor cross-nationaal onderzoek mogelijk werd gemaakt.

Het Steinmetz-archief werd in 2005 geïntegreerd in DANS, waarbij de volledige collectie nu is opgenomen in het Data Station SSH. Deze uitgebreide collectie is toegankelijk via https://edu.nl/thupy/.

Het huidige DANS Data Station SSH zet deze internationale erfenis voort als een centrum voor data-archivering en -publicatie, en als een waardevolle bron voor andere onderzoekers.

https://ssh.datastations.nl

CLARIAH

Vervolg van pagina 1

derzoeksinfrastructuur voor de sociale wetenschappen", aldus Van Miert. "Daarnaast willen ook, weer samen met ODISSEI, een nieuw projectvoorstel indienen voor NWO's Roadmap voor grootschalige wetenschappelijke infrastructuur."

Daarvoor is het wel noodzakelijk dat CLARIAH blijft bestaan. Van Miert: "We gaan daarom niet door als project, maar als een consortiumorganisatie: CLARIAH-NL." Deze organisatie zal blijvend de ontwikkeling, het beheer en het gebruik van de digitale infrastructuur op nationaal niveau coördineren en stimuleren.

Alle organisaties die betrokken zijn bij de geesteswetenschappen kunnen tegen een jaarlijks bedrag lid worden van CLARIAH-NL. "De Nederlandse faculteiten Geesteswetenschappen hebben al aangegeven lid te worden. Daarnaast is er plek voor instellingen op het gebied van onderzoek, erfgoed en/of infrastructuur. Denk bijvoorbeeld aan de KNAW-instituten, musea, de Koninklijke Bibliotheek, het eScience Center en SURF."

Inspraak

De leden adviseren het managementteam, dat het dagelijks bestuur van de organisatie voor zijn rekening zal nemen. Een supervisory board, bestaande uit afgevaardigden van de leden, bewaakt de doelstellingen van CLARIAH-NL en heeft de eindverantwoordelijkheid.

Waarom zou een organisatie lid moeten worden van CLARIAH-NL? "Je zit aan tafel op het moment dat er nieuwe landelijke infrastructurele projecten worden ontwikkeld. Dat doen we als CLARIAH-NL ten behoeve van onderzoekers en instellingen en dan is het wel zo fijn om mee te praten over waarover het moet gaan en wat nodig is. Echt co-creatie."

Daarnaast hebben leden inspraak op het beheer van de bestaande infrastructuur, aldus Van Miert. "Van het budget zullen we niet de complete bestaande infrastructuur voor altijd in de lucht kunnen houden. Dat betekent dat we life cycle-keuzes moeten maken: wat ga je afschalen en wat ga je doorontwikkelen? Dat zijn soms lastige beslissingen en dan is juist fijn als de gebruikers meedenken."

https://www.clariah.nl/

GEHOORD & BIJGEWOOND

CBS-community dag bij Amsterdam UMC

Hilde van Oirschot

Op 1 februari 2024 organiseerde Amsterdam UMC een symposium over het koppelen van eigen (cohort) data aan CBS microdata. Dit was tevens de kick-off voor de oprichting van een CBS-community. Het symposium vond plaats in de Amstelzaal van het VUmc. Zo'n 200 geïnteresseerden hadden zich aangemeld.

Het symposium richtte zich specifiek op de mogelijkheden om eigen (medische) cohorten te koppelen aan microdata ten behoeve van het verbeteren van (patiënten)zorg en maatschappelijke gezondheid.

Na een introductie over het werken met microdata werd ingegaan op de ethische kant van het koppelen van medische cohorten aan CBS microdata. Tot slot werden resultaten van drie onderzoeken gepresenteerd waarbij het koppelen van cohort data aan CBS microdata tot nieuwe inzichten heeft geleid en tot aanpassingen in zorgbeleid. Bekijk het filmpje:

https://vimeo.com/911190315/cf6ccc9cb4?share=copy

COLOFON **Uitgever**: E-data & Research. **Redactieadres:** Anna van Saksenlaan 51, 2593 HW Den Haag, 088 0034666, edata@dans.knaw.nl, edata.nl, @eDataResearch. **Hoofd-/eindredacteur**: Ingrid Korver. **Redactie**: Ricarda Braukmann, Mathilde Jansen, Kasia Karpinska, Astrid de Kock, Hilde van Oirschot, Erica Renckens, Jetze Touber, Mara Verheijen. **Redactiesecretariaat**: Ingrid van der Elst. **Aan dit nummer werkten mee**: Alastair Dunning, Valentijn Gilissen, Ivo Gorissen, Andrea Scharnhorst, Amber Vangerven, Samantha Willemsen. **Opmaak**: Colette Sloots, Haarlem. **Productie**: Platform P, Rotterdam.
Oplage: 4200 papier, 6000 digitaal. **ISSN**: 1872-0374. We hebben getracht alle belanghebbenden met betrekking tot het gebruikte beeldmateriaal te benaderen. Degenen die menen recht te kunnen doen gelden, kunnen zich tot ons wenden. E-data & Research kosteloos ontvangen? Laat het weten via edata.nl. E-data & Research liever (tijdelijk) op een ander adres ontvangen? Geef wijzigingen aan ons door via edata@dans.knaw.nl.

Het samenbrengen van sociaalwetenschappelijke onderzoeksinfrastructuren

Infra4NextGen brengt Europese initiatieven samen

Sociaalwetenschappelijke onderzoeksinfrastructuren in Europa beschikken over een schat aan gegevens, relevant voor beleidsmakers en onderzoekers. Deze gegevens zijn echter momenteel verspreid en soms lastig toegankelijk.

Mara Verheijen

Europese vlaggen. *Credits: Pixabay*

Een nieuw project (Infra4NextGen, gefinancierd door de Europese Commissie) brengt de belangrijkste sociaalwetenschappelijke initiatieven van Europa samen. De resultaten hiervan zullen van nut zijn voor zowel beleidsmakers als academici.

Tijdens het 4-jarige project Infra4NextGen worden de uitkomsten van belangrijke sociaalwetenschappelijke onderzoek infrastructuren samengebracht. Infra4NextGen zal bestaande onderzoeksdiensten een nieuwe bestemming geven en aanpassen ter ondersteuning van de vijf thema's van het NextGenerationEU programma. Deze thema's zijn (1) Ga voor groen, (2) Digitaal, (3) Gezond, (4) Sterk en (5) Gelijk. Hiermee streeft NextGenerationEU ernaar dat Europa een groenere, veerkrachtigere en meer digitale toekomst opbouwt.

Organisatie en samenwerking

Het project wordt gecoördineerd door het European Social Survey European Research Infrastructure Consortium (ESS ERIC). In het kernteam zitten CESSDA ERIC, de European Values Study (EVS) en het Gender and Generations Programme (GGP).

In het reeds bestaande CROss-National Online Survey (CRONOS) Panel worden data verzameld en vragen gesteld over de NextGenEU-thema's. Dit panel zal voor de eerste keer met de DataCTRL tooling, ontwikkeld door Centerdata, worden gehost. Zo zal Centerdata de vragenlijsten programmeren, worden de panels van de verschillende Europese landen beheerd met de Survey CTRL tool en verwerken we de data.

Dataportaal

Geharmoniseerde en samengevoegde uittreksels uit bestaande datasets, zullen worden geproduceerd door GESIS – Leibniz Instituut voor Sociale Wetenschappen (vertegenwoordiger van CESSDA ERIC), om hiermee de last voor analisten te verminderen en de steekproefomvang te vergroten. De bestaande data uit deze onderzoeken worden vervolgens geanalyseerd en samengevat in een reeks beleidsrelevante tabellen en visualisaties met commentaar. Deze zullen worden gepresenteerd op een speciaal onlineportaal.

Deze initiële analyse zal later in het project worden aangevuld met nieuwe data die over elk onderwerp worden verzameld via het online CRONOS-webpanel. Dit zal gebeuren in vijf fases in elf landen; Oostenrijk, België, Tsjechië, Finland, Frankrijk, Hongarije, IJsland, Portugal, Slovenië, Zweden en Groot-Brittannië. De weging van de panelgegevens zal worden uitgevoerd door het Instituut voor Sociaal en Economisch Onderzoek van de Universiteit van Essex (vertegenwoordiger van ESS ERIC).

