Kwaliteit van onderzoeksdata

Zelfs in de gewone pers is de kwaliteit van onderzoeksdata recent een hot item geworden. Juist naar dit aspect hebben Leo Waaijers en Maurits van der Graaf vorig jaar voor SURF een onderzoek gedaan. In deze aflevering rapporteren zij hun bevindingen

Leo Waaijers en Maurits van der Graaf

In een voor SURF uitgevoerd onderzoek naar de kwaliteit van onderzoeksdata hebben we nagegaan wat het begrip kwaliteit in operationele zin betekende in de diverse stadia van de levenscyclus van die data.[1,2] Dit onderzoek was een vervolg op een SURFshare-studie naar de organisatie van onderzoeksdata in Nederland. Daarin werd het complexe landschap van datacenters, data-archieven en datarepositories geschetst met het oog op het beleid van universiteiten en onderzoeksinstituten op dit gebied.[3]

Onze uitgebreide enquête onder Nederlandse hoogleraren en UHD's (universitair hoofddocenten) was gebaseerd op een studie van de literatuur, aangevuld met zestien interviews met dataprofessionals, zowel datamanagers als onderzoekers. We beschrijven hier kort de resultaten en plaatsen die in de context van de ontwikkeling van een Collaborative Science Data Infrastructure.

Wat is kwaliteit?

De literatuur over onderzoeksdata is pas goed op gang gekomen met een OECD-studie uit 2007, *Principles and Guidelines for Access to Research Data from Public Funding.* Deze belangrijke maar tamelijk saaie studie benadrukt het grote belang van onderzoeksdata als zelfstandige bron van informatie en neemt open toegang daartoe als uitgangspunt. De studie vormt het startpunt van een groeiende stroom publicaties op dit terrein. Het viel op dat kwaliteit als eis veelvuldig wordt genoemd in de literatuur, maar dat dit vrijwel nergens operationeel wordt gemaakt. Een uitzondering is een RIN-rapport (Research Information Network) uit 2008 dat een begin maakt met het koppelen van het begrip kwaliteit aan de verschillende stadia van de levenscyclus van data: productie, management en hergebruik. Latere literatuur volgt dit onderscheid regelmatig, maar werkt het niet uit. Ook in de interviews die wij afnamen werd deze driedeling als bruikbaar onderschreven en nader ingevuld met tal van relevante aspecten. Daarop hebben ook wij deze driedeling als uitgangspunt voor ons onderzoek genomen (zie kader op rechterpagina).

Op de valreep van ons onderzoek verscheen het rapport *Riding the wave. How Europe can gain from the rising tide of scientific data.*[4] Het werd geschreven in opdracht van de Europese Commissie door de High Level Expert Group on Scientific Data en speelde als visiedocument een belangrijke rol in ons onderzoek. Hierop komen we apart nog terug.

Enquête

Voor de enquête kregen wij medewerking van DANS. Wij konden in totaal 2811 hoogleraren en UHD's benaderen. De respons was 14 procent, wat hoog is voor dit soort enquêtes. Daarnaast kwamen er opvallend veel geschreven reacties terug. Een eerste conclusie was dan ook dat het onderwerp leeft in academische kring. Voorts viel de hoge expertise op van de respondenten: 95 procent was zelf peer reviewer, 57 procent zat in de redactie van een tijdschrift, 71 procent was zelf dataproducent, 60 procent maakte onderzoeksdata beschikbaar voor gebruik door derden en 50 procent was zelf gebruiker van data van anderen.

In onze enquête stelden we negen concrete maatregelen voor ter verbetering van de kwaliteit van onderzoeksdata. De vragen waren gericht op steun of afwijzing van deze maatregelen. Daarbij werd onderscheid gemaakt naar de disciplines Technische en Natuurwetenschappen, Sociale en Geesteswetenschappen en de Levenswetenschappen.

