

Preferred formats

DANS publiceerde in 2008 het Datakeurmerk (zie ook pagina 1 'Datakeurmerk gaat internationaal'). Een van de richtlijnen van het Datakeurmerk dat DANS vorig jaar publiceerde, geeft aan dat een dataproducent zijn gegevens moet aanleveren in een door het archief voorgeschreven formaat. In de afgelopen decennia zijn er echter heel wat bestandsformaten gekomen en weer (bijna) verdwenen. Een bekend voorbeeld is Word Perfect, ooit razend populair. Voor DANS zelf was het publiceren van de richtlijnen aanleiding om nog eens goed te kijken naar haar eigen lijst van 'preferred formats'.

Een werkgroep binnen DANS heeft deze lijst geactualiseerd. De werkgroep heeft dertien bestandssoorten onderkend die voor het beheer en beschikbaar stellen van onderzoeksgegevens in de alfa- en gammawetenschappen van belang zijn. Enkele daarvan zijn opgemaakte tekst, spreadsheets, images, statistische bestanden en databases. Voor elke bestandssoort heeft de werkgroep vastgelegd in welk formaat of welke formaten op dit moment het meeste vertrouwen bestaat als het gaat om de bruikbaarheid op de lange termijn. Aanvullend zijn enkele andere gangbare bestandsformaten genoemd die

op betrouwbare wijze om te zetten zijn naar een duurzaam formaat.

Een bestandsformaat dat de laatste jaren wat betreft duurzaamheid zeer veel vertrouwen heeft gekregen is PDF/A, dat echter als nadeel heeft dat de erin verpakte data maar zeer beperkt herbruikbaar zijn. Ook voor het Open Document Format (ODF) zijn de verwachtingen hoog. ODF kent een goede balans tussen duurzaamheid en herbruikbaarheid. Daarom hebben PDF/A en ODF eveneens een plaats gekregen in de lijst. In de komende tijd zal de lijst van preferred formats worden gepubliceerd.

Daarbij heeft DANS enerzijds op het oog dataproducenten bewust maken van de noodzaak om de bruikbaarheid op lange termijn in het oog te houden en daar ook een extra inspanning voor te verrichten. Anderzijds wil het instituut de producenten niet ontmoedigen om data die ze al beschikbaar hebben bij haar te deponeren. Andere actiepunten zijn eventuele aanwijzingen voor het aannemen van preferred formats en het aanbieden van gereedschappen hiervoor. Wie geïnteresseerd is in de lijst van preferred formats, of daarover graag wil meedenken, kan contact opnemen met henk.koning@dans.knaw.nl. (Henk Koning)

Dataverse maakt delen eenvoudig

Eind vorig jaar gaf projectleidster dr. Merce Crosas van de Harvard University bij DANS in Den Haag een presentatie over het Dataverse Network Project. Dataverse biedt instellingen, onderzoeksgroepen en onderzoekers de mogelijkheid om onderzoeksdata op een eenvoudige manier te archiveren en te delen met anderen. Het gebruik van de software is gratis en de gebruiker kan de onderzoeksdata op de Harvard-server opslaan, een eigen versie van Dataverse installeren of volstaan met een beschrijving van een databestand met een link naar het bestand.

De aanwezigen bij de presentatie waren getuige van een inspirerende bijeenkomst met een levendige discussie over alle ins and outs van het archiveren en delen van onder-



LUCAS PASTEUNING

Dataverse projectleidster Merce Crosas op bezoek bij DANS

zoeksdata. Uit het groeiend aantal (gerenommeerde) instellingen dat gebruik maakt van Dataverse blijkt dat Dataverse snel aan populariteit en bekendheid wint.

Een van de verklaringen van dit succes is dat het team achter Dataverse goed heeft gekeken naar de eisen en wensen van onderzoekers. Onderzoekers willen namelijk controle houden over wie de data gaat gebruiken en voorwaarden kunnen stellen aan het gebruik van de data. Merce Crosas liet zien dat dit in Dataverse eenvoudig in te stellen is en dat downloadstatistieken van de databestanden standaard beschikbaar zijn. Een belangrijk onderdeel van Dataverse is dat een onderzoeker geciteerd kan worden met een databestand. Verder kan elk databestand van een echtheidsstempel worden voorzien.

Een van de instellingen die gebruik maakt van Dataverse is de Universiteit van Tilburg, die het gebruikt voor de opslag en het delen van bestanden binnen het Network of European Economists Online (NEEO, www.nereus4economics.info/neo.html). NEEO is een Europees project waarin zestien universiteiten uit acht landen samenwerken om publicaties en de onderzoeksdata die aan de basis liggen van de publicatie, open access beschikbaar en toegankelijk te maken. Doorslaggevend voor de keuze voor Dataverse was de eenvoud in het gebruik en de mogelijkheid om de databestanden via een internationale standaard voor sociaalwetenschappelijk onderzoek (Data Document Initiative, DDI) te kunnen beschrijven. (Rob Grim)

Workshop over Trusted Digital Archives

Hoe kunnen de kwaliteit en betrouwbaarheid van digitale onderzoeksarchieven gegarandeerd kan worden en wanneer is een instelling een Trusted Digital Archive? Dat waren op 30 januari belangrijke vragen in een workshop die bij DANS in Den Haag werd gehouden. Er is de laatste tijd een aantal richtlijnen en *best practices* ontwikkeld die heel verschillend zijn en variëren van breed geformuleerd tot zeer gedetailleerd. Ze hebben echter met elkaar gemeen dat ze criteria bevatten waaraan een *trusted digital repository* zou moeten voldoen.