Nieuwe data over elk NextGenEU-thema zullen worden verzameld en snel beschikbaar worden gesteld via het gevestigde CROss-National Online Survey (CRONOS) Panel. Dit panel wordt beheerd door Centerdata. Sikt – Noors agentschap voor gedeelde diensten in onderwijs en onderzoek (vertegenwoordiger van ESS ERIC) zal ervoor zorgen dat de CRONOS-data worden verwerkt en gemakkelijk toegankelijk zijn voor de onderzoeksgemeenschap via een speciaal dataportaal met alle data georganiseerd per NextGenerationEU-thema.

Meer lezen over Infra4NextGen, zie: ***https://www.centerdata.nl/actueel/nieuw-project-infra4nextgen en https://next-generation-eu.europa.eu***

Dutch Data Prize

In 2024 wordt de Dutch Data Prize voor de achtste keer uitgereikt aan een individu of een team dat onderzoeksdata FAIR maakt. De prijs is een waardevolle erkenning voor de bijdragen van onderzoekers aan hun eigen vakgebied en aan het principe van FAIR (Findable, Accessible, Interoperable, Reusable) data.

Samantha Willemsen

De Dutch Data Prize (voorheen Nederlandse Dataprijs) is onderdeel van Research Data Nederland (RDNL), een consortium bestaande uit 4TU.ResearchData, DANS, Health-RI en SURF, en wordt sinds 2010 om de twee jaar uitgereikt in drie domeinen: Social Science and Humanities, Natural and Engineering Sciences en Life Sciences and Health Science.

Eigen dataset nomineren

Tot en met half augustus is het mogelijk om een eigen dataset te nomineren of een dataset die door een ander individu of onderzoeksgroep is geproduceerd. Dit kan via de website van RDNL (www.researchdata.nl). Aan de hand van criteria bepaalt een zevenkoppige jury - een voorzitter en twee juryleden per domein - welke drie datasets per domein - in totaal negen - het halen tot de finale. Op 17 oktober presenteren de negen finalisten hun pitch, gevolgd door de prijsuitreiking. Het evenement vindt plaats in combinatie met de FAIR-IMPACT National Roadshow. Deze twee versterken elkaar door hun focus op FAIR. Waar tijdens de roadshow sprekers de laatste ontwikkelingen op dit vlak presenteren en er workshops worden gegeven hoe dit toe te passen in de praktijk, heeft de awardshow de focus op de uitvoering van het principe.

Vertaalmachines

Maarten Marx is één van de eerste winnaars van de Nederlandse Dataprijs, die zijn tijd ver vooruit was met het FAIR maken van data. Meedoen aan de Dutch Data Prize was voor Marx een controle of de FAIR principes goed waren toegepast. "Door jezelf op te geven, moet je een aantal zaken controleren en dat is ook meteen een check op je werk. Ik vond dat al heel waardevol." Het FAIR principe is iets waar hij als informaticus vanaf het begin af aan al het nut van inzag. "Het argument dat het veel werk is, vind ik te verwaarlozen. Natuurlijk is het meer werk, maar het levert de wetenschap echt iets op. Je weet nooit waar jouw data later voor gebruikt kan worden. Zo hebben de notulen van het Europees Parlement ervoor gezorgd dat we nu vertaalmachines hebben." Marx motiveert daarom iedereen om de tijd te investeren om data FAIR te maken: "Als het bij de bron meteen goed wordt aangepakt, kost het weinig tijd en is het van onschatbare waarde."

Naast de award en de erkenning is er prijzengeld om data meer FAIR te maken en hergebruik van data aan te moedigen. De winnaars kunnen hiermee bijvoorbeeld een symposium organiseren of hun data online beter toegankelijk maken.

Op de website www.researchdata.nl is alle informatie te vinden hoe een dataset te nomineren. Meer informatie over de FAIR-IMPACT National Roadshow is te vinden op https://fair-impact.eu

SINDS KORT BESCHIKBAAR

Dit overzicht toont databestanden die recent beschikbaar zijn gekomen bij CBS, Centerdata en DANS.

CBS

• AANVULVERGOEDPERSOONMND-BEDRAGBUS

Dit bestand vertegenwoordigt alle personen die in een bepaald jaar van hun (ex-)werkgever ontslagvergoeding/transitievergoeding of een aanvulling op hun uitkering werknemersverzekering ontvangen. Bestand per jaar beschikbaar over de periode 2021-2022.

• PENSNABESTPERSOONMND-BEDRAGBUS

Dit bestand vertegenwoordigt alle personen van 55 jaar of ouder die in een bepaald jaar Nabestaandepensioen hebben ontvangen. Bestand per jaar beschikbaar over de periode 2021-2022.

• PENSPIJLER2PERSOONMND-BEDRAGBUS

Dit bestand vertegenwoordigt alle personen die in een bepaald jaar tweede pijler pensioen ontvangen. Het gaat hierbij grotendeels om pensioenen die werknemers bij een pensioenfonds hebben opgebouwd. Andere vormen van tweede pijler pensioen die zijn meegenomen omvatten uitkeringen in het kader van de vitaliteitsregeling en uitkeringen in het kader van de Algemene Pensioenwet Politieke Ambtsdragers (APPA). Bestand per jaar beschikbaar over de periode 2021-2022.

• PENSPIJLER3PERSOONMND-BEDRAGBUS

Dit bestand vertegenwoordigt alle personen die in een bepaald jaar derde pijler pensioen, oftewel lijfrentes, ontvangen. Bestand per jaar beschikbaar over de periode 2021-2022.

• GEWASPERCELEN

Het bestand bevat een opsomming van alle in gebruik zijnde gewaspercelen op Nederlandse agrarische bedrijven op peildatum 15 mei van het jaar. De percelen met de daarop geteelde gewassen zijn door de bedrijven geregistreerd in het perceelsregister van de Rijksdienst voor ondernemend Nederland (RVO). Het perceelsregister is verplicht voor subsisdies in het kader van het Gemeenschappelijk Landbouw Beleid (GLB). Bestanden beschikbaar over de periode 2017-2022.

Deze publicaties zijn beschikbaar via www.cbs.nl/microdata Bezoek deze site of scan de QR-code.

Centerdata

• Nationale Monitor geldzorgen

Bron: AdobeStock

In de vorige versie van E-data (februari 2024) stond een stuk beschreven over de nationale monitor geldzorgen. Met deze monitor brengen onderzoekers de financiële situatie van Nederlanders in kaart en achterhalen ze waar Nederlanders zich zorgen over maken. Hierbij kijken ze ook naar financiële weerbaarheid. Resultaten van nieuwe metingen worden elk kwartaal bekendgemaakt op de website van Wijzer in geldzaken.

lissdata.nl

Geldzorgen. *Credits: Pixabay*

De data wordt ook beschikbaar gesteld via LISS Data Archive. Meting van september 2023 –
DOI: https://doi.org/10.57990/h1rw-yg28
Meting van december 2023 –
DOI: https://doi.org/10.57990/1zcd-nk89

Ook sinds kort beschikbaar:

Studies LISS panel

- Brandt, M., Turner-Zwinkels, F – September 2019 – February 2020 – Attitudes and opinions (all 12 waves).
DOI: https://doi.org/10.57990/99c8-3994
- Centerdata – April 2023 – Work and Schooling (Wave 16).
DOI: https://doi.org/10.57990/vvb5-4c11
- Centerdata – July 2023 – Economic Situation: Housing (Wave 16).
DOI: https://doi.org/10.57990/vq4j-2g10
- Centerdata – October 2023 – Social Integration and Leisure (Wave 16).
DOI: https://doi.org/10.57990/2gze-ha88
- Centerdata – November 2023 – Health (Wave 16).
DOI: https://doi.org/10.57990/evk0-mk06
- De Werd, R., van Dalen, H – July 2023 – Population trends.
DOI: https://doi.org/10.57990/mfap-3g80
- Ten Kate, J., de Koster, W., van der Waal, J – February 2022 – Discrete choice experiment (part 1).
DOI: https://doi.org/10.57990/02qq-c015
- Ten Kate, J., de Koster, W., van der Waal, J – March 2022 – Discrete choice experiment (part 2).
DOI: https://doi.org/10.57990/7vrq-tp94
- Van Hoboken, J – February 2020 – Experience with illegal internet behavior.
DOI: https://doi.org/10.57990/xppk-qw59

