Onderzoeksdata delen als norm

De digitale schatkamer van DANS

Heidi Berkhout ■

Wetenschappers besteden veel aandacht aan het verzamelen en verwerken van data. Maar hoe worden datasets opgeslagen? Tijdelijk, vergankelijk? Of duurzaam, blijvend, FAIR? En hoe kunnen onderzoeksgegevens het beste worden hergebruikt? DANS, het Nederlands instituut voor permanente toegang tot digitale onderzoeksgevens, helpt onderzoekers en dataprofessionals bij deze en andere vragen op het gebied van Research Data Management.

In één dataset zit een schat aan informatie, klaar om ontdekt te worden door de onderzoeker. Maar na publicatie van een artikel, of zodra een proefschrift is afgerond, verdwijnen datasets vaak in een la of op een computerschijf. De kans is groot dat na een paar jaar de data zijn vergeten. Of erger, dat de data onvindbaar of niet meer bruikbaar zijn, bijvoorbeeld door verouderde software.

Rijkdom

Data verdienen het om bewaard te worden. Voor de onderzoeker en voor zijn of haar opvolgers. Want door data te delen, kunnen ook anderen kennis uit de data halen door ze te analyseren of te gebruiken voor vervolgonderzoek. Verzamelde data vertegenwoordigen dus een enorme rijkdom. Precies voor dat doel is DANS (Data Archiving and Networked Services) in 2005 opgericht. De Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) en de Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO) namen daartoe het initiatief.
In vijftien jaar is DANS uitgegroeid tot hét Nederlands instituut voor permanente toegang tot digitale onderzoeksgegevens. Meer dan 50 medewerkers stimuleren onderzoekers om hun digitale onderzoeksgegevens vindbaar, toegankelijk, interoperabel en herbruikbaar te maken, ofwel FAIR (Findable, Accessible, Interoperable and Reusable).[1] DANS doet dit door gecertificeerde diensten aan te bieden en training en consultancy te verzorgen voor alle stappen van de 'research life cycle'.

Optimaal hergebruiken van data

De doelen van DANS sluiten aan bij nationale en internationale ambities. Zo stelde de vroegere staatssecretaris OCW Sander Dekker (nu minister voor Rechtsbescherming) in 2013 dat in 2020 alle wetenschappelijke artikelen in Europa vrij toegankelijk moeten zijn. Hij zette hierbij ook in op het optimaal kunnen hergebruiken van onderzoeksdata. Ook de Europese Commissie zet hoog

Alle Nationale Kiezersonderzoeken zijn via DANS beschikbaar. (Bron: Nationale beeldbank/Harm van der Geest, DOI: 10.17026/dans-zcv-uybr)

Internationale data-infrastructuren

Voorbeelden van projecten en infrastructuren op het gebied van researchdata:

- In EOSC-hub wordt gewerkt aan één online omgeving – de zogenoemde European Open Science Cloud ofwel EOSC – waarbinnen onderzoekers data kunnen delen en combineren, ook over disciplinaire grenzen heen (eosc-hub.eu).
- OpenAIRE ondersteunt het Open Science-beleid van de Europese Commissie met ICT-diensten, maar nadrukkelijk ook met een menselijk kennisnetwerk (openaire.eu).
- Het project FAIRsFAIR werkt aan internationale kwaliteitsstandaarden en beleid om wetenschappelijke data FAIR te maken (fairsfair.eu).
- Het FREYA-project werkt aan een infrastructuur voor persistent identifiers, om onderzoekers, hun projecten, publicaties en data met elkaar te verbinden (project-freya.eu).
- Vanuit het ARIADNE-project is een onderzoeksinfrastructuur ontwikkeld met ruim twee miljoen archeologische datasets (ariadne-infrastructure.eu).
- Het CESSDA-consortium van Europese data-archieven bouwt een infrastructuur voor sociale wetenschappers met data, tools, advies en de mogelijkheid om kennis uit te wisselen (cessda.eu).
- CLARIAH is een Nederlands infrastructuurproject in de geesteswetenschappen (clariah.nl).
- EHRI staat voor European Holocaust Research Infrastructure. Het belangrijkste doel is om de in verschillende landen verspreide collecties over de Holocaust via één portal te ontsluiten, en daarmee het wetenschappelijk onderzoek over de Holocaust te ondersteunen (ehri-project.eu).
- De Research Data Alliance (RDA) is een internationale, op leden gebaseerde organisatie die zich richt op de ontwikkeling van een infrastructuur en het verminderen van de sociale en technische belemmeringen voor data-uitwisseling en -hergebruik (rd-alliance.org/rda-europe).
- De Social Sciences & Humanities Open Cloud (SSHOC) heeft als doel om initiatieven van de huidige Europese onderzoeksinfrastructuren beter op elkaar en op de Europese Open Science Cloud te laten aansluiten (sshopencloud.eu).

