Over kwaliteit van onderzoeksdata

Colofon

Over kwaliteit van onderzoeksdata

SURFfoundation
PO Box 2290
NL-3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60

info@surf.nl
www.surf.nl

Auteurs
Maurits van der Graaf – *Pleiade Management en Consultancy*
Leo Waaijers – *Open Access Consultant*

Eindredactie
Marnix van Berchum – *SURFfoundation*
Brenda van der Laan – *Communicatiewinkel*

Dit rapport is geschreven in het kader van het SURFshare-programma.

SURF is de ICT-samenwerkingsorganisatie van het hoger onderwijs en onderzoek (www.surf.nl).
Deze publicatie is digitaal beschikbaar via de website van SURFfoundation:
www.surffoundation.nl/publicaties

December 2010
ISBN *9789078887096*

Inhoudsopgave

Managementsamenvatting **5**

Management summary **7**

1 Inleiding **9**
1.1 SURFshare 9
1.2 Verkennend onderzoek 9
1.3 Resultaten 9

2 Wat kan kwaliteitsbeoordeling van onderzoeksdatasets inhouden? **11**
2.1 Definitie kwaliteit datasets 11
2.2 Datacreatie onderzoeksdataset en kwaliteit 11
2.2.1 Nauwkeurigheid van de data 12
2.2.2 Methodische correctheid 12
2.3 Datamanagementsaspecten van kwaliteit onderzoeksdataset 12
2.3.1 Metadata en documentatie 12
2.3.2 Formats 13
2.3.3 Rechten 13
2.4 Wetenschappelijke kwaliteit: 'scholarly merit' 13
2.5 Conclusies 14

3 Huidige situatie van kwaliteitsbeoordeling onderzoeksdatasets **15**
3.1 Inleiding 15
3.2 Datasets geproduceerd door onderzoeksfaciliteiten en overheidsinstituten, veelal met het oog op gebruik door derden 16
3.3 Replicatiedatasets en supplementary data 16
3.3.1 Supplementary data 16
3.3.2 Koppeling tijdschriftartikelen en replicatiedatasets 17
3.4 Data-archieven 17
3.4.1 Data-archieven bij tijdschriften 18
3.4.2 'Echte' data-archieven 18
3.5 Niet (rechtstreeks) beschikbaar gestelde datasets 18
3.5.1 Datasets voor eigen gebruik 19
3.5.2 Datasets voor grootschalig(er) gebruik 19
3.6 Datapublicaties 19
3.7 Conclusies 20

4 Wat zijn wenselijke opties voor kwaliteitsbeoordeling onderzoeksdatasets? **23**
4.1 Inleiding 23
4.1.1 Enquête 23
4.2 Peer review datasets als onderdeel peer review publicaties 24
4.3 Datapublicaties: peer-reviewed publicaties over datasets 25
4.4 Commentaren over kwaliteit door hergebruikers van datasets 26
4.5 Citeren van datasets 26
4.6 Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium 27
4.7 Populariteit van negen opties voor kwaliteitsbevordering van datasets 28
4.8 Achtergrondkenmerken van de respondenten 31
4.9 Relatie hergebruikers, dataproducenten en ervaring met het ter beschikking stellen van de eigen datasets 32

5 Conclusies en aanbevelingen **33**
5.1 Conclusies 33
5.1.1 Verschillen tussen disciplines 35
5.2 Aanbevelingen: een per vakgebied verschillende aanpak 35
5.2.1 Voor het gehele wetenschapsgebied 35
5.2.2 Voor de Technische en Natuurwetenschappen en de Sociale en Geesteswetenschappen 36

5.2.3 Voor de Levenswetenschappen en de Sociale en Geesteswetenschappen 36
5.2.4 In andere vorm implementeren 36
5.2.5 Voor de wetenschappelijke tijdschriften 37
5.2.6 Voor individuele onderzoeksinstellingen (Levenswetenschappen en Sociale en Geesteswetenschappen) 37

Appendix A - Bibliografie en bloemlezing 39

Appendix B - Overzicht geïnterviewden 57

Appendix C - Aanvullende gegevens enquête 59

Managementsamenvatting

Hedendaags wetenschappelijk onderzoek resulteert niet alleen in publicaties, maar in toenemende mate ook in onderzoeksdata. Voor hergebruik moeten deze verzamelingen, net als publicaties, vindbaar en toegankelijk zijn. Ook kwaliteit speelt bij onderzoeksdata een belangrijke rol.

In het kader van het SURFshare-programma is een verkennend onderzoek uitgevoerd naar de kwaliteitsbeoordeling van onderzoeksdata. Dit onderzoek bestond uit een literatuurstudie, 16 interviews met sleutelfiguren in dit veld en een enquête onder een representatieve steekproef van universitaire hoogleraren en universitaire hoofddocenten in Nederland.

De voornaamste resultaten van dit onderzoek zijn als volgt:

Wat houdt kwaliteitsbeoordeling van onderzoeksdata in?

- Kwaliteitsbeoordeling van datasets focust vooral op kwaliteitszorg bij de creatie van de dataset, en op de metadata. De kwaliteitstoetsen op deze aspecten zijn bedoeld om de toegankelijkheid van de dataset voor hergebruikers te verhogen.
- Een kwaliteitsbeoordeling van de wetenschappelijke waarde ('scholarly merit') beperkt zich tot het bepalen of de dataset mogelijk zinvol te hergebruiken is door andere wetenschappers.

Wat is de huidige situatie van kwaliteitsbeoordeling van datasets?

- Bij datasets gecreëerd in onderzoeksfaciliteiten en/of overheidsinstituten blijken een of meer mechanismen operationeel te zijn die de kwaliteit van de datasets waarborgen: peer review vooraf, validatie en datacleaning door het datacentrum van de onderzoeksfaciliteiten, feedback van gebruikers en betrokkenheid van peers bij de aansturing van de onderzoeksfaciliteiten. Aanvullende kwaliteitstoetsen lijken in deze gevallen dan ook overbodig en zijn in deze studie buiten beschouwing gebleven.
- Voor de opslag en beschikbaarstelling van datasets gecreëerd door onderzoeksgroepen zijn een reeks van mogelijkheden:
  - De datasets kunnen bewaard kunnen worden in een data-archief. Bij het moment van overdracht vindt in een aantal gevallen een kwaliteitstoets plaats: op de metadata, op de documentatie bij de dataset en een toets of de inhoud van de dataset binnen de scope van het data-archief past.
  - De datasets kunnen als replicatiedataset (of supplementaire data) bij een tijdschriftartikel gepubliceerd worden. Deze data kunnen 'meegenomen' worden in de peer review van de publicatie, maar in de praktijk lijkt dit slechts op beperkte schaal plaats te vinden.
  - Er zijn twee datapublicatietijdschriften bekend waarin artikelen datasets uitvoerig beschrijven. Dit gaat via peer review. Het artikel en de publicatie van de dataset telt op deze wijze mee in het 'academic record' van de auteurs.
  - De datasets kunnen door de onderzoeksgroep zelf worden bewaard en beschikbaar gesteld.