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

DANS

De volgende datasets zijn open access beschikbaar in de DANS Data Stations:

- R. Keeton; M. Abu; S. Boly; E. Gurmu; B. Ketsomboon; J. Owour; D. Reckien (Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente), 2024, 'Database of perceived cause-effect relations in migration aspirations and decisions', https://doi.org/10.17026/SS/PWKZME, DANS Data Station Social Sciences and Humanities, V2
- L. van Wissen; N. Vriend; A. Nijssen; L. Vereecken; M. den Engelse (Noord-Hollands Archief), 2024, 'FAIR Photos - CLARIAH FAIR Data Call 2023', https://doi.org/10.17026/SS/7VD3ME, DANS Data Station Social Sciences and Humanities, V1
- D. Farace; S. Biagioni; C. Carlesi (GreyNet International), 2024, 'Realizing the potential of grey literature by recognizing its publishers', https://doi.org/10.17026/SS/GWIFFK, DANS Data Station Social Sciences and Humanities, V1
- B. van Dijk; M.J. van Duijn; S. Verberne; M.R. Spruit (Leiden University), 2024, 'ChiSCor: Children's Story Corpus', https://doi.org/10.17026/SS/TGPDJF, DANS Data Station Social Sciences and Humanities, V2
- E. Bouw (Hogeschool Utrecht), 2024, 'Replication Data for: Onderwijs over disciplinaire grenzen heen', https://doi.org/10.17026/SS/NN2U39, DANS Data Station Social Sciences and Humanities, V1
- D. Brukx; T. Termorshuizen; J. Lodewick (ResearchNed; Ministerie van OCW), 2024, 'LAKS-monitor 2023', https://doi.org/10.17026/SS/6MVUR5, DANS Data Station Social Sciences and Humanities, V1
- S. van Dorsselaer (Department of ISW: University Utrecht; Trimbos-instituut), 2024, 'Jong na Corona: Welzijn van jongeren 2022 en inzet van NP Onderwijsmiddelen door scholen', https://doi.org/10.17026/SS/LDM0TJ, DANS Data Station Social Sciences and Humanities, V2
- S. Stigter; M. Bundgaard (University of Amsterdam), 2024, 'Interview Mark Manders on Room, Constructed to Provide Persistent Absence', https://doi.org/10.17026/SS/V49Q09, DANS Data Station Social Sciences and Humanities, V2
- P. Bitter (Gemeente Alkmaar), 2024, 'Opgraving Alkmaar Koorstraat 'Gulden Vlies' 2022 (22KST)', https://doi.org/10.17026/AR/FTQNCH, DANS Data Station Archaeology, V2
- K. Pepers (Aeres University of Applied Sciences), 2024, 'RhoC validation data from 'Validation of a new Soil Bulk Density sensor", https://doi.org/10.17026/PT/BYVPLB, DANS Data Station Physical and Technical Sciences, V1
- A. Ural-Janssen; C. Kroeze; E. Meers; M. Strokal (Wageningen University & Research), 2024, 'Data from: Large reductions in nutrient losses needed to avoid future coastal eutrophication across Europe', https://doi.org/10.17026/PT/OUJR9H, DANS Data Station Physical and Technical Sciences, V1
- M. Broekman; J. Hilbers; S. Hoeks; M. Huijbregts; A. Schipper; M. Tucker (Radboud University), 2024, 'Data of: Environmental drivers of global variation in home range size of terrestrial and marine mammals', https://doi.org/10.17026/LS/GICIFU, DANS Data Station Life Sciences, V2
- M. Berger; M. van Hartingsveldt; H. van Dongen; L. Valent; M. Sol; K. van Dijk; W. Dirks; C. Doorenbosch (De Haagse Hogeschool; Hogeschool van Amsterdam; Hogeschool Utrecht; Vrije Universiteit Amsterdam), 2024, 'OPTIMA - Naar een optimale match tussen kind en rolstoel', https://doi.org/10.17026/LS/E1XH3X, DANS Data Station Life Sciences, V2
- B. Dejaco (HAN), 2024, 'Repliaction data for the study on Experiences of Physiothera-pists considering VR for shoulder rehabilitation', https://doi.org/10.17026/LS/COWIPU, DANS Data Station Life Sciences, V2

Bezoek de DANS Data Stations via dans.knaw.nl/nl/data-stations/ of scan de QR-code.

AGENDA

12 juni • Den Haag
NPSO jaarlijkse dag
Het thema dit jaar is non-respons en lastig bereikbare groepen.
npso.net

13 juni • Leiden
CLARIAH Conferentie 2024
Ontdek de impact van CLARIAH in de geesteswetenschappen! Doe mee met hands-on sessies, verken het INEO-portaal en krijg inzicht in de CLARIAH-NL consortiumorganisatie.
clariah.nl/events/clariah-conference-2024

17 oktober • Den Haag
FAIR IMPACT National Roadshow
Gericht op de RDM gemeenschap, data support staff, onderzoekers en beleidsmakers van universiteiten en/of onderzoeksinstituten.
fair-impact.eu/events/national-road-shows

17 oktober • Den Haag
Uitreiking Dutch Data Prize 2024
Tot medio augustus kun je een (eigen) dataset nomineren. Ga hiervoor naar www.researchdata.nl. Op 17 oktober presenteren negen finalisten hun pitch.
www.researchdata.nl

22 oktober • Maastricht
National Open Science Festival
Dit jaar zijn de Universiteitsbibliotheek van Maastricht en Open Science Community Maastricht de gastheren.
openscience.nl/en/open-science-festival

14 november • Den Haag
DANS Open Dag
Thema van deze dag is Open Science. Met workshops over citeren van data, machine-leesbare licenties, toepassen van FAIR-principes, etc.
dans.knaw.nl

PiCo maakt koppeling historische persoonsgegevens mogelijk

Nieuwe standaard CBG voorziet in behoefte

Het Centraal Bureau voor Genealogie (CBG) ontwikkelde een standaard voor het registreren van persoonsgegevens voor het optimaliseren van genealogisch onderzoek. Als Linked Data-standaard is deze breed inzetbaar.