De voorgestelde maatregelen betroffen:

1 peer review van datasets als onderdeel van de peer review van artikelen;
2 opzetten van datapublicaties, dat wil zeggen aparte peer reviewed publicaties over datasets;
3 citeerbaar maken van separate datasets;
4 Open Access bieden tot datasets;
5 commentaar opvragen van hergebruikers, te publiceren bij de dataset;
6 opzetten van trainingen in datamanagement voor onderzoekers;
7 organiseren van periodieke data-audits;
8 een dataparagraaf verplichten in aanvragen voor projecten;
9 een gedragscode ontwikkelen voor omgaan met datasets.

De uitkomsten waren verbluffend helder (zie tabel op rechterpagina). Een drietal maatregelen kreeg brede steun in alle disciplines: het opzetten van datapublicaties, het citeerbaar maken van datasets en het opvragen van commentaar bij hergebruikers. Kortom, iedereen wil erkenning en waardering voor het beschikbaar stellen van datasets. Dit sloot wonderwel aan bij een van de aanbevelingen van het *Riding the wave* rapport: 'Develop and

	Technische en Natuurwetenschappen	Sociale en Geesteswetenschappen	Levenswetenschappen
Opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	+++	+++	++
Citeren van datasets	+++	+++	++
Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	+++	++	+++
Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	++++	++	
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets		++	+++
Peer review van de dataset als onderdeel van peer review van de publicatie	+	+	++
Aanbieden van datamanagementtrainingen		+	++
Instellen van periodieke data-audits	----	--	+
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	---	--	-

De mate van acceptatie van de voorgestelde maatregelen per discipline

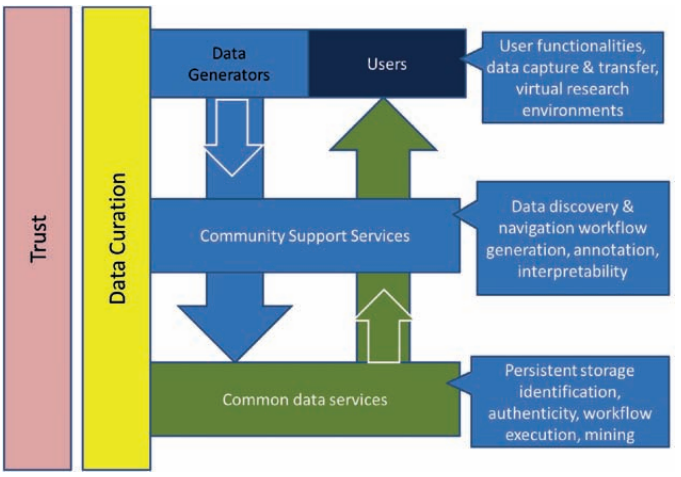
De scores in de tabel zijn bepaald door het percentage respondenten dat van de optie een stimulans vindt uitgaan te verminderen met het percentage van de respondenten dat bezwaren heeft tegen de optie. De percentages zijn afgerond op tien punten en ieder tiental aangeduid met + respectievelijk een -.

use new ways to measure data value, and reward those who contribute it'. Opmerkelijk was voorts dat peer review van datasets als onderdeel van de peer review van artikelen algemeen hoog scoorde, maar dat deze score grotendeels teniet werd gedaan door de verwachting dat dit onhaalbaar zou zijn vanwege de nu al bestaande overbelasting van het peer-systeem. Eveneens algemeen was de afwijzing van een verplichte dataparagraaf in onderzoeksvoorstellen. De commentaren wezen op vrees voor bureaucratie.

De overige vier maatregelen oogstten een uiteenlopende, maar wel duidelijke ontvangst per discipline. De tabel geeft een overzicht.