Opties voor kwaliteitsbeoordeling van onderzoeksdatasets:

- Kwaliteitsbeoordeling voorafgaand aan de beschikbaarstelling:
  - *als onderdeel van de peer review van het artikel:* uit de enquête blijkt dat veel wetenschappers dit weliswaar wenselijk vinden, maar niet haalbaar achten omdat de peer reviewers reeds overbelast zijn.
  - *in de peer review van de datapublicatie:* uit de enquête blijkt dat veel wetenschappers het starten van tijdschriften voor dergelijke datapublicaties op hun vakgebied toejuichen.
- Kwaliteitsbeoordeling na de beschikbaarstelling:
  - *commentaren over kwaliteit door hergebruikers van datasets:* hergebruikende onderzoekers voegen commentaren over de kwaliteit van (onderdelen van) de datasets bij de dataset. Deze commentaren worden vervolgens ter beschikking gesteld aan andere potentiële hergebruikers. Uit de enquête blijkt dat velen dit als een wenselijke optie zien.
  - *citeren van datasets:* de hergebruikende wetenschappers kunnen de dataset in hun publicaties citeren wanneer daar duidelijke regels voor zijn. Het aantal citaties kan dan als maat fungeren voor de kwaliteit van de dataset. Ook dit zien veel wetenschappers als een wenselijke optie.

Wenselijkheid van opties voor verbetering van de kwaliteit van datasets: naar een disciplinegerichte aanpak
De enquête heeft negen opties voor verbetering van de kwaliteit van datasets voorgelegd aan de respondenten. Zij konden aangeven van welke drie opties zij de meeste stimulans zien uitgaan voor hun vakgebied en bij welke drie opties zij de meeste bezwaren hebben. De resultaten zijn uitgesplitst naar het vakgebied van de respondenten: Technische en Natuurwetenschappen, Sociale en Geesteswetenschappen, en Levenswetenschappen.

Drie opties zijn populair bij alle disciplines, terwijl weinig respondenten hierin bezwaren zien. Onze aanbeveling is deze opties voor het hele wetenschapsgebied te stimuleren:
- Het opzetten van tijdschriften voor datapublicaties
- Het citeren van datasets
- Commentaren over kwaliteit door hergebruikers

Twee opties stuiten op grote weerstanden. Onze aanbeveling is deze opties niet in deze vorm te implementeren:
- Een verplichte datamanagementparagraaf in onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers
- Het instellen van periodieke data-audits

Voor de overige opties is onze aanbeveling een disciplinegerichte aanpak:
- Het bevorderen van Open Access-beschikbaarstelling van datasets, eventueel na een embargoperiode, is in eerste instantie vooral relevant voor de Technische en Natuurwetenschappen en de Sociale en Geesteswetenschappen. Bij de Levenswetenschappen stuiten regels over Open Access-toegankelijkheid van datasets op grote weerstanden. Deze weerstanden - waarschijnlijk gebaseerd op overwegingen van privacy en ethiek - dienen eerst weggenomen te worden door afspraken en regels.
- Het vaststellen van een gedragscode over datamanagement en het beschikbaar stellen van datasets is in eerste instantie vooral relevant voor de Levenswetenschappen en de Sociale en Geesteswetenschappen. Deze gedragscode dient vooral aandacht te bestede aan het omgaan met privacy en ethiek.

Management summary

Scientific and scholarly research nowadays results not only in publications but increasingly also in research data. Such collections of data need to be traceable and accessible in just the same way as actual publications. Quality also plays a major role in research data.

As part of the SURFshare programme, an exploratory study has been carried out to the quality assurance of research datasets. The study consisted of a literature review, 16 interviews with key figures in the field and an online survey among a representative sample of university professors and associate professors in the Netherlands. The main results of this study are as follows:

What is quality assurance of research datasets?

- Quality assurance of research datasets focuses mainly on the documentation of the creation of the dataset and on the metadata. Quality assessments with regard to these aspects serve the accessibility of the dataset for reusing scholars.
- A quality assessment with regard to the scholarly merit of the dataset is mostly limited to assessing the validity for other scholars.

What is the present situation with regard to quality assessment of research datasets?

- The quality of datasets created by large research facilities and/or governmental institutions appears to be assured by a number of mechanism, that are often operational in these facilities: peer review of the data request, validation and data cleaning by the research facility's data centre, feedback of users and involvement of peers in the government of the research facilities. Therefore, additional quality assessments seem to be superfluous and are omitted in this study.
- Other datasets are created by research groups and/or individual researchers. There are several options to store and make these datasets available:
  - The datasets can be deposited in a data archive. With the ingest of the dataset, the quality of the documentation and metadata may be assessed and there may be a check on the dataset falling within the scope of the data archive.
  - The datasets can be deposited as a replication dataset (or as supplementary data) in conjunction with a journal article. In principle, these datasets should be part of the peer review process of the journal article; in practice, this appears to take place only on a very limited scale.
  - There are two journals dedicated to data publications: the dataset is a such described in an article, peer-reviewed and thus, the article and the publication of the dataset is part of the academic record of the authors.
  - Datasets can be stored and made assessable by the research groups themselves.

Options for quality assessments of research datasets

- Quality assessment before publication of the dataset:
  - *Part of the peer review process of the journal article.* It can be concluded from the results of the survey that many scientists deem this desirable, but unfeasible because of the overload of the peer review system.
  - *Part of the peer review of a data publication journal.* Many scientists participating in the survey would welcome journals dedicated to data publications in the field of work.
- Quality assessment after publication of the dataset:
  - *Commentaries about quality by reusing scholars.* Scholars who reuse a research dataset will be asked to tag their comments on the quality of (parts of) the dataset on it. These commentaries can be used by other scholars who want to reuse the dataset. Many scientists participating in the survey found this a desirable option.
  - *Citing of dataset.* Scholars, who reuse a research dataset, should cite these datasets in their publications when clear rules have been set up on how to cite datasets. The number of citations then can function as a quality indicator. Again, many see it as a desirable option.

Desirability of various options for improving the quality of research datasets: towards a discipline-oriented approach

In the survey, 9 options for improving the quality of datasets were presented to the participants. Respondents were asked to indicate 3 options which would most stimulate the quality of datasets in their field and 3 they would most object. The results are presented per discipline: Physical Sciences and Engineering, Social Sciences and Humanities and Life Sciences. Three options are popular among all 3 disciplines, while there are fewer respondents with objections. Therefore, it is recommended to advance these options for the entire scholarly community:

- journals dedicated to data publications
- citations of datasets
- commentaries on quality by reusing scholars

Two options for improving the quality of research datasets meet many objections and therefore it is recommended not to implement those in this form:

- a mandatory data management paragraph in research grant proposals
- periodical data audits

With regard to other options, a discipline-oriented approach is recommended:

- The advancement of Open Access availability of research datasets, immediately or after an embargo period, is first of all relevant to the areas of Physical Sciences and Engineering and to a lesser extent to the Social Sciences and Humanities. With regard to Life Sciences, any advancement of Open Access availability of datasets will meet large resistance, probably based on concerns about privacy and ethics. This resistance among life scientists should be addressed firstly.
- A Code of Conduct on data management and on how to make research datasets available to others is in first instance relevant to the Life Sciences and to a lesser extent the Social Sciences and Humanities. In this Code of Conduct, especially issues with regard to privacy and ethics should be addressed.