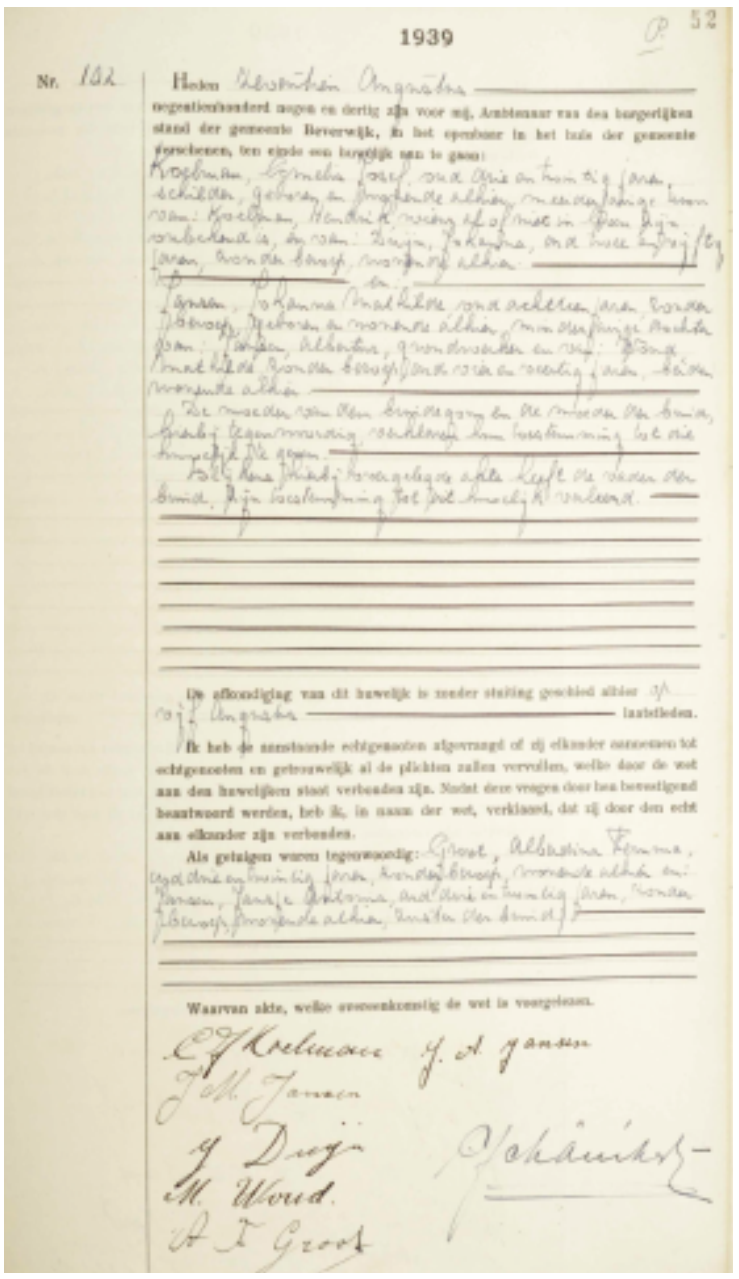
Mathilde Jansen

Linked Data-standaarden voor persoonsgegevens waren er wel al, maar tot nu toe waren deze allemaal geënt op het heden, vertelt Pieter Woltjer, datamanager bij het CBG. Een standaard voor historische persoonsgegevens was er nog niet, en daar heeft het CBG nu verandering in gebracht. “Met Schema bijvoorbeeld, de Linked Data-standaard die gebruikt wordt door Google en Facebook, kun je wel vastleggen dat iemand een partner heeft, maar niet meerdere in de tijd. En je kunt wel een geboortedatum vastleggen, maar geen leeftijd op een bepaald moment in de tijd. Terwijl dat typerend is voor historische bronnen. Dus historische bronnen kun je er moeilijk mee beschrijven.”

Maar nu is er dus PiCo, die voorziet in een behoefte die er ook was bij andere instellingen. Zo zijn het Instituut voor Internationale Sociale Geschiedenis (IISG) en WO2NET betrokken geweest bij de reviewfase omdat zij ook te maken hebben met historische persoonsgegevens. “Doordat PiCo is gebaseerd op Linked Open Data en op bestaande internationale standaarden kunnen andere instellingen makkelijk eigenschappen toevoegen. Dus WO2NET kan gebeurtenissen toevoegen die specifiek zijn voor de Tweede Wereldoorlog en het IISG inkomensgegevens, als ze dat zouden willen.”

Persoonsreconstructie

Het model is in eerste instantie ontwikkeld voor de website WieWasWie van het CBG, waar iedereen terechtkan voor stamboomonderzoek. Op dit platform zijn heel veel Nederlandse archieven samengebracht. Dit heeft als voordeel dat iedereen die zijn familiegeschiedenis wil uitzoeken, niet allemaal verschillende archieven af hoeft.


1939

Nr. 182

Heden [illegible] negentienhonderd negen en dertig zijn voor mij, Ambtenaar van den burgerlijken stand der gemeente Beverwijk, in het openbaar in het huis der gemeente verschenen, ten einde een huwelijk aan te gaan:

[illegible] oud [illegible] jaren, schilder, geboren en woonachtig alhier, meerderjarige [illegible] van: [illegible] Hendrik, weduwnaar of man in [illegible] verbonden is, en van: [illegible], Jakobina, oud [illegible] jaren, wonende [illegible], zonder beroep, meerderjarige alhier

en

[illegible] Johanna Mathilde, oud [illegible] jaren, zonder beroep, geboren en woonachtig alhier, meerderjarige dochter van: [illegible], Albertus, grondwerker en wijlen [illegible] Mathilde, wonende [illegible], ongehuwde [illegible] oud [illegible] jaren, zonder beroep, meerderjarige alhier

De meerdere [illegible] van den bruidegom en de moeder der bruid, hebben bij [illegible] verklaring van toestemming tot dit huwelijk gegeven.

Het [illegible] huwelijk [illegible] heeft de vader der bruid, bij [illegible] verklaring tot dit huwelijk verleend.

De afkondiging van dit huwelijk is reeds stelling gehad alhier op [illegible] laatstleden.

Ik heb de aanstaande echtgenoten afgevraagd of zij elkander aannemen tot echtgenoten en getrouwelijk al de plichten zullen vervullen, welke door de wet aan den huwelijken staat verbonden zijn. Nadat deze vragen door hen bevestigend beantwoord werden, heb ik, in naam der wet, verklaard, dat zij door den echt aan elkander zijn verbonden.

Als getuigen waren tegenwoordig: Groot, Albertina Femma, oud [illegible] en twintig jaren, [illegible], woonachtig alhier en Jansen, Jantje Antonia, oud [illegible] en twintig jaren, [illegible], [illegible], woonachtig alhier, vrienden der bruid.

Waarvan akte, welke overeenkomstig de wet is voorgelezen.

C. J. Kelmaar J. A. Jansen

J. M. Jansen

J. Duijn

M. Wout

A. T. Groot

Burgerlijke stand van de gemeente Beverwijk, archiefdeel van (dubbele) registers van de huwelijksakten van de gemeente Beverwijk, 1939. *Credits: Noord-Hollands Archief, Wiewaswie.nl*

Om de gegevens van al die archieven op een eenduidige manier bij elkaar te krijgen, heeft het CBG tien jaar geleden de standaard A2A ontwikkeld. PiCo is de opvolger, legt Woltjer uit: “Het grote verschil is dat A2A een standaard was puur gericht op losse persoonsvermeldingen: alle gegevens die een persoon identificeren zoals ze in diverse bronnen staan vermeld zoals aktes van de burgerlijke stand of notariële archieven. In PiCo brengen we ze samen.”

Op WieWasWie staan vandaag de dag 220 miljoen persoonsvermeldingen, vertelt Woltjer. Dat zijn dus allemaal losse persoonsgegevens. “Mijn overgrootvader staat daar minstens drie keer in: via zijn geboorteakte, zijn huwelijksakte en zijn overlijdensakte. Misschien zelfs wel tien keer. A2A bood nog geen manier om die vermeldingen te koppelen. Archieven leverden tot nu toe data aan zoals ze op de bron vermeld stonden. Maar vandaag de dag zeggen we: als we die persoon tien keer hebben, dan willen wij een reconstructie van die persoon maken, die verwijst naar alle tien die bronnen. Dat kan bijvoorbeeld met algoritmes of crowdsourcing. Naar zo’n persoonsreconstructie werken we nu toe.”

‘Wanneer je straks zoekt naar je overgrootvader, kom je hem maar één keer tegen’

WieWasWie 2.0

PiCo is dus niet alleen een standaard voor het registeren van persoonsvermeldingen, maar ook voor reconstructies die gemaakt zijn op basis van de persoonsvermeldingen. “Met PiCo zijn we nu een nieuwe WieWasWie aan het ontwikkelen, die gebaseerd is op persoonsreconstructies. Dat betekent dat wanneer jij straks zoekt naar jouw overgrootvader, je hem niet tien keer tegenkomt maar één keer, met een verwijzing naar alle bronnen waarin hij voorkomt. In de ideale wereld hoef je dan niet meer verder te zoeken. PiCo maakt het bovendien mogelijk om relaties te leggen tussen personen, dus je vindt dan ook meteen de partner van je overgrootvader en zijn kinderen enzovoorts.”