in op de toegankelijkheid van onderzoeksresultaten. Publicaties en datasets die met Europees onderzoeksgeld zijn gefinancierd, moeten vrij toegankelijk beschikbaar zijn. Om dit mogelijk te maken, financiert de Europese Commissie naast onderzoeksprojecten ook projecten die de benodigde infrastructuur ontwikkelen en op elkaar aansluiten. Projecten om dit te bereiken zijn bijvoorbeeld: EOSC-hub, OpenAIRE, FAIRsFAIR en FREYA (zie kader 'Internationale data-infrastructuren'). Nationaal en internationaal is het delen van data inmiddels prioriteit, waarbij een verantwoorde, toegankelijke dataopslag basisvoorwaarde is, al is het maar om data een langere levensduur en verhoogde waarde te geven.

Huiverig

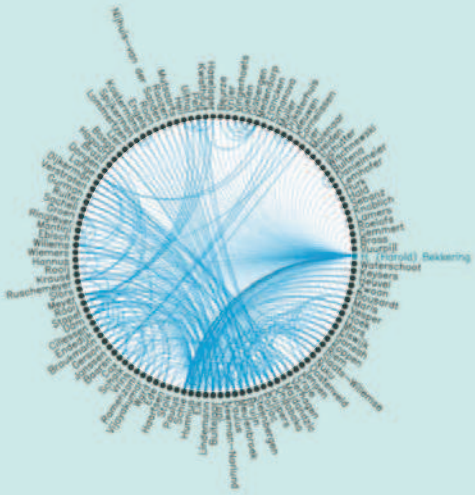
Toch zijn sommige wetenschappers nog huiverig. Waarom eigenlijk? Omdat de kans bestaat dat andere onderzoekers uit de dataset andere conclusies trekken. Misschien strijkt hij of zij dan met de eer van jarenlang onderzoek, of misschien interpreteert een ander de data wel verkeerd en wordt de oorspronkelijke onderzoeksgroep erop aangekeken. Zo'n vergrootglas op eigen werk is natuurlijk spannend. 'It is a paradoxical state of affairs: data are very close to the central workflows of science, they are much less encumbered with copyrights [than publications], and yet they are far less eagerly shared.'[2] Hiertegenover staat de winst die te behalen valt: data duurzaam en veilig opslaan en beschikbaar houden. Dat geeft iets extra's aan het werk, aan het onderzoek, aan de invulling van de rol als wetenschapper en aan diens zichtbaarheid – wat mede van belang is om nieuwe projecten gefinancierd te krijgen. Door data via een archief te delen ofwel te publiceren is het onderzoeksproces ook beter verifieerbaar.

Duurzaam opslaan

Maar hoe kunnen onderzoekers data duurzaam opslaan en publiceren? Zelf deponeren kan in Nederland via EASY, het online data-archief van DANS. Inmiddels telt EASY ruim 116.000 datasets, ruwweg eenzelfde aantal wordt jaarlijks door onderzoekers opgevraagd. In EASY krijgt een nieuwe dataset automatisch een Digital Object Identifier (DOI) toegekend. De onderzoeker of de deponeerder kan zelf zijn of haar Digital Author Identifier (DAI) invullen. DOI en DAI zijn persistent identifiers, die kunnen worden gebruikt om duurzaam te refereren naar de dataset en naar de 'maker' of creator ervan. Binnenkort kan men ook de persoonsidentifiers ORCID en ISNI toevoegen. Met het toevoegen van de identifiers ORCID en ISNI draagt DANS bij aan belangrijke internationale ontwikkelingen op het gebied van persistent identifiers. Door bij het deponeren de Dublin Core metadata zo volledig mogelijk in te vullen, worden en blijven gearchiveerde data goed >>