Onze conclusies op basis van deze resultaten luiden dan ook als volgt:

- De mogelijkheden om datasets te citeren en/of in datapublicaties te beschrijven en/of commentaren over kwaliteit te leveren bij al gepubliceerde datasets, voorzien in een behoefte bij alle disciplines en deze mogelijkheden zouden dus met spoed ontwikkeld moeten worden.
- Tegelijk zijn er tussen de disciplines ook verschillen waar te nemen. In de Technische & Natuurwetenschappen wordt Open Access-beschikbaarstelling van datasets als een goede optie gezien. In de Levenswetenschappen daarentegen ligt dit vanwege de privacygevoeligheid en andere ethische kwesties ingewikkelder. Hier zou eerst een gedragscode voor het omgaan met de beschikbaarstelling van datasets ontwikkeld moeten worden. De Sociale en Geesteswetenschappen zitten tussen deze twee andere disciplines in.
- Trainingen op het gebied van datamanagement zouden volgens een meerderheid van alle respondenten in een behoefte voorzien. Echter, qua populariteit leeft deze optie het sterkst bij respondenten uit de Levenswetenschappen, redelijk bij de Sociale en Geesteswetenschappen en het minst bij de respondenten uit de Technische en Natuurwetenschappen.



Figuur 1.
The collaborative data infrastructure - a framework for the future.
Uit: *Riding the wave*, p. 31

Aspecten van kwaliteit van onderzoekdatasets

- Kwaliteitszorg bij het creëren van de data: het gaat dan om nauwkeurigheid van de data (afstelling van de meetapparatuur en dergelijke) en om de methodische correctheid van de dataverzameling.
- Datamanagement: het betreft onder andere toegankelijkheid (is de dataset goed beschreven door middel van metadata en documentatie?) en duurzaamheid (is het format van de data in de toekomst nog leesbaar?).
- De wetenschappelijke waarde ('the scholarly merit'): het gaat hierbij vooral om de vraag of de dataset betekenis kan hebben voor andere wetenschappers in verband met toekomstig onderzoek.

Collaborative Science Data Infrastructure

Het rapport *Riding the wave. How Europe can gain from the*

rising tide of scientific data werd geschreven in opdracht van de Europese Commissie door de High Level Expert Group on Scientific Data. Dit rapport verscheen in oktober 2010 en ontwikkelt een visie voor het jaar 2030, het jaar waarin betrouwbare onderzoeksdata onbelemmerd toegankelijk en herbruikbaar zijn in een zogenaamde *Collaborative Science Data Infrastructure* (zie figuur op pagina 39 voor een schets). Deze visie wordt inmiddels in de wetenschappelijke wereld breed onderschreven. Ook een studie van Knowledge Exchange, het samenwerkingsverband van SURF, JISC in de UK, DFG in Duitsland en Deff in Denemarken, naar een data-actieprogramma voor de komende periode in de vier landen, neemt *Riding the wave* als uitgangspunt. Onderzoekers zijn zowel de dataproducenten als in veel gevallen ook de dataconsumenten in het steeds data-intensiever wordende wetenschappelijke onderzoek. Zij vormen dus de spil van de in *Riding the wave* beschreven Collaborative Science Data Infrastructure. Kwaliteit is voor wetenschappers essentieel en het verhogen van de kwaliteit van onderzoekdatasets zal ons inziens essentieel blijken bij het realiseren van de visie van *Riding the wave*.

Daarbij zijn ons inziens twee zaken cruciaal:

- *Het delen en beschikbaar stellen van onderzoeksdata dient onderdeel te worden van de wetenschappelijke cultuur*. Om dit te bereiken dient het publiceren van datasets mee te tellen in het academische 'record' van de wetenschappers. Kwaliteit moet dan gehonoreerd worden: bijvoorbeeld door een goede citatiescore of een datapublicatie.
- *Datalogistiek en datamanagement dient een integraal onderdeel van de academische professie te worden.* Opleidingen voor onderzoekers moeten worden aangepast aan data-intensief onderzoek door het opnemen van trainingen op het gebied van datamanagement en datalogistiek. Hier ligt ook een belangrijke rol voor bibliotheken: bij het opzetten en in stand houden van de Collaborative Science Data Infrastructure zijn *data librarians* als een nieuwe vorm van bibliothecarissen van groot belang.[5]

Het lijdt voor ons geen twijfel dat een tijdperk is ingetreden waarin (publicatie van) onderzoeksdata een vergelijkbare rol voor wetenschappelijke vooruitgang zal spelen als publicaties. Dataprofessionals zullen minstens zo'n belangrijke rol spelen als informatieprofessionals. Een idee voor een wijziging van de titel van dit tijdschrift? <

Leo Waaijers is Open Access-consultant. Maurits van der Graaf is senior consultant bij Pleiade Management and Consultancy.