Er hoort wel een disclaimer bij, waarschuwt Woltjer. “We gaan dit soort reconstructies maken, maar we weten van tevoren niet of dat ook voor alle personen in onze WieWasWie-database gaat lukken. Misschien komt jouw overgrootvader wel in een andere akte voor, maar is zijn naam daarin heel anders gespeld. Dan is het voor een algoritme heel moeilijk om te zeggen: dat is hem.” Om toch uiteindelijk die koppelingen te maken, wil het CBG ook gebruik gaan maken van de kennis van de genealogen die dagelijks hun website bezoeken. Een bezoeker kan dan zelf aangeven dat een afwijkende naamsvermelding ook zijn of haar overgrootvader betreft, bij wijze van spreken.

Om deze nieuwe versie van WieWasWie mogelijk te maken, is het van belang dat alle archieven hun data in PiCo gaan aanleveren. Om dit te bewerkstelligen is het CBG nu in gesprek met de leveranciers van de Nederlandse collectiebeheersystemen. Zodat alle persoonsdata straks op een uniforme manier zijn opgeslagen. Nog mooier zou het zijn als buitenlandse instellingen dit systeem ook overnemen, want hoe meer systemen met elkaar kunnen ‘praten’, hoe makkelijker het wordt om stamboomonderzoek te doen. “Wij willen dat iedereen stamboomonderzoek kan doen. En familiegeschiedenis is bijna nooit beperkt tot de grens. Dus hoe meer landen het op een vergelijkbare manier ontsluiten, hoe makkelijker het wordt om ook je internationale roots terug te vinden.”

Wie geïnteresseerd is of meer wil weten over PiCo kan contact opnemen met Pieter Woltjer, pieter.woltjer@cbg.nl

https://www.wiewaswie.nl/
https://github.com/CBG-Centrum-voor-familiegeschiedenis/PiCo/

KORT

DANS ontwikkelt richtlijn publiceren datasets met beperkte toegang

Veel waardevolle datasets, met name in de sociale wetenschappen, bevatten persoonsgegevens en kunnen daarom niet vrij toegankelijk worden gemaakt. Om hergebruik van deze gegevens toch mogelijk te maken, bieden archieven zoals DANS opties voor beperkte toegang. In de DANS Data Stations kunnen gebruikers dergelijke ‘restricted access’ datasets pas downloaden na toestemming van de eigenaar.

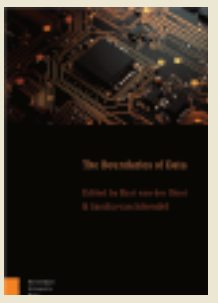
Ter ondersteuning van data-eigenaren in het publiceren van data met beperkte toegang, heeft DANS een specifieke richtlijn ontwikkeld. Deze richtlijn begeleidt gebruikers door alle stappen die nodig zijn voor het voorbereiden van het deponeren. Tevens bevat het tips voor het opstellen van een protocol dat de voorwaarden voor hergebruik beschrijft. Door deze informatie zorgvuldig te documenteren, wordt duidelijk wie onder welke voorwaarden toegang tot de data mag krijgen. De richtlijn is beschikbaar op Zenodo (RB).

https://doi.org/10.5281/zenodo.10887484
https://mlcommons.org/2024/03/croissant_metadata_announce
https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets

GELEZEN

The Boundaries of Data
Bart van der Sloot, Sascha van Schendel (eds), (2024), The Boundaries of Data, Amsterdam University Press

Het boek The Boundaries of Data onderzoekt de uitdagende vraag welk effect huidige en toekomstige technische ontwikkelingen hebben op het gegevensbeschermingskader en de bescherming van de verschillende soorten gegevens met betrekking tot anonimisering, pseudonimisering, aggregatie en identificatie van gegevens. Een breed internationaal gezelschap van deskundigen benadert dit vraagstuk vanuit verschillende perspectieven. In een bijdrage van het CBS leveren Peter-Paul de Wolf, Ivo Gorissen, Daan von Berg en Michel Zaaijer een lezenswaardige bijdrage over de door de CBS-wet gecreëerde mogelijkheden om onderzoekers toegang te bieden tot microdata van het CBS. Dit alles met inachtneming van de hoogst mogelijke bescherming van data gebaseerd op het ‘five safes framework’. (IG)



The Boundaries of Data | Amsterdam University Press (aup.nl)

Datasets makkelijker vindbaar en bruikbaar

Dataverse geschikt maken voor Machine Learning met Croissant

Op 6 maart 2024 kondigde MLCommons (een consortium dat werkt aan standaarden voor kunstmatige intelligentie) de introductie aan van Croissant. Dit is een metadatastandaard, bedoeld om datasets geschikt te maken voor machine learning (ML).
Jetze Touber en Vyacheslav Tykhonov

Door datasets te voorzien van metadata volgens de Croissant-standaard, worden de datasets makkelijker vindbaar en bruikbaar voor allerlei tools en platforms. Dit is belangrijk om kunstmatige intelligentie (AI) en database systemen met elkaar te integreren. Zowel het bedrijfsleven (onder meer Google, Kaggle, Hugging Face) als academische instellingen (onder meer Harvard University, DANS) dragen bij aan de ontwikkeling van Croissant.

Eenvoudiger vindbaar

Data vormen de kern van alle ML- en AI-toepassingen. Tot op heden was er echter nog geen gestandaardiseerde methode voor het organiseren en ordenen van de data en de bestanden waaruit elke dataset bestaat. Als gevolg hiervan was het vinden, begrijpen en gebruiken van datasets voor ML een lastige en tijdrovende klus. Een van de doelstellingen van Croissant is dat datasets makkelijker kunnen worden opgespoord en beoordeeld op hun geschiktheid voor toepassing in ML-systemen.
Het vocabulaire van Croissant is een uitbreiding van schema.org, een machineleesbare standaard om gestructureerde data te beschrijven, gebruikt voor meer dan 40 miljoen datasets op het web. Door voort te bouwen op schema.org worden de datasets vindbaar via specifieke zoekmachines voor datasets, zoals Google Dataset Search.
Croissant is gemakkelijk te implementeren omdat het niet nodig is om de data zelf, of de manier waarop ze worden weergegeven, te veranderen. In plaats daarvan voegt Croissant een laag metadata toe die de inhoud van de dataset op een gestandaardiseerde manier weergeeft en de belangrijkste kenmerken ervan beschrijft.

‘De data hoef je niet te veranderen, er wordt een laag metadata toegevoegd’

Brede implementatie

DANS heeft bijgedragen aan de ontwikkeling van de Croissant-specificatie door input te leveren met betrekking tot FAIR-data management, de herkomst van data en verantwoordelijk AI gebruik. Deze aspecten specificeert Croissant gelaagd in de metadata. De toevoeging van een semantische laag, opgebouwd volgens de FAIR-principes, zal de kwaliteit van data op de lange termijn verbeteren.
DANS zet zich verder, samen met het Harvard Institute for Quantitative Social Science (IQSS), in om de Croissant metadatastandaard in Dataverse op te nemen. Dit betekent dat alle research data repositories gebouwd op de Dataverse software vanaf de volgende release hun metadata kunnen aanbieden in Croissant. De verwachting is dat data geproduceerd door academische en commerciële partijen, waar ze ook zijn opgeslagen, door middel van Croissant naadloos kunnen worden gecombineerd. Het zal academische onderzoekers ook toegang geven tot data uit het bedrijfsleven.

Als je meer wilt weten over Croissant, ga dan naar de website van MLCommons of Google Dataset Search.
https://mlcommons.org/2024/03/croissant_metadata_announce
https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets

KORT

TDCC Challenge projecten

In mei en juni presenteren verschillende groepen hun ideeën voor TDCC-SSH Challenge projecten. De ideeën richten zich op het oplossen van uitdagingen in de sociale en geesteswetenschappen zoals: Kennis en expertise over data management, beschikbaarheid van FAIR data en software, en het (her)gebruik van datasets met gevoelige informatie.
Anders dan bij andere NWO financieringsstromen wordt bij deze projecten vanaf het begin samengewerkt met verschillende partners. Projectideeën worden openbaar gedeeld en op basis van feedback van de gemeenschap verder ontwikkeld. In juni wordt bepaald welke ideeën het meeste potentieel hebben om de uitdagingen van het veld aan te pakken. Deze ideeën worden dan de komende maanden verder uitgewerkt tot projectaanvragen die de consortia in november kunnen indienen bij NWO. (RB)
Een overzicht van de project ideeën en meer informatie is te vinden op
https://tdcc.nl/nwo-tdcc-call-2023/tdcc-ssh-project-development-process/

Onderzoek naar herkenning van overheidswebsites

Helpt een uniforme domeinnaamextensie?