NARCIS

Onderzoeksportal NARCIS maakt inzichtelijk hoeveel wetenschappelijke uitgaven in Nederland open access-publicaties zijn. In 2018 bijvoorbeeld, was van meer dan 97.000 in dat jaar uitgegeven publicaties, ruim de helft open access beschikbaar. Het betreft zowel tijdschriftartikelen, als proefschriften, conferentiepapers, rapporten et cetera.
Meer dan twee miljoen researchobjecten (publicaties, datasets en onderzoeksprojecten) in NARCIS hebben alle een of meerdere persistent identifiers. Dit maakt het mogelijk om publicaties en projecten van één onderzoeker aan elkaar te linken, wat de vindbaarheid en toegankelijkheid van onderzoek verbetert.



NARCIS maakt het mogelijk om publicaties en projecten van een onderzoeker aan elkaar te linken. (Bron: narcis.nl)

Ook veel archeologische data zijn via DANS te raadplegen. (Bron: ADC Archeoprojecten (DOI: 10.17026/dans-x5n-srs7)

>> vindbaar, citeerbaar en verbonden aan de deponeerder. Rijke metadata helpen geïnteresseerden om te beoordelen of data voor hen relevant zijn, en het archief zorgt ervoor dat de data toegankelijk en bruikbaar blijven. Dit draagt bij aan FAIR-data.

Schat aan datasets

De bij DANS opgeslagen datasets zijn enorm divers. Zo zijn bijvoorbeeld de ervaringen van Nederlandse oorlogsveteranen digitaal vastgelegd in audiobestanden met bijbehorende transcripten. Als men wil weten wat het weer was bij Kaap de Goede Hoop op 6 januari 1763: het staat in een van de gearchiveerde scheepsjournaals uit de periode 1750 tot 1850. Ook alle Nationale Kiezersonderzoeken (NKO) zijn via DANS beschikbaar. Het NKO wordt sinds 1971 gehouden rond de Tweede Kamerverkiezingen en is een rijke bron voor onderzoek naar kiesgedrag door de tijd heen.

Verder zijn veel archeologische data beschikbaar, zoals scans van analoge veldtekeningen, objectfoto's en veldfoto's en tabellen met determinaties van vondsten, die ook gedigitaliseerd zijn. DANS beschikt bijvoorbeeld over de data van een opgraving aan de Grote Marktstraat in Den Haag, waarbij goed bewaard gebleven vondsten uit de zeventiende en achttiende eeuw aan het licht kwamen tijdens de bouw van de Nieuwe Haagse Passage.

Ook zijn de foto's uit Nederlandse geschiedenisschoolboeken online beschikbaar. Uit deze verzameling van ruim 5.000 foto's valt op te maken hoe de Nederlandse geschiedenis verbeeld wordt in lesmateriaal.

224 jaar na de eerste volkstelling in Nederland zijn de data over bevolkingsgrootte, leeftijden, geslacht en beroepsgroepen steeds beter online te raadplegen. Onlangs zijn meer dan 30.000 afbeeldingen met onderzoeksgegevens van de Volks- en beroepentelling 1947 naar Excelbestanden overgezet en via EASY beschikbaar gesteld. Een mooie digitale toevoeging op het oorspronkelijke materiaal.[3]

Twee laatste voorbeelden zijn de dataset 'Mapping the Dutch Vaccination Debate on Twitter: Identifying Communities, Narratives, and Interactions', een onderzoek naar het Nederlandse vaccinatiedebat op Twitter gedurende een aantal maanden in 2017[4], en een dataset met tweets over de Brexit de ochtend na het referendum[5].

Meer weten?

Het blad *E-data & Research* (edata.nl) wordt uitgegeven door CentERdata, CLARIAH, KNAW Humanities Cluster, Koninklijke Bibliotheek, het Rijksmuseum, ODISSEI en DANS. Het blad staat vol ervaringsverhalen over het delen van onderzoeksdata en tools en wordt op verzoek gratis verzonden.