De workshop was specifiek gericht op de toepasbaarheid van al deze richtlijnen in de sociale wetenschappen en in de humaniora. Na korte inleidingen over de richtlijnen werden deze kritisch onder de loep genomen door onderzoekers uit de genoemde wetenschapsgebieden en informatiespecialisten werkend bij (data)archieven of andere repositories.

De aandacht ging daarbij vooral uit naar het door DANS ontwikkelde 'Data Seal of Approval' (DSA), of



www.datasealofapproval.org/node/6

Datakeurmerk. De richtlijnen van het Datakeurmerk werden door hun ruime formulering en flexibiliteit als breed toepasbaar beoordeeld. Als belangrijke pluspunten werden bovendien het creëren van bewustwording van het digitale duurzaamheidsprobleem genoemd en het scheppen van een algemeen referentiekader. Daardoor is het Data Seal bruikbaar voor uiteenlopende onderzoeksgroepen en organisaties. Wel is in de praktijk nadere operationalisering nodig. Methoden voor digitale data-archivering zoals TRAC, DRAMBORA of NESTOR kunnen daarbij van pas komen, ook al zijn deze technischer, specifiekere en gedetailleerder van aard. (Heiko Tjalsma)

<http://thedata.org/>

Sinds kort beschikbaar

Het overzicht toont een aantal databestanden die recent voor onderzoekers beschikbaar zijn gekomen bij CBS en DANS. Een volledig overzicht van de CBS-bestanden is te vinden op www.cbs.nl/microdata. De bij DANS beschikbare databestanden komen

van diverse andere onderzoeksinstellingen. Deze kunnen kosteloos worden gedownload vanuit DANS EASY: <http://easy.dans.knaw.nl>. Via DANS kunnen ook de beveiligde microdata van het CBS kosteloos geleverd worden: www.dans.knaw.nl/nl/data/cbs/overzicht/

Centraal Bureau voor de Statistiek	Periode
Werkloosheidsuitkeringen (WW)	1 ^e helft 2008
Bijstandsuitkeringensstatistiek (BUS en BUS-TRANS)	1 ^e helft 2008
Bijstandsfraude en bijstandsdebiteuren statistiek (BFS-BDS)	1 ^e helft 2008
Registratie Arbeidsongeschiktheid (AO)	1 ^e helft 2008
Statistiek reïntegratie gemeenten (SRG)	1 ^e helft 2008
AOW-uitkeringen (AOW)	1 ^e helft 2008
Pensioenaanspraken (PA)	2005
Inkomens panel onderzoek (IPO)	2006
Consumenten prijzen index (CPI)	2007
Enquête Beroepsbevolking (EBB)	2007
Landbouwteiling (LBT)	2007
Milieukosten van bedrijven (MKB)	2006
Productiestatistieken van diverse bedrijfstakken	2000-2006

Beschikbaar via DANS EASY	Periode
<i>Archeologie</i>	
Wonen en begraven nabij Elst (Gld.)	2006
Archeologisch onderzoek te Opperdoes Kluiten-Zuid	2007
Romeinen aan de Ring	2006
Neolithische bewoningsresten te Leidschendam	2006
Prehistorische bewoning op het World Forum gebied – Den Haag	2007
Erven uit de vroege ijzertijd en de Late Middeleeuwen	2007
Heesche landwerken	2007
(Allen Archeologisch Onderzoeksbureau Leiden, Archol b.v.)	
<i>Geschiedenis</i>	
Census Nederlands Toneel (CENETON, dr A.J.E. Harmsen)	1500-1803
<i>Sociale wetenschappen</i>	
Arbeidsaanbodpanel (OSA – Universiteit Tilburg)	2004
High-school pupils about themselves and their contacts (Radboud Universiteit, Hidde Bekhuis)	2007
ICT gebruik – POLS ICT	
Consumenten Conjunctuur Onderzoek – CCO POLS (Beide CBS – beveiligde microbestanden)	2008

COLOFON

e-data@research is het kwartaalblad in Nederland over data en onderzoek in de alfa- en gammawetenschappen. Het verschijnt onder auspiciën van DANS, het Huygensinstituut, het Internationaal Instituut voor Sociale Geschiedenis, het Centraal Bureau voor de Statistiek, de Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica. Toezending kosteloos aan relaties van de stakeholders en op verzoek aan studenten in de alfa- en gammarichtingen. Oplage: 7500. *e-data@research* is online te raadplegen op www.edata.nl

Uitgever: Stichting Uitgeverij *e-data@research*, Postbus 93067, 2509 AB Den Haag

Redactieadres: Postbus 93067, 2509 AB Den Haag; t (070)3494450 f (070)3494451 e edata@dans.knaw.nl

Redactie: Peter Boot, Ivo Gorissen, Martijn de Groot (hoofd/eindredacteur), Inge Angevaare, Jetske van der Schaaf, Luuk Schreven

Aan dit nummer werkten mee: Sinisa Boksic, Leen Breure, E.H. Dooijes, Rob Grim, Thijs Hermesen, Henk Koning, Jan Kooistra, Bart

de Nil, Dirk Roorda, Eric de Ruijter, Ruurd Schoonhoven, Mariken Teeuwen, Heiko Tjalsma.

Redactiesecretariaat: Lucas Pasteuning, Jetske van der Schaaf

Vormgeving en opmaak: Ellen Bouma

Productie: Uitgeverij Aksant, Amsterdam

Druk: PlantijnCasparie, Almere

ISSN: 1872-0374