Voor burgers is het niet altijd duidelijk of online overheidsinformatie daadwerkelijk afkomstig is van de overheid. Men kan daardoor denken dat een website van de overheid is terwijl dit niet het geval is (of andersom). ***Mara Verheijen***

Andere aanbieders, maar ook fraudeurs, maken daar gebruik van. Bijvoorbeeld door websites te maken die sterk lijken op een overheidswebsite, maar dat dus niet zijn. Het onderzoek laat zien dat het toevoegen van een uniforme domeinnaamextensie, zoals ‘.overheid.nl’ of ‘.gov.nl’, aan een overheidswebsite burgers kan helpen om overheidswebsites beter te herkennen.

Online experiment

Centerdata voerde dit onderzoek uit in opdracht van Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. Er werd op een objectieve manier onderzocht of een domeinnaamextensie zorgt voor betere herkenning van overheidswebsites en welke domeinnaamextensie het best werkt. Hiervoor werd in het LISS panel van Centerdata een experiment afgenomen, waaraan meer dan 1.700 respondenten deelnamen.
Tijdens het experiment werden de respondenten gevraagd om websites te beoordelen.

Gebruik van internet. *Credits: Pixabay*

Eerst konden ze een aantal websites zónder uniforme domeinnaamextensie beoordelen. Vervolgens ontvingen de respondenten uitleg over domeinnaamextensies en beoordeelden opnieuw een aantal websites, dit keer met de extensie ‘.overheid.nl’ of ‘.gov.nl’.

Belangrijkste bevindingen

- Tussen de 67% en 79% van de respondenten vond het makkelijker om te bepalen of een website, e-mailbericht, social media bericht of app van de overheid was als er een uniforme domeinnaamextensie gebruikt werd.
- Tussen de 73% en 82% vond het handig als er een domeinnaamextensie werd gebruikt.
- Herkenning van overheidswebsites was hoger nádat men uitleg kreeg dan daarvoor (zoals in de huidige situatie, zonder domeinnaamextensie).
- Men is ook beter in het herkennen van websites die niet van de overheid zijn na de uitleg.
- Voor de herkenning van frauduleuze websites die op die van de overheid lijken maar dat niet zijn, helpt de domeinnaamextensie ‘.gov.nl’ mensen beter te herkennen dat dit géén overheidswebsites zijn.
- Subjectief gezien geven burgers een voorkeur aan de extensie ‘.overheid.nl’.
- Respondenten die een uitgebreide vs. korte uitleg kregen over de domeinnaamextensies, herkennen overheidswebsites beter als zijnde van de overheid én websites die niet van de overheid zijn als zijnde niet van de overheid.
- Kwetsbare groepen, zoals mensen met weinig digitale vaardigheden, lager opgeleiden, mensen met een migratieachtergrond en laaggeletterden, zijn eveneens geholpen met het gebruik van een uniforme domeinnaamextensie. Zij blijven minder goed in het herkennen van overheidswebsites, ook na de uitgebreide uitleg.

Conclusie

Op basis van dit onderzoek blijkt dat een uniforme domeinnaamextensie burgers helpt om snel en makkelijk te beoordelen of een website daadwerkelijk van de overheid is. Het verdient de aanbeveling om voor websites, e-mailberichten, social media berichten en apps een uniforme domeinnaamextensie te gebruiken, zodat het overheidsportfolio voor de burger beter te herkennen is.

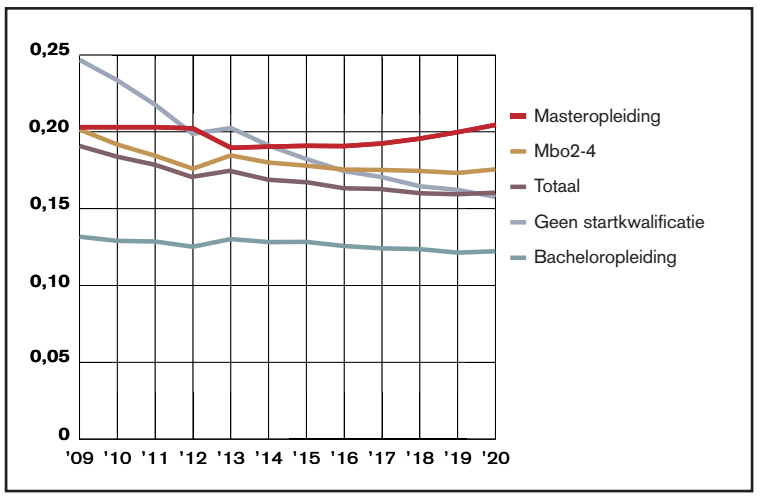
Het volledige rapport is te vinden op de website van de Rijksoverheid: ***https://tiny-url.com/yzedjuhz***
Eind 2024 *zal de dataset ook beschikbaar worden gemaakt in het* LISS *data archive:* ***https://www.dataarchive.lissdata.nl/***

Nieuwe toepassing van de CBS-persoonsnetwerken

Opleidings- en herkomst-segregatie in beeld gebracht

Het CBS heeft op basis van administratieve overheids-registers persoonsnetwerken geconstrueerd van familie, huisgenoten, klasgenoten, buren en collega's. Bronnen zijn onder andere de Basis-registratie Personen, de Polisadministratie en onder-wijsinformatie van DUO.

Marjolijn Das (CBS/EUR/LDE Centre for BOLD Cities), Fijnanda van Klingeren (CBS), Jan van der Laan (CBS), Edwin de Jonge (CBS)



Figuur 1. Segregatie naar opleidingsniveau, 2009-2020

Vanuit deze bronnen is wel bekend dat mensen collega, familielid enzovoorts van elkaar zijn, maar niet of mensen ook daadwerkelijk contact met elkaar hadden. De netwerken moeten daarom gezien worden als een reservoir van potentiële contacten tussen mensen. De netwerkbestanden zijn beschikbaar in de Remote-Access-omgeving van het CBS en kunnen onder strikte voorwaarden door externe onderzoekers voor allerlei soorten wetenschappelijk onderzoek worden gebruikt. De privacy van individuen wordt hierbij altijd beschermd door het CBS. De data bevatten geen direct identificerende kenmerken en zij kunnen en mogen uitsluitend voor onderzoek worden gebruikt, niet voor opsporing of profilering. Resultaten zijn dus nooit te herleiden naar individuen.

Het afgelopen jaar is door het CBS met deze netwerken onderzoek gedaan naar segregatie: hoe gescheiden leven mensen van verschillende herkomstgroepen van elkaar? En hoe zit dat bij mensen met verschillende opleidingsniveaus? Segregatie is een belangrijk maatschappelijk thema. Sociale groepen zijn er altijd geweest. Maar al te sterke segregatie wordt als ongunstig beschouwd, onder andere omdat het polarisatie kan versterken.

Netwerkanalyse

Het meten van segregatie via netwerkanalyses heeft veel meerwaarde. Gewoonlijk richt segregatie-onderzoek zich op segregatie binnen één specifieke context, meestal de buurt of de school. Met de persoonsnetwerken wordt segregatie bepaald in vijf verschillende dagelijkse contexten tegelijk. Daarnaast worden niet alleen directe contacten, maar ook indirecte contacten in het onderzoek betrokken: personen die een, twee of nog meer stappen verwijderd zijn in het netwerk. Dit doet recht aan het feit dat informatie, normen en waarden en voorbeeldgedrag niet alleen direct worden overgedragen maar juist als het ware door een netwerk heen sijpelen. De mate van segregatie wordt uitgedrukt in een segregatiescore. Deze houdt tevens rekening met de grootte van de verschillende groepen in de omgeving. Meer informatie over de berekening van deze score is hier te vinden.