Beide datasets kunnen worden gevonden in EASY en zijn voor de lange termijn opgeslagen bij DANS. Voor hergebruik worden ze beschikbaar gesteld via het platform Mendeley Data.

Dit zijn slechts enkele voorbeelden van de tienduizenden datasets die in EASY te vinden zijn, beschikbaar gesteld door wetenschappelijke onderzoekers voor wetenschappelijke onderzoekers en anderen. Iedereen kan zoeken in EASY[6] en een groot deel van de datasets downloaden. Soms zijn er bijvoorbeeld privacyredenen die ervoor zorgen dat alleen geregistreerde gebruikers toegang krijgen. Bij de vermelding 'Open Access' is de dataset voor iedereen toegankelijk.

Bestandsformaten

Bij het deponeren van onderzoeksdata stuurt DANS aan op het zo veel mogelijk gebruiken van voorkeursbestandsformaten. Zo'n formaat biedt op de langere termijn de beste garanties qua bruikbaarheid, toegankelijkheid en duurzaamheid van data. Het zijn formaten die veel worden gebruikt, open specificaties hebben en onafhankelijk zijn van specifieke software, ontwikkelaars of leveranciers. Data in voorkeursbestandsformaten kunnen zonder meer in EASY worden gedeponeerd. Als het in de toekomst nodig is, zet DANS de data om naar een ander formaat. Zo blijven data bruikbaar. Op de website van DANS staat een overzicht van de verschillende voorkeursformaten. Mocht een onderzoeker een ander bestandsformaat willen deponeren, dan helpen de datamanagers van DANS om te kijken wat hiervoor de beste oplossing is.[7]

Betrouwbare plek voor FAIR-data

Het internationale keurmerk CoreTrustSeal[8] zorgt ervoor dat data-archieven als een 'trustworthy data repository' worden beschouwd. Dit houdt in dat ze voldoen aan de eisen die worden gesteld aan een betrouwbaar data-archief op het gebied van kwaliteit, duurzaamheid en toegankelijkheid. Een onderzoeker die datasets toevertrouwt aan een archief met het CoreTrustSeal, kan erop vertrouwen dat de data in veilige handen zijn. Inmiddels zijn wereldwijd ruim 140 repositories gecertificeerd met een CoreTrustSeal, waaronder EASY van DANS.

Open als het kan, beschermd als het moet

Zowel overheid als wetenschap leggen steeds meer nadruk op het belang van open toegang tot uit publieke middelen gefinancierde gegevens. Bij het deponeren in EASY kan de maker-deponeerder van de data zelf bepalen wie onder welke voorwaarden toegang heeft tot de data. Vanuit de gedachte 'open als het kan, beschermd als het moet' adviseert DANS onderzoekers om, wanneer mogelijk, gebruik te maken van CC0 of CC-BY van Creative Commons.[9] CC0 laat de juridische en technische belemmeringen voor hergebruik van data geheel vervallen doordat de deponeerder afstand doet van alle mogelijk op de dataset berustende rechten. CC-BY vereist naamsvermelding en staat gebruikers toe de gehele inhoud van de dataset vrij verder te verspreiden. Overigens schrijft de Nederlandse Gedragscode voor Wetenschappelijke Integriteit[10] bronvermelding voor bij hergebruik

Oude schoolboeken vertellen hoe de geschiedenis verbeeld werd in lesmateriaal. (Bron: IISG/Cornelis Leenheer (DOI: 10.17026/dans-zfn-u8k4)

van onderzoeksmateriaal. Ook wie CC0-datasets gebruikt, wordt dus geacht die correct te citeren. Daarom pleiten Kraaikamp et al. (2019) voor CC0 als licentie voor datasets. Dat dit soms moeite en goed overleg kost, illustreren zij met de dataset 'Text database of the Hebrew Bible ETCBC3', die onder meer een geannoteerde versie van de Hebreeuwse bijbel bevat. De spanning tussen open access en intellectueel eigendom van de annotaties leidde aanvankelijk tot deponering in EASY met beperkte toegang. Inmiddels is de betreffende dataset[11] voor iedereen beschikbaar dankzij de toepassing van CC0 en worden er vrijelijk apps op ontwikkeld.