1 Inleiding

Hedendaags wetenschappelijk onderzoek resulteert niet alleen in publicaties, maar in toenemende mate ook in onderzoeksdata: gegevensverzamelingen waarop het onderzoek gebaseerd is, maar die na of naast de publicatie een zelfstandig leven leiden. Voor hergebruik moeten deze verzamelingen, net als publicaties, vindbaar en toegankelijk zijn. Ook kwaliteit speelt bij beide soorten onderzoeksresultaten een belangrijke rol. Voor publicaties is dit aspect in de loop der tijd – niet altijd onomstreden – geoperationaliseerd via peer review en citatie-indexen, maar voor onderzoeksdata staat dit nog in de kinderschoenen.

1.1 SURFshare

SURFfoundation houdt zich in het SURFshare-programma op verschillende manieren bezig met onderzoeksdata. Met als doel de nieuwe faciliteiten voor de opslag en het toegankelijk maken van onderzoeksdata te verkennen. Zo is onder meer het Onderzoeksdata Forum opgericht. Binnen dit forum werken een aantal organisaties, zoals DANS, de NCDD en de 3TU, samen en kijken naar de mogelijkheid om onderzoeksdata gekoppeld aan publicaties te publiceren in verrijkte publicaties.

1.2 Verkennend onderzoek

Om een beeld te krijgen van de kwaliteitsbeoordeling van onderzoeksdata heeft SURFfoundation een verkennend onderzoek laten uitvoeren. Het doel van dit onderzoek is de belangrijkste knelpunten en belanghebbenden te identificeren en te komen tot aanbevelingen voor mogelijke vervolgstappen voor de betrokken stakeholders. Deze aanbevelingen bespreekt SURFfoundation met de stakeholders en worden indien mogelijk in beleid omgezet. De nadruk ligt bij deze studie op de beoordeling van de data zelf en niet op kwaliteitseisen ten aanzien van data-archieven, zoals onder andere uitgewerkt in het Datakeurmerk van DANS.

Het onderzoek maakt gebruik van een literatuurstudie, 16 interviews met sleutelfiguren en een enquête onder wetenschappers in Nederland. De studie is uitgevoerd door Leo Waaijers (Open Access consultant) en Maurits van der Graaf (Pleiade Management en Consultancy) onder supervisie van Gerard van Westrienen, Wilma Mossink en Marnix van Berchum (SURFfoundation) en Laurents Sesink (DANS).

1.3 Resultaten

Dit rapport presenteert de resultaten van het onderzoek als volgt:

- Hoofdstuk 2: de verschillende mogelijkheden voor kwaliteitsbeoordeling van onderzoeksdatasets op basis van de bevindingen van de literatuurstudie en de interviews.
- Hoofdstuk 3: de huidige situatie van kwaliteitsbeoordeling van onderzoeksdatasets, voornamelijk gebaseerd op de interviews.
- Hoofdstuk 4: de verschillende opties voor kwaliteitsbeoordeling van de onderzoeksdatasets (resultaten van de interviews en van de enquête onder wetenschappers).
- Hoofdstuk 5: conclusies en aanbevelingen voor verschillende stakeholders.
- Appendices: de literatuurstudie en ander achtergrondmateriaal.

2 Wat kan kwaliteitsbeoordeling van onderzoeksdatasets inhouden?

2.1 Definitie kwaliteit datasets

Bij aanvang van deze studie werden de volgende aspecten van kwaliteit van onderzoeksdatasets onderscheiden (zie Appendix A, referentie #16):

- Kwaliteitszorg bij het creëren van de data (zie 2.2)
- Datamanagement: duurzaamheid en toegankelijkheid, o.a. metadata, documentatie, formats, rechten (zie 2.3)
- Inhoudelijke kwaliteit: 'the scholarly merit of the dataset' (zie 2.4)

Alle geïnterviewden herkenden deze aspecten van kwaliteit en de literatuur onderscheidt deze of vergelijkbare aspecten ook.

Onderstaande outlines van de werkwijzen van een hergebruikende sociale wetenschapper en een hergebruikende bio-informaticus geven inzicht in het belang van de genoemde aspecten van kwaliteit.

Secondary analysis in social sciences (uit referentie 7)

- The 1st step is to search, choosing and obtain the data. Users match their research interest with descriptions of available datasets and decide which dataset to choose.
- After obtaining the data, users do standard checks on the data. Data checks include:
  - file verification (among others: verify definitions of variables)
  - sample verification (degree to which the data reflect the sampling procedures described, investigate the effects of nonresponse, data loss and missing data affect the data)
  - how are non-responses and missing data handled.
- Data manipulation: constructing new variables based on variables in the existing data, recoding data, and emerging several files etc.
- Data analysis: applying analysis techniques to the newly created dataset.

Hergebruik datasets in de bio-informatica

In geval van een meta-analyse worden meerdere datasets van derden over hetzelfde onderwerp naast elkaar gezet en geanalyseerd. Hoe wordt de kwaliteit van een dataset van een en ander in deze gevallen beoordeeld? De respondent geeft aan dat voor vrijwel alle datasets in de databank van het NCBI geldt dat er een publicatie aan gekoppeld is. De respondent schetst de volgende gang van zaken:

- Een eerste kwaliteitstoets is de reputatie van het betreffende tijdschrift waarin het artikel is gepubliceerd.
- Een tweede toets is dat de in de publicatie beschreven methode wordt bestudeerd en beoordeeld. Wanneer het geanalyseerde data betreft, dan wordt bekeken hoe men de analyse heeft uitgevoerd. Gemakkelijker is er als het ruwe data betreft: de scans van de microarrays. Dan kan de hergebruiker dezelfde toetsen op kwaliteit uitvoeren als deze op de eigen ruwe data zou uitvoeren.
- Wat houden deze laatste kwaliteitstoetsen in? Het gaat om gedetailleerde, technische toetsen, die moeilijk uit te leggen zijn aan een leek. Het principe is als volgt: elke microarray is een weergave van een fluorescent beeld, dat de gen-expressie representeert. Eén van de mogelijk uit te voeren toetsen op kwaliteit van de dataset is om de totale genexpressie te bepalen, waarbij men ervan uitgaat dat die in vrijwel alle gevallen ongeveer gelijk zal zijn. Mocht de overall genexpressie van een bepaalde meting veel hoger of lager zijn, dan duidt dit op een fout in de meting of in het sample.

2.2 Datacreatie onderzoeksdataset en kwaliteit

Een elementair maar zinvol onderscheid kan worden gemaakt tussen:

1. data die worden voortgebracht door apparatuur
2. data die het resultaat zijn van registratie van menselijk gedrag en bevindingen.