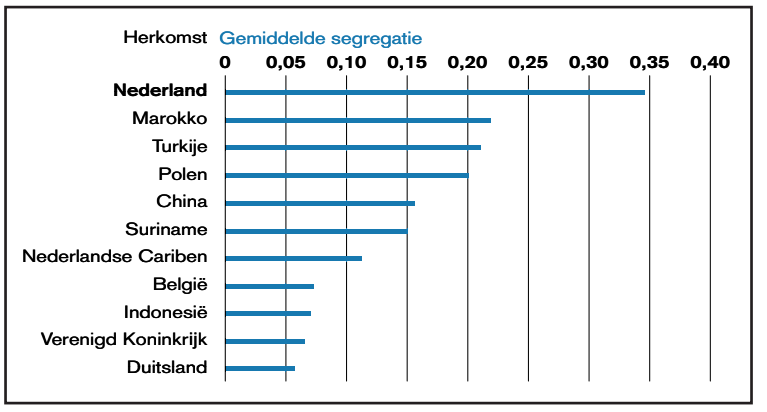
Opleidingssegregatie

De opleidingssegregatie is tussen 2009 en 2020 afgenomen. Dat komt vooral door een steeds lagere segregatie onder laagopgeleiden (mensen zonder startkwalificatie). Dit is over de tijd heen een steeds kleinere en meer selectieve groep geworden. Er zijn tegenwoordig minder banen voor mensen zonder startkwalificatie en het onderwijsbeleid zet sterk in op het behalen van een startkwalificatie. Voor de groepsgrootte wordt gecorrigeerd in de segregatiescore, maar het is goed mogelijk dat mensen zonder startkwalificatie vroeger meer een eigen sociale groep vormden en tegenwoordig meer tussen anderen wonen en leven. In 2020 was de segregatie het hoogst onder mensen met een masteropleiding en het laagst onder mensen met een bacheloropleiding. In de Randstad is er gemiddeld meer segregatie dan in de rest van Nederland. Dat komt met name doordat de segregatie van mensen met een masteropleiding en van degenen zonder startkwalificatie in de Randstad groter is dan buiten de Randstad. Dit past bij het beeld van de stad als gebied van grotere sociale contrasten.

Herkomstsegregatie

Wat betreft herkomstsegregatie blijkt dat alle herkomstgroepen meer mensen van hun eigen herkomst in hun netwerk hebben dan op grond van toeval verwacht mag worden. Elke groep leeft dus in zekere mate gesegregeerd. Dit geldt het sterkst voor mensen met een Nederlandse herkomst, zij leven gesegregeerder dan andere herkomstgroepen. Daarbij geldt ook nog: hoe hoger hun inkomen, hoe meer gesegregeerd het netwerk. Voor veel andere herkomstgroepen, zoals de Turkse, Marokkaanse en Nederlands-Caribische herkomstgroepen, is dat juist andersom: mensen met een hoger inkomen hebben een minder gesegregeerd netwerk dan mensen met een lager inkomen. Een mogelijke verklaring is dat er in 'rijkere' buurten en in werkkringen met veel hoogbetaalde banen verhoudingsgewijs meer mensen met een Nederlandse herkomst wonen en werken.

https://www.cbs.nl/nl-nl/longread/statistische-trends/2024/herkomstsegregatie-in-nederland-een-netwerkanalyse
https://www.cbs.nl/nl-nl/nieuws/2023/15/opleidingssegregatie-in-nederland-gedaald
https://odissei-data.nl/nl/2023/10/het-persoonsnetwerk-van-het-cbs-verbeteringen-uitbreidingen-en-toepassingen/



Figuur 2. Segregatie naar herkomstgroep, 2020

SANE - Secure ANalysis Environment

Vertrouwen opbouwen in veilige gegevensuitwisseling

Gevoelige datasets met vertrouwelijke informatie kunnen niet zomaar worden gedeeld voor wetenschappelijk onderzoek. SANE biedt een gecontroleerde omgeving voor het delen en analyseren van dergelijke gegevens.

Evgeniia Krichever en Lucas van der Meer

Een van de grootste uitdagingen bij het werken met gevoelige gegevens is het beheer ervan en het voorkomen van datalekken. Hierdoor zijn veel gegevensaanbieders, zeker buiten de academische wereld, terughoudend om hun data beschikbaar te stellen voor wetenschappelijk onderzoek uit angst de controle over de data te verliezen. Maar wat als het delen van gevoelige gegevens niet alleen veilig is, maar ook via een eenvoudig en gestandaardiseerd proces zou kunnen? Samen met SURF en CLARIAH heeft ODISSEI een SANE oplossing ontwikkeld voor het aangaan van deze uitdaging, en draagt daarmee bij aan wetenschappelijk onderzoek zonder grenzen.

Wat is SANE?

Secure ANalysis Environment (SANE) is een AVG-conforme onderzoeksomgeving voor het delen en analyseren van gevoelige datasets. Hiermee behouden gegevensaanbieders de volledige controle over hun data. Deze oplossing helpt aanbieders de juiste balans te vinden tussen enerzijds het delen van waardevolle datasets die niet openbaar beschikbaar zijn voor wetenschappers, en de zekerheid dat er tijdens de analyse niks raars gebeurt met hun data.

Deze omgeving heeft twee varianten - Tinker en Blind SANE - en maakt verschillende manieren van werken met gevoelige data mogelijk, afhankelijk van het type onderzoeksproject en de vereisten van de aanbieders.

In Tinker SANE krijgt de onderzoeker toegang tot gevoelige gegevens en kan hij er mee 'spelen'. In Blind SANE kan de onderzoeker de gegevens niet zien en wordt de analyse op de achtergrond uitgevoerd. In beide varianten controleert de dataleverancier of de outputresultaten geen gevoelige informatie bevatten voordat ze geëxporteerd kunnen worden naar de computer van de onderzoeker.

Make it SANE? Wat de European Data Governance Act en Data Act betekenen voor het delen van gevoelige gegevens

Naar verwachting zal de invoering van de European Data Governance Act en Data Act ertoe leiden dat datasets uit alle sectoren beter beschikbaar komen voor de publieke en academische sector. Dit zal leiden tot nieuwe onderzoeksvragen die ten goede komen aan de samenleving. Omdat de precieze uitwerking van deze regelgeving nog onduidelijk is, is er ruimte voor gegevensaanbieders en onderzoekers om de SANE-methode voor het delen van gevoelige gegevens te hanteren bij de implementatie van de twee wetten.

SANE draait op het SURF Research Cloudplatform (gecertificeerd volgens ISO 27001) en volgt het Five Safes-principe. Hiermee ontstaat een volledig afgesloten omgeving, en gegevensaanbieders behouden de controle: ze kunnen voorkomen dat onderzoekers extra data uploaden en kunnen hun activiteiten monitoren. Zo ontstaat een platform voor het uitwisselen van datasets zonder dat deze openbaar beschikbaar zijn.

SANE is een 4-jarig samenwerkingsproject gefinancierd door het Platform Digitale Infrastructuur Social Sciences & Humanities (PDI-SSH). SANE is momenteel toegankelijk voor gebruik door Nederlandse onderzoeks- en onderwijsinstellingen, met ad-hoc toevoeging van organisaties daarbuiten. Bezoek voor meer informatie de SURF website.

www.surf.nl

Datahergebruik kent zes dimensies van afstand

For data reuse, think distance

Van ontvangers van onderzoeksbeurzen en auteurs van wetenschappelijke artikelen wordt tegenwoordig verwacht dat ze hun data delen, zowel vanuit het oogpunt van overheidsbeleid als vanuit het oogpunt van best practice.
Christine L. Borgman en Paul T. Groth

Richtlijnen voor het delen van data, zoals de FAIR-principes (Findable, Accessible, Interoperable, Reusable) instrueren de onderzoekers over hoe ze hun datasets beschikbaar kunnen maken voor anderen. De FAIR-principes en andere beleidsverklaringen over het delen van data impliceren dat onderzoeksdata waardevolle entiteiten zijn die moeten worden beheerd, nuttig zijn voor anderen, en hergebruikt zullen worden. Deze aannames leggen echter een zware last op onderzoekers en data archieven die hen ondersteunen. Onderzoeksprojecten genereren vaak veel grotere hoeveelheden data dan praktisch bewaard kunnen of misschien zouden moeten worden. Het vrijgeven of documenteren van alle data en apparatuur op een manier die ze voor iedereen, voor altijd herbruikbaar maakt, is zelden mogelijk of haalbaar. Naast het streven naar maximalisatie van datahergebruik, is het daarom van essentieel belang om de haalbaarheid van deze aanpak te bespreken.