Online diensten

Sinds 2005 begeleidt DANS onderzoekers, dataprofessionals, andere data-archieven, onderzoeksinstellingen en onderzoeksfinanciers bij vragen op het gebied van datamanagement, certificering en onderwerpen zoals FAIR-data, open access en software sustainability. DANS biedt hiertoe diverse trainings- en adviestrajecten én de volgende diensten:

- **DataverseNL**: al tijdens het onderzoek kunnen onderzoeksdata opgeslagen, gedeeld en gepubliceerd worden via DataverseNL;
- **EASY**: na afloop van het onderzoek kunnen onderzoeksdata duurzaam opgeslagen en gedeeld worden via het online archiveringssysteem EASY. EASY is CoreTrustSeal-gecertificeerd;
- **NARCIS**: onderzoeksinformatie over onderzoeksprojecten, onderzoekers, onderzoeksorganisaties, data en (open access-) publicaties is te vinden in het nationale wetenschapsportal NARCIS.

Open en FAIR is de toekomst

Begin 2019 maakte minister Van Engelshoven (OCW) drie ambities voor de Nederlandse wetenschap voor de komende vier jaar bekend. Open science komt in deze wetenschapsvisie veelvuldig voor. In haar wetenschapsbrief *Nieuwsgierig en betrokken – de waarde van wetenschap* wordt als een van de kansen het nieuwe financieringsprogramma Horizon Europe (2021–2027) genoemd, met open science als de norm voor alle onderdelen. Binnen open science is het 'hergebruiken van onderzoeksdata' een van de speerpunten.
Ook NWO vindt dat onderzoeksresultaten die zijn betaald uit publieke middelen, wereldwijd vrij toegankelijk moeten zijn. Dit geldt zowel voor wetenschappelijke publicaties als voor andere vormen van wetenschappelijke output. Ook onderzoeksgegevens moeten in principe met anderen gedeeld kunnen worden. Op die manier kan waardevolle kennis worden benut door onderzoekers, bedrijven en maatschappelijke instellingen. De diensten van DANS ondersteunen dit.

Research Data Netherlands

Research Data Netherlands (RDNL) biedt met Essentials 4 Data Support een cursus voor wie meer wil weten over Research Data Management, in groepsverband of online. Kijk op researchdata.nl voor meer informatie. RDNL is een samenwerkingsverband van 4TU.ResearchData, SURFsara en DANS, met als missie het bevorderen van duurzame toegankelijkheid en verantwoord hergebruik van wetenschappelijke onderzoeksgegevens.

Gigantische bron

Het verleden is een gigantische bron voor de toekomst. Data-archieven zijn en blijven onmisbare bouwstenen in onderzoeksnetwerken. Het in 2017 verschenen COAR-rapport *Next Generation Repositories*[12] doet aanbevelingen voor deze bouwstenen en hoe ze de verschillende soorten content toegankelijk kunnen houden voor zowel mens als machine. Herbert Van de Sompel, Chief Innovation Officer van DANS, is een van de auteurs. Binnenkort verschijnt een vervolgdocument dat beschrijft hoe het een en ander past onder de FAIR-paraplu. We willen immers dat betrouwbare repositories bijdragen aan het FAIR maken en houden van onderzoeksdata. ■

Noten

1. Wilkinson et al. (2016), 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data* (2016), volume 3, DOI:10.1038/sdata.2016.18.
2. Kraaikamp et al. (2019), 'On Open Access to Research Data: Experiences and reflections from DANS. Grey Literature Conference Proceedings GL20', http://hdl.handle.net/20.500.11755/17980ef4-02fc-4a34-9e27-2f2dd54d0e25.
3. DOI: 10.17026/dans-zs3-cf4m
4. DOI:10.17632/fjvk93bc5m.1
5. DOI: 10.17632/x9wkrghz23.1
6. Zie: easy.dans.knaw.nl
7. Zie: dans.knaw.nl/over-dans/diensten/easy
8. Zie: coretrustseal.org
9. Zie: creativecommons.nl
10. Zie: vsnu.nl
11. DOI: 10.17026/dans-x8h-y2bv
12. Zie: coar-repositories.org

Heidi Berkhout ■ communicatieadviseur DANS.