Dit onderscheid is in de context van dit onderzoek zinvol omdat het kwaliteitsvraagstuk voor beide categorieën data verschillende invalshoeken kent.

1. Bij de eerste categorie staat de nauwkeurigheid van de apparatuur en het raffinement van de toegepaste algoritmiek centraal.
2. Bij de tweede categorie spelen vooral methodologische vraagstukken een primaire rol.
3. Dit onderscheid wordt nog eens geïllustreerd door de kwaliteitstoetsen bij hergebruik, zoals op de vorige pagina weergegeven in de outlines.

Het genoemde onderscheid valt niet zonder meer samen met de scheiding tussen exacte en geesteswetenschappen. Zo behoren gedigitaliseerde tekstcorpora uit de geesteswetenschappen tot de eerste categorie data en werken de sociale en economische wetenschappen met dataverzamelingen uit beide categorieën.

2.2.1 Nauwkeurigheid van de data

Het gaat dan onder andere om de nauwkeurigheid van meetgegevens. Daarbij heeft men bijvoorbeeld de calibratiegegevens van de meetapparatuur nodig. Het kan namelijk voorkomen dat de calibratie van bepaalde meetapparatuur wordt veranderd of dat er een beter algoritme is ontwikkeld om een en ander te berekenen. Dit kan dus betekenen dat de dataset opnieuw uitgerekend moet kunnen worden met de verbeterde calibratiegegevens en/of het verbeterde algoritme.

Een respondent stelt dat het vaak onmogelijk is om achteraf vast te stellen of de data op zichzelf nauwkeurig zijn vastgesteld. Deze respondent gaf als voorbeeld dat het niet te achterhalen is of bij een enquête de vraag man/vrouw juist is ingevuld. Wel kan men achterhalen dat er iets niet klopt aan de codering van deze vraag. Met andere woorden: een controle op de nauwkeurigheid van de datameting wordt vaak een controle op de nauwkeurigheid van de datadocumentatie.

2.2.2 Methodische correctheid

Het gaat dan om vragen als:

- Is de gekozen methode voor dataverzameling de meest passende voor dit onderzoeksdoel?
- Is die methode correct toegepast (bijv. random samples, double blindness)?
- Beschrijft de dataset adequaat het te onderzoeken fenomeen (o.a. representativiteit)?
- Hoe zijn integriteitaspecten voor de data afgehandeld bij het opschonen van de dataset (bv tegenstrijdige data, veronderstelde meetfouten, onvolledige data, etc.)?
- Is aan de toepasselijke ethische voorwaarden voldaan (bijv. op het gebied van privacy, regels over dierproeven e.d.)?

De bij een dataset behorende documentatie geeft een verantwoording van deze zaken en is essentieel voor een oordeel over de kwaliteit van de dataset.

2.3 Datamanagementsaspecten van kwaliteit onderzoeksdataset

Bij dit aspect staat het begrip kwaliteit vooral voor (duurzame) toegankelijkheid. Daarbij worden de volgende elementen genoemd:

2.3.1 Metadata en documentatie

Enkele respondenten noemen het maken en aanleveren van een goede beschrijving van de dataset een drempel voor onderzoekers. Toch wordt ook gezegd dat dit niet meer dan enkele uren werk zou hoeven te zijn. Het grootste struikelblok voor wetenschappers is de documentatie van de dataset. Het gaat dan om zaken als codeboeken, beschrijvingen van de respondenten of vragenlijsten en dergelijke. Dit is vooral problematisch voor datasets die in principe opgezet zijn

voor eigen gebruik. Vaak ontbreekt dan de documentatie en is het een groot knelpunt om deze achteraf te maken. Ook speelt hier het 'tacit knowledge'-probleem: onderzoekers die de dataset hebben gecreëerd, vinden een aantal zaken daarvan zo logisch dat zij die niet opschrijven.

2.3.2 Formats

Data-archieven stellen dat het dataformat van de dataset zodanig dient te zijn dat deze voor andere onderzoekers hanteerbaar is. In de interviews zijn overigens geen gevallen genoemd van onleesbare dataformats. Wel is aandacht gevraagd voor de zogenaamde 'grid-gegevens' in geografische datasets. Het gaat dan om de locatiegegevens van het meetpunt, dat op verschillende manieren beschreven kan worden. Binnen het NODC-netwerk zijn gegevenssets zodanig bewerkt dat deze per meetpunt bevraagd kunnen worden (dus zonder alle afzonderlijke datasets apart te doorzoeken).

2.3.3 Rechten

Vooral bij geesteswetenschappen speelt bij sommige datasets het probleem van auteursrechten. Voorbeeld: een bepaald tekstcorpus is gemengd met een andere tekst, waarop de onderzoeker geen auteursrechten heeft. Bij sociale en medische wetenschappen spelen privacykwesties vaak een rol. Ook kunnen contractuele afspraken over geheimhouding (bijvoorbeeld in geval van bedrijfsgegevens of voor derdegeldstroomonderzoek) een rol spelen.

2.4 Wetenschappelijke kwaliteit: 'scholarly merit'

De meeste geïnterviewden zien een beoordeling van de 'scholarly merit' van een onderzoeksdataset als ondoenlijk/onpraktisch en/of als theoretisch onmogelijk en/of als niet op zijn plaats.

Ondoenlijk/onpraktisch:
- Enkele respondenten geven aan dat dit veel werk zou inhouden. Volgens een respondent zou hij in dezelfde tijd zelf een artikel kunnen schrijven.
- Een andere respondent geeft aan dat een peer review tot een ja/nee-beoordeling leidt (het onderzoeksvoorstel wordt wel/niet toegekend, het artikel wordt wel/niet gepubliceerd). Dit is minder relevant bij een dataset: deze is er immers al.

Theoretisch onmogelijk:
- Sommige respondenten vinden het theoretisch onmogelijk om een dataset op scholarly merit te beoordelen. Het gaat immers om de onderzoeksvraag in relatie met de dataset. Hergebruik van datasets is juist gericht op andere, toekomstige onderzoeksvragen.[1]

Niet op zijn plaats:
- Een aantal respondenten geeft aan dat het onderzoeksvoorstel, dat ten grondslag ligt aan de dataset, in veel gevallen al door een peer review proces is beoordeeld. Ook worden de publicaties al door een peer review beoordeeld. Een (extra) beoordeling van de dataset zou dan niet op zijn plaats zijn.

Respondenten gaven aan dat wanneer er toch een beoordeling op de scholarly merit van een dataset plaatsvond, dat uitsluitend gebeurde op de betekenis van de dataset, zoals onder andere

[1] Dit is overigens een grote vrees van een respondent, die er met nadruk op wijst dat het gevaarlijk is om datasets voor onderzoeksvragen te gebruiken waarvoor zij niet geschikt zijn. Deze respondent heeft dan ook sterke twijfels bij het Open Access beschikbaar stellen van onderzoeksdatasets en vind dat om de genoemde redenen de hergebruiker vooraf afspraken dient te maken met de datasetproducent.

blijkt uit de uniciteit van de dataset en/of de bruikbaarheid van de dataset voor toekomstig onderzoek.[2]

2.5 Conclusies

1. De kwaliteitsbeoordeling van onderzoeksdatasets focust vooral op de in de documentatie beschreven kwaliteitszorg bij de creatie van de dataset, en op de metadata. De kwaliteitstoetsen op deze aspecten zijn bedoeld om de toegankelijkheid van datasets voor hergebruikers te verhogen.

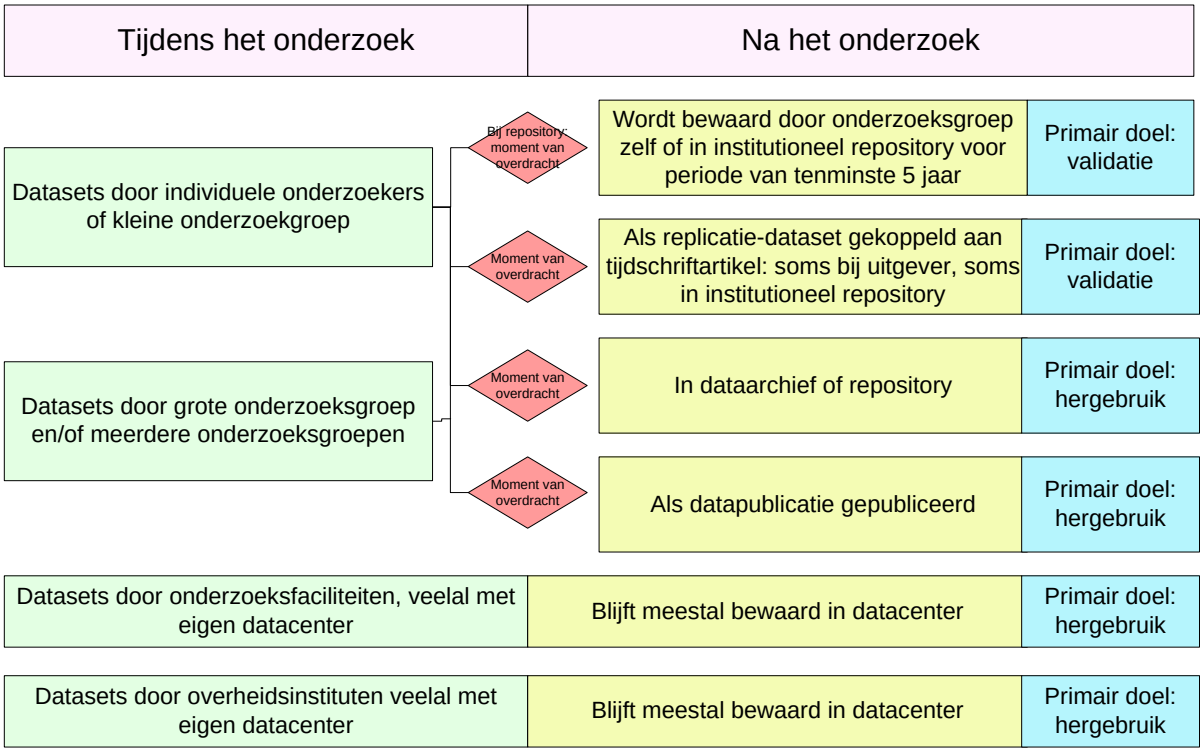
2. Een kwaliteitsbeoordeling van de wetenschappelijke waarde (scholarly merit) beperkt zich tot het bepalen of de dataset mogelijk zinvol te hergebruiken is door andere wetenschappers. Een directe kwaliteitsbeoordeling op de werkelijke wetenschappelijke waarde vindt niet plaats en wordt door de gesprekspartners ook afgewezen: ofwel op theoretische gronden ofwel op praktische gronden.

[2] Zie Earth System Science Data, textbox op pag. 20 en de ICPRS quality appraisal: *'substantive current value for research and instruction, or enduring value or unique in some way'*, http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/appraisal.jsp

3 Huidige situatie van kwaliteitsbeoordeling onderzoeksdatasets

3.1 Inleiding

Het rapport 'Organisatorische aspecten duurzame opslag en beschikbaarstelling onderzoeksdata'[3] van SURFfoundation presenteert het landschap van organisatievormen als volgt:



Het rapport maakt onderscheid tussen twee soorten onderzoeksdatasets:

- datasets gecreëerd in onderzoeksfaciliteiten en/of overheidsinstituten waar in principe geen overdracht van de dataset plaatsvindt
- door onderzoeksgroepen gemaakte datasets die kunnen worden bewaard in een data-archief, gepubliceerd als datapublicatie, gepubliceerd als replicatiedataset (of supplementaire data) bij een tijdschriftartikel of worden bewaard door de onderzoeksgroep zelf. Behalve in het laatste geval, vindt hier wel overdracht van de dataset plaats.

De situatie wat betreft de kwaliteitsbeoordeling van datasets blijkt sterk samen te hangen met de genoemde organisatievormen.

[3] Beschikbaar via http://www.surffoundation.nl/nl/publicaties/Pages/Organisatorischeaspectenduurzameopslagenbeschikbaarstellingonderzoeksdata.aspx

3.2 Datasets geproduceerd door onderzoeksfaciliteiten en overheidsinstituten, veelal met het oog op gebruik door derden

Voor deze studie zijn vertegenwoordigers geïnterviewd van de LOFAR (radiotelescoop), het KNMI (meteorologische meetgegevens), de NODC (netwerk van datacentra van oceanografische instituten), CenterData (producent van het LISS-panel, longitudinale gegevens van een panel van 5000 Nederlandse huishoudens), en de HSN (Historische Steekproef Nederlandse Bevolking).[4]

De wijze waarop deze onderzoeksfaciliteiten van zeer verschillende disciplines met kwaliteit van de datasets omgaan, blijkt in grote lijnen overeen te komen:

- **Peer review vooraf:** Bij LOFAR en het LISS-panel dienen onderzoekers verzoeken in voor waarnemingstijd respectievelijk een vragenlijst. Deze worden vooraf beoordeeld door een peer review op 'scholarly merit'.

- **Validatie en data-cleaning door datacentrum:** Een gecreëerde dataset wordt gevalideerd voordat deze aan de gebruikers beschikbaar wordt gesteld. Zo gaat het bijvoorbeeld bij astronomische gegevens voor een check of alle apparaten wel werkten, en of er geen storingen hebben plaatsgevonden (bijvoorbeeld: mobiel telefoongebruik in de buurt van radio antennes)[5]. Een ander voorbeeld wordt genoemd bij longitudinale enquêtes: hier kan bijvoorbeeld een respondent een foute invoer hebben gedaan, zodat deze in 1 jaar 10 jaar ouder is geworden. Over deze fouten moet een besluit genomen worden (of een waarde negeren of het antwoord weglaten): in de praktijk blijken de wetenschappers dit besluit vaak zelf te willen nemen.

- **Feedback van gebruikers:** Eventuele opmerkingen over de kwaliteit van de datasets zullen door gebruikers snel gemeld worden. Bij LOFAR kregen de onderzoekers bij een onvoldoende kwaliteit van de dataset opnieuw waarnemingstijd toegewezen. Vertegenwoordigers van de onderzoeksfaciliteiten gaven aan opmerkingen over kwaliteit van de gebruikers te verwelkomen om de kwaliteit van hun output te verbeteren. De HSN werkt met releases: in een nieuwe release zijn meestal data toegevoegd, maar ook data verbeterd op grond van nieuwe gegevens.

- **Betrokkenheid van peers bij onderzoeksfaciliteit**: Bij de aansturing van de grote onderzoeksfaciliteiten zijn wetenschappers veelal betrokken als leden van adviescommissies, bestuur en dergelijke. Ook is er bij een aantal onderzoeksfaciliteiten sprake van visitatiecommissies, waarin de gang van zaken door peers wordt doorgelicht.

3.3 Replicatiedatasets en supplementary data

3.3.1 Supplementary data

Veel tijdschriften bieden de mogelijkheid aan auteurs om 'supplementary information' bij het artikel te publiceren. Samenhangend hiermee is er een ontwikkeling naar verrijkte publicaties, waarbij aanvullende gegevens bij de tekst van de publicatie ter beschikking worden gesteld. Het gaat bijvoorbeeld om grafieken en visualisaties, spreadsheets en/of videofilms, maar ook om gegevens van chemische formules, kristallografische datasets en dergelijke. Uitgevers bieden deze supplementary data aan op hun digitale platform, in veel gevallen met beperkingen (bijvoorbeeld niet meer dan 100 MB en zonder garanties van digitale duurzaamheid).

[4] Voor een overzicht van de geïnterviewden, zie Appendix B.

[5] Dit opschonen/ valideren van datasets lijkt universeel. De term 'ruwe data' is dan ook niet helemaal juist: het gaat in de meeste gevallen om opgeschoonde, gevalideerde data, die bewaard worden. Deze kunnen mogelijk beter met de term 'basis onderzoeksdata' aangeduid kunnen worden. De oorspronkelijke ruwe data worden meestal niet bewaard.

Deze supplementary data moeten in principe worden meegenomen in de peer review van de publicatie zelf. Sommige respondenten geven aan dat in de praktijk peer reviewers dit niet of nauwelijks doen. Andere respondenten geven aan dat het in veel gevallen om zeer beperkte datasets gaat, die niet vergelijkbaar zijn met de datasets die in data-archieven en -centra zijn opgeslagen[6].

3.3.2 Koppeling tijdschriftartikelen en replicatiedatasets

Editorial boards van wetenschappelijke tijdschriften eisen in toenemende mate dat onderzoekers de datasets, die ten grondslag liggen aan de publicatie (zogenaamde replicatiedatasets), deponeren. Volgens het ICPRS gaat het om datasets met alle data en informatie, die voor een andere onderzoeker noodzakelijk is om het onderzoek te repliceren. De zogenaamde Brusselse verklaring van de uitgeversorganisaties ALPSP en STM[7] stelt:

1. ('ruwe') datasets vallen niet onder het copyright
2. vrije beschikbaarheid van datasets is wenselijk
3. opslag van datasets is niet een taak voor uitgevers
4. koppeling met publicaties is wenselijk.

De afbakening van de aan de publicatie gekoppelde datasets en de uit het (gehele) onderzoek resulterende datasets is overigens niet altijd goed te maken. De koppeling van publicaties en datasets lijkt dan ook voor beide vormen van datasets gebruik te kunnen worden.

Uit de interviews blijkt dat dit volop in ontwikkeling is:

- **Opslag replicatiedatasets in data-archieven**
  Het ICPSR heeft hiervoor het publication related archive ingericht. Onderzoekers kunnen daarin de bij publicaties behorende datasets deponeren. Ook het UKDA biedt middels de UKDA-store[8] een vergelijkbare mogelijkheid voor onderzoekers aan. Deze store geeft onderzoekers de mogelijkheid hun datasets zelf te deponeren om aan de verplichtingen van het tijdschrift te voldoen. Zie voor meer informatie 3.3.

- **Koppeling replicatiedatasets en publicaties**
  In ScienceDirect (het platform van Elsevier Science, de grootste uitgever van wetenschappelijke tijdschriften) is recentelijk een mechanisme ontwikkeld met een data-archief. Hierin wordt de koppeling tussen de publicatie in het tijdschrift en de bijbehorende dataset in een data-archief automatisch gelegd. Men hoopt deze koppeling met meerdere data-archieven te kunnen realiseren en uiteindelijk via DataCite met alle data-archieven (DataCite wordt dan gezien als de CrossRef voor datasets).

Wat betreft de kwaliteitsbeoordeling heeft dit in beide gevallen niet of nauwelijks consequenties: peer review is gefocust op de publicatie en de koppeling naar de dataset geeft de peer reviewers de mogelijkheid om naar de achterliggende data te kijken.

3.4 Data-archieven

Uit de interviews met de respondenten van de zijde van de data-archieven komt naar voren dat bij een opname in het data-archief enkele kwaliteitstoetsen plaatsvinden.

[6] Ook in de Brusselse verklaring van ALPSP en STM wordt gesteld dat supplementary data meestal 'smaller subsets' betreft van grotere datasets.
[7] Databases, data sets and data accessibility – views and practices of scholarly publishers; STM, ALPSP, June 2006.
[8] http://store.data-archive.ac.uk/store/

3.4.1 Data-archieven bij tijdschriften

Bij het ICPSR en het UKDA kunnen onderzoekers (replicatie)datasets zelf deponeren in een voorportaal van het data-archief. Er vindt geen kwaliteitscontrole plaats over deze soorten datasets in data-archieven. Het ICPSR heeft hiervoor het publication-related archive ingericht en het UKDA de UKDA store. Direct na het deponeren is de dataset beschikbaar en kan de betreffende onderzoeker een link toevoegen aan de publicatie. Deze mogelijkheden van ICPSR en het UKDA zijn vooral ingericht om onderzoekers te ondersteunen bij hun verplichtingen met betrekking tot de beschikbaarstelling van onderliggende datasets bij tijdschriftartikelen en om de workload voor het data-archief te beperken.

De respondenten van beide data-archieven geven aan dat op basis van diverse kwaliteitscriteria prioriteiten worden gesteld wat betreft de in het 'echte' data-archief op te nemen datasets. Dit gebeurt mede met het oog op de beperkte middelen. Immers, opname in het data-archief kost menskracht op het gebied van o.a. ontsluiting en datacuratie.

3.4.2 'Echte' data-archieven

Voor opname in het 'echte' data-archief bestaat de kwaliteitstoets in veel gevallen uit:
- een toets of de dataset inhoudelijk in het data-archief past
- een toets op de (volledigheid van de) metadata
- een toets op de (volledigheid van de) documentatie (zie tekstbox hieronder).

Quality criteria ICPSR
- ICPSR strongly prefers data collections that have comprehensive technical documentation providing ample information on sampling procedures, weighting, recoding rules, skip patterns, constructed variables, and data collection procedures.
- ICPSR prefers data in the most complete and original form, with the exception of data extracts specifically intended for instructional purposes.
- Lower quality data will be considered for inclusion if the data have unique historical value

Sommige data-archieven vragen voorafgaande aan de opname om een sample voor deze kwaliteitstoetsen. Anderen (zie hierboven) gebruiken een soort voorportaal. Hierbij kan de hele dataset gedeponeerd worden en blijven ook een aantal datasets achter.

Wat betreft de kwaliteit van de huidige aangeboden datasets zijn de ervaringen van de respondenten vergelijkbaar:
1. Onderzoekers hebben moeite met het opstellen van de metadata bij hun dataset. Volgens enkele respondenten kost dat overigens slechts enkele uren werk.
2. Wat betreft de documentatie van de dataset is er een duidelijk knelpunt: veel onderzoekers hebben geen training gehad in datamanagement. Dit is ook zelden een onderdeel van het curriculum voor onderzoekers in opleiding. Het documenteren van data achteraf is zeer tijdrovend en in sommige gevallen vrijwel niet mogelijk. Een aantal data-archieven biedt dan ook trainingen aan op het gebied van datamanagement. Het UKDA - geholpen door een mandaat van de betreffende onderzoeksfinancier - benadert onderzoekers direct na de toekenning van het onderzoeksvoorstel om deze te begeleiden in het opzetten van de dataset. Dit geeft volgens de betreffende respondent goede resultaten.

3.5 Niet (rechtstreeks) beschikbaar gestelde datasets

Een respondent introduceert een indeling gebaseerd op het gebruik van de dataset:
1. primair opgezet voor eigen, meestal eenmalig gebruik door individuele of kleine groepen wetenschappers
2. primair opgezet voor (langdurig) gebruik door een grotere groep wetenschappers

3.5.1 Datasets voor eigen gebruik

Deze respondent geeft enkele voorbeelden van datasets die opgezet zijn voor eigen gebruik. Het kan een dataset zijn die opgesteld is door een promovendus die hierover een aantal artikelen publiceert. 'Daarmee is het verhaal wel verteld'. Wel is het belangrijk dat de publicaties kloppen met de data die verzameld zijn. Het gaat dan dus voornamelijk om validatie en veel minder om hergebruik. De kwaliteitscontrole dient in principe door de begeleiders van de promovendus te gebeuren. De respondent denkt niet dat de peer reviewers van de publicaties geneigd zijn om de data te gaan analyseren. 'In die tijd kan ik zelf een ander artikel schrijven'. Wel kunnen inconsistenties door reviewers worden opgemerkt. Een andere mogelijkheid - vaak bij grotere instituten - is een kwaliteitscommissie. De respondent beschrijft een voorbeeld van een kwaliteitscommissie die onderzoekers naar recente publicaties vraagt: zou jij voor morgenochtend kunnen aangeven hoe deze tabel of figuur in je meest recente artikel is geproduceerd? Daar werd veel van geleerd: het bleek in veel gevallen niet zo eenvoudig voor de betreffende onderzoekers te zijn, terwijl de onderzoekers dit direct paraat zouden moeten hebben. Ook werden er audits op datasets uitgevoerd door deze kwaliteitscommissie, waarbij bestaande datasets werden onderzocht.

Volgens deze respondent komen de voor eigen gebruik gemaakte, beperkte datasets heel soms voor hergebruik in aanmerking. Dit kan gebeuren bij retrospectieve meta-analyses, waarin aan de hand van verschenen publicaties datasets worden opgevraagd, samengevoegd en opnieuw geanalyseerd. Hierbij blijkt vaak dat veel individuele onderzoekers hun boekhouding niet op orde hebben. Ook blijkt de kwaliteit van de data sterk te verschillen. Daarnaast hebben onderzoekers soms het gevoel dat ze iets weggeven en zijn daarom niet direct bereid hun dataset ter beschikking te stellen. Tenslotte, onderzoekers veranderen van baan en dan is de dataset vaak verdwenen.

3.5.2 Datasets voor grootschalig(er) gebruik

Daarnaast zijn er datasets die door de grotere groepen onderzoekers worden opgezet met het oog op gebruik door velen. De betreffende respondent noemt als voorbeeld een onderzoeksgroep die aan een 'levende' dataset werkt, vaak met meerdere generaties onderzoekers en een prospectief afgesproken dataset met meerdere groepen. In dit laatste geval zetten verschillende onderzoeksgroepen (bijvoorbeeld vanuit verschillende landen) met dezelfde definities en variabelen datasets op. Deze datasets worden naderhand samengevoegd voor secundaire analyses. Het gaat bijvoorbeeld om cohorten diabetespatiënten uit verschillende landen.

Bij dergelijke datasets horen mechanismen om de kwaliteit te garanderen en een reglement, waarin afgesproken wordt wie wat mag doen. Deze grote datasets leveren stromen publicaties op. Vanuit het perspectief van de respondent kunnen andere onderzoekers de dataset gebruiken, mits voorafgaande daaraan afspraken worden gemaakt. Onderdeel van die afspraken zou moeten zijn dat hergebruikers rapporteren wat er aan het gebruikte deel van de dataset verbeterd kan worden. Dan zijn die gegevens ook beschikbaar zijn voor eventuele nieuwe hergebruikers en voor wanneer de dataset wordt aangevuld met nieuwe data.

3.6 Datapublicaties

Earth System Science Data is een tijdschrift met peer-reviewed artikelen over datasets. De doelstelling van het tijdschrift is om de intellectuele inspanningen van dataproducenten en de inspanningen om deze datasets voor anderen ter beschikking te stellen tot uiting te laten komen in het 'academic record'. Dit kan alleen door een peer-reviewed publicatie. De oprichters van het tijdschrift zagen een behoefte bij onderzoekers hieraan en hun eerste ervaringen (het tijdschrift bestaat een kleine twee jaar) onderschrijven deze behoefte: onderzoekers van laboratoria met grote datasets worden door middel van dit tijdschrift gestimuleerd om hun onderzoeksdatasets te publiceren (soms nadat deze samengevoegd zijn met onderzoeksdatasets van collega-laboratoria).

Een ander tijdschrift dat op vergelijkbare wijze peer-reviewed datapublicaties publiceert is Acta Crystallography E. Andere tijdschriften voor datapublicaties zijn bij de opstellers van het rapport niet bekend.

Earth System Science Data stelt vier vragen in het peer review proces:
1. Is het artikel adequaat om de publicatie van een dataset te ondersteunen?
2. Is de dataset betekenisvol: uniek, bruikbaar en compleet?
3. Is de dataset van hoge kwaliteit (in de zin van datacreatie en datamanagement)?
4. Is de dataset te begrijpen en te hergebruiken door het lezen van het artikel en het downloaden van de dataset?

Wat betreft de 'scholarly merit' van de dataset hanteert ESSD het criterium van 'significance'. Zie de textbox voor toelichting.

Significance

Is there any potential of the data being useful? This is obviously the most problematic decision to take. There are at least three sub-criteria to evaluate:

- Uniqueness: It should not be possible to replicate the experiment or observation on a routine basis. Thus, any dataset on a variable supposed or suspected to reflect changes in the Earth System deserves to be considered unique. This is also the case for cost-intensive data sets which will not be replicated due to financial reasons. A new or improved method should not be trivial or obvious.
- Usefulness: It should be plausible that the data, alone or in combination with other datasets, can be used in future interpretations, for the comparison to model output or to verify other experiments or observations. Other possible uses mentioned by the authors will be considered.
- Completeness: A dataset or collection must not be split artificially, e.g., to increase the possible number of publications. It should contain all data that can be reviewed without undue multiplication of workload and can be re-used in one context by a reader.

from: http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html

3.7 Conclusies

De conclusies over de kwaliteitsbeoordeling van onderzoeksdata zijn in grote lijnen (gerangschikt naar de (huidige) wijze van beschikbaarstelling):

- **Datasets geproduceerd door onderzoeksfaciliteiten en overheidsorganisaties - meerdere mechanismes om de kwaliteit te monitoren en te verbeteren met het oog op hergebruik:** Er is in sommige gevallen peer review vooraf. De validatie en datacleaning gebeurt door het datacentrum van de onderzoeksfaciliteiten volgens vaste protocollen. Er zijn feedbackmechanismen met de gebruikers van de datasets. En in veel gevallen zijn peers betrokken bij de aansturing van de onderzoeksfaciliteit.

- **Supplementary data en replicatiedatasets - kwaliteitstoetsen aanwezig maar mogelijk onvoldoende toegepast:** Deze datasets - gerelateerd aan publicaties - zijn vaak subsets van grotere datasets. De supplementary data horen in principe meegenomen te worden in het peer review proces van de publicatie, maar in de praktijk lijkt dit slechts op beperkte schaal plaats te vinden. Dit geldt ook voor de zogenaamde replicatiedatasets: deze staan open voor de peer reviewers van een publicatie, maar worden volgens de gesprekspartners zelden daadwerkelijk geraadpleegd.

- **Datasets ondergebracht bij data-archieven - kwaliteitstoetsen in veel gevallen aanwezig:** In het proces van opname in het data-archief van een dataset voeren medewerkers van het data-archief kwaliteitstoetsen uit op de metadata en de documentatie bij de dataset. Ook toetsen zij of de inhoud van de dataset binnen de scope van het data-archief past. Enkele data-archieven hebben tevens een 'voorportaal' ingericht. Hierin kunnen onderzoekers hun datasets deponeren zonder voorafgaande kwaliteitscontroles. Een selectie van de

gedeponeerde datasets wordt vervolgens op kwaliteit gecontroleerd en opgenomen in het data-archief. De criteria voor deze selectie zijn niet geëxpliciteerd.

- **Niet rechtstreeks beschikbaar gestelde datasets - kwaliteitstoetsen beperkt aanwezig:** Dit is een groot en weinig bekend gebied. Op basis van de interviews lijkt het zinnig om onderscheid te maken tussen datasets die primair voor eigen gebruik zijn opgesteld door een of enkele onderzoekers en datasets die door grotere groepen onderzoekers zijn opgesteld. Bij de in dit onderzoek besproken gevallen van deze grotere datasets zijn kwaliteitsnormen geëxpliciteerd en mechanismes ingesteld om te monitoren en te handhaven. Wat betreft de primair voor eigen gebruik opgestelde datasets zal in de praktijk vooral de documentatie vaak tekortschieten (onder andere door het tacit knowledge probleem). De bredere beschikbaarstelling van beide categorieën datasets zal sowieso in veel gevallen problematisch zijn.

- **Datapublicaties - kwaliteitstoetsen via peer review gericht op datasets:** Dit is een relatief nieuw fenomeen, waarvan er twee gevallen bekend zijn. Het betreft peer-reviewed tijdschriften waarin de publicaties een dataset inhoudelijk en wat betreft de kwaliteitsnormen beschrijven. De gedachte is om de intellectuele inspanningen bij het creëren van datasets een plaats te geven in het academische record van de producenten van de dataset door middel van een peer-reviewed publicatie en eventuele citaties daarvan en op deze manier het publiceren van datasets te stimuleren.

- Tenslotte merkt een aantal gesprekspartners op dat er behoefte is aan meer training op het gebied van datamanagement om onderzoekers bij het verbeteren van de kwaliteit van hun onderzoek datasets te ondersteunen. Ook kunnen data-audits onderzoekers de ogen openen voor de huidige gebreken in hun datamanagement[9].

[9] Zie voor een methode van een data audit: http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework

4 Wat zijn wenselijke opties voor kwaliteitsbeoordeling onderzoeksdatasets?

4.1 Inleiding

Uit de literatuurstudie en interviews kwam een aantal mogelijke kwaliteitstoetsen voor onderzoeksdatasets naar voren:

- De optie peer review van datasets als onderdeel van peer review van publicaties (zie 4.2).
- De optie peer-reviewed publicaties over datasets (zie 4.3).
- De optie commentaren over de kwaliteit van datasets door hergebruikers (zie 4.4). Dit betreft in feite een soort peer review achteraf: aan hergebruikende onderzoekers wordt gevraagd om commentaren over de kwaliteit van de gehele datasets of van onderdelen ervan bij de dataset te voegen. Deze commentaren (ook wel recensies of annotaties genoemd) worden ter beschikking gesteld aan andere potentiële hergebruikers.
- De optie citeren van datasets(zie 4.5). Deze optie kan worden gepresenteerd als een soort kwaliteitsbeoordeling achteraf: de hergebruikende wetenschappers kunnen de dataset in hun publicaties citeren (enkele initiatieven hebben als doel ervoor te zorgen dat datasets citeerbaar zijn). Het aantal citaties kan als maat fungeren voor de kwaliteit van de dataset.
- De optie ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium(zie 4.6). Hiervoor zijn twee mogelijkheden: (1) trainingen op het gebied van datamanagement en (2) periodieke data audits, waarin de totstandkoming en beheer van datasets worden beoordeeld.

4.1.1 Enquête

De wenselijkheid van deze mogelijkheden is getoetst in een enquête onder een steekproef universitaire wetenschappers. Deze steekproef is representatief voor de universitaire gemeenschap in Nederland. Grote onderzoeksfaciliteiten en/of overheidsinstituten zijn expliciet uitgesloten van de enquête aangezien is gebleken dat daar reeds diverse kwaliteitstoetsen en -mechanismen van onderzoeksdatasets operationeel zijn (zie hoofdstuk 3).

Circa 2800 hoogleraren en universitaire hoofddocenten (UHD) zijn uitgenodigd deel te nemen aan dit onderzoek. Zij ontvingen op