Wie, waarom, hoe

Het delen van onderzoeksdata is complex, arbeidsintensief, duur en vereist investeringen in infrastructuur van meerdere belanghebbenden. Hoewel het beleid voor open science zich voornamelijk richt op het vrijgeven van data, blijkt dat het hergebruik ervan moeilijk, kostbaar en mogelijk zelfs niet altijd plaatsvindt. Het zou daarom verstandig zijn om bij investeringen in databeheer te overwegen wie de data mogelijk zal hergebruiken, op welke manier, waarom, voor welke doeleinden, en wanneer. Aangezien onderzoekers niet alle potentiële hergebruiksscenario's kunnen voorzien, is het doel van deze studie het identificeren van factoren die belanghebbenden kunnen helpen bij het nemen van beslissingen over investeringen in onderzoeksdata, het identificeren van potentiële hergebruikers en hergebruiksmogelijkheden, en het verbeteren van processen voor data-uitwisseling.

Afstand

Op basis van empirisch onderzoek naar het delen en hergebruiken van data is een theoretisch concept ontwikkeld dat de 'afstand' tussen de maker en hergebruiker van data inzichtelijk maakt. Daarbij zijn zes afstandsdimensies geïdentificeerd die van invloed zijn op het effectief overdragen van kennis: domein, methoden, samenwerking, curatie, doeleinden, en tijd en temporaliteit. Hoewel deze dimensies voornamelijk sociaal van aard zijn, spelen bijbehorende technische aspecten een rol bij het verkleinen – of vergroten – van de afstand tussen makers en hergebruikers. Onze theoretische benadering van de afstand tussen makers en potentiële hergebruikers van data leidt tot aanbevelingen aan vier categorieën van belanghebbenden - makers, hergebruikers, archivarissen en financieringsinstanties - over hoe het delen en hergebruiken van data effectiever kan worden gemaakt.

Borgman, C. L., & Groth, P. T. (2024). From Data Creator to Data Reuser: Distance Matters (arXiv:2402.07926). arXiv. https://doi.org/10.48550/arXiv.2402.07926

KORT

Synthetische data voor onderwijsdoeleinden

In samenspraak met Universiteit Leiden is er als stageopdracht onderzoek gedaan naar synthetische datasets die inzetbaar zijn in het onderwijs. Op basis van de eisen en wensen is er een generatiemethode, om synthetische data mee te genereren, gezocht waarmee de balans tussen de privacy, getrouwheid en geschiktheid doeltreffend blijft. Dit heeft geresulteerd in een synthetische dataset gebaseerd op geaggregeerde data, die Universiteit Leiden gebruikt als ondersteuning van een vak programmeren. (AV)

COLUMN

Wie zorgt er echt voor uw digitale collecties?

Het duurde maar een paar uur. De verwoestende brand van 2018 in het Nationaal Museum van Brazilië vernietigde, naast vele andere kostbare voorwerpen, "de collectie met betrekking tot inheemse talen, inclusief de opnames sinds 1958, de gezangen in alle uitgestorven talen, de Curt Nimuendajú-archieven en de etnologische en archeologische referenties van alle etnische groepen in Brazilië sinds de 16e eeuw."

Snel vooruit naar oktober 2023. Een ransomware-aanval legt een van de grootste bibliotheken ter wereld, de British Library, lam.
In tegenstelling tot de brand in Brazilie is de fysieke collectie volledig ongedeerd gebleven. Maar de manier waarop de gebruikers van de bibliotheek toegang konden krijgen tot deze boeken verdween effectief. Een waarnemer merkte op: "voor zover het de digitale wereld betreft, bestaat het [de collectie] niet".
De digitalisering van culturele collecties, van wetenschap en onderzoek, betekent dat instellingen die verantwoordelijk zijn voor langetermijntoegang en preservering enkele radicaal nieuwe bondgenoten nodig hebben om hun missie te handhaven. Fysieke ruimtes hadden nog steeds behoefte aan door regels geobsedeerde brandveiligheidsbeambten, leveranciers van efficiënte sprinklersystemen en curatoren met een voorliefde voor robuuste vitrines. Maar wie hebben we precies nodig voor onze digitale ruimtes? En hebben we een duidelijk idee van wie dat is?

Een verontrustend feit: de verantwoordelijkheid voor de collectie ligt bij u, maar de verantwoordelijkheid voor de digitale bestanden ligt eigenlijk ergens anders. De bits en bytes staan op een server op een ICT-afdeling een paar kilometer verderop, of een serverpark in Flevoland, of weet ik waar onder het toeziend oog van een groot Amerikaans technologiebedrijf.
Gaat er iets mis? U pakt de telefoon en hoopt dat er iemand vriendelijk opneemt. Misschien hebt u geluk - u hebt een goede band opgebouwd met de IT-medewerker en deze zal het probleem snel oplossen. Of misschien hebt u al een verstandige beleidsaanpak: de zaken worden gladgestreken door de leverancier voordat u het weet.

Maar er blijven nieuwe problemen komen. Nieuwe patches, nieuwe cyberdreigingen, nieuwe personeelsziekten die het IT-team treffen. Bij instellingen waar de culturele collectie onderdeel is van een groter geheel (bijvoorbeeld onderdeel van een universiteit of overheidsinstelling), kan dat nog lastiger zijn. Hoe belangrijk is de culturele collectie in vergelijking met de e-mails van het personeel? Of het HR-systeem? Wat krijgt als eerste prioriteit? Zoals een cybersecurity-expert het samenvat, zijn culturele instellingen "het meest kwetsbaar voor cyberverstoringen".
De penibele positie van culturele instellingen vraagt om een nieuwe manier van denken. In plaats van volledig afhankelijk te zijn van IT-afdelingen die al dan niet onze waarden delen, moeten we onze eigen infrastructuren beheren die bescherming bieden op lange termijn. Sommige van deze infrastructuren bestaan al: het eDepot bij de Koninklijke Bibliotheek. Of het internationale CLOCKSS-consortium van onderzoeksbibliotheken en uitgevers.

Dergelijke infrastructuren moesten misschien worden aangepast aan het tijdperk van cyberbeveiliging. Hoe kunnen dergelijke infrastructuren evolueren om een gefedereerde dienst mogelijk te maken die geen single point of failure heeft? Hoe kunnen we zorgen voor voortdurende onmiddellijke toegang, zelfs als deze wordt bedreigd door kwaadwillende actoren? En hoe kunnen ze werken op regionaal, nationaal en internationaal niveau? Zulke vragen stellen ons in staat om na te denken over en te handelen naar deze urgente dreiging. Dit is de collectieve uitdaging waar we voor staan. En als we die niet aangaan, lopen we het risico dat onze collecties ophouden te bestaan.
Alastair Dunning

Alastair Dunning

Alastair Dunning is Hoofd Onderzoeksdiensten bij de bibliotheek van de Technische Universiteit Delft. Zijn professionele zorgen omvatten het bestrijden van de hydra van non-interoperabiliteit, het verslaan van de draak van dataverlies en het vinden van het elixer van duurzaamheid.
Eerder werkte hij voor Europeana (de aggregator van cultureel erfgoed) en Jisc, de Britse liefdadigheidsinstelling voor ICT-ondersteuning in het universitair en voortgezet onderwijs.
https://allthingslinguistic.com/post/177712815507/folks-theres-nothing-left-from-the-linguistics
https://www.publicbooks.org/how-to-lose-a-library/