Noten

1] Over kwaliteit van onderzoeksdata; SURFshare studie; 2010:

2] Leo Waaijers, Maurits van der Graaf. Quality of Research Data, an Operational Approach, *D-Lib Magazine*, January/February 2011. tinyurl.com/455c2rx

3] Organisatorische aspecten duurzame opslag en beschikbaarstelling onderzoeksdata; SURFshare-studie; 2010. tinyurl.com/67savjz.

4]] *Riding the wave. How Europe can gain from the rising tide of scientific data.* Final report of the High Level Expert Group on Scientific Data. October 2010. cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

5] Zie bijdrage van Rob Grim in deze serie (*InformatieProfessional* 11/12-2010) en het recente Witboek Dataprofessionals in Nederland van SURFfoundation. tinyurl.com/6geg8rx.

COLOFON

ISSN: 1385-5328
InformatieProfessional is een uitgave (15de jaargang) van Otto Cramwinckel Uitgever, Herengracht 416, 1017 BZ Amsterdam.
www.informatieprofessional.nl

redactieadres
InformatieProfessional, Herengracht 416, 1017 BZ Amsterdam, tel. 020-4276583, fax 020-6383817, e-mail redactie@informatieprofessional.nl.

redactie
Jos van Dijk, Alice Doek, Alice de Jong, Marie-José Klaver (nieuwsredacteur), Carin Klompen, Marieke Kramer, Hans van der Laan, Jenny Mateboer, Paul Nieuwenhuysen, Ronald de Nijs (eindredacteur), Eric Sieverts, Jeroen Tegelaar en Ans ter Woerds.

vormgeving
Eric van den Berg, Tom van Staveren, graphicisland@upcmail.nl.

medewerkers aan dit nummer
Josje Calff, Maurits van der Graaf, Mark Jansen, Bernardette de Lange, Edgar Meij, Jeannette M. Soeters-Keman, Leo Waaijers en Bert Zeeman.

redactieadviesraad
Drs. P. Evers, drs. C. Groeneveld (voorzitter), drs. Ch. L. Citroen, prof.dr. J.S. Mackenzie Owen.

abonnementen
Otto Cramwinckel Uitgever, Herengracht 416, 1017 BZ Amsterdam, tel. 020-6276609, fax 020-6383817. InformatieProfessional verschijnt maandelijks (10 x per jaar, januari/februarinummer en juli/augustusnummer gecombineerd). Abonnementsprijs € 85,-. Instellingen met meer dan één abonnement op hetzelfde adres betalen voor het tweede en volgende abonnement € 57,50. Nieuwe abonnementen: abonnementen worden per jaargang afgesloten. Het abonnement wordt jaarlijks in het eerste kwartaal gefactureerd.
Beëindiging abonnement: Abonnementen kunnen uiterlijk 1 december van het lopende abonnementsjaar worden opgezegd. Bij niet-tijdige opzegging wordt het abonnement automatisch voor een jaar verlengd. Studentenabonnement € 50,- (maximale duur is vier jaar), losse nummers € 11,-. Leden van de NVB komen in aanmerking voor een gereduceerd abonnementstarief.
Meer informatie biedt het NVB-secretariaat, Mariaplaats 3, 3511 LH Utrecht, tel. 030-2330050, email info@nvbonline.nl.

advertentieverkoop
Otto Cramwinckel Uitgever, tel. 020-6276609, fax 020-6383817.

ADVERTENTIE-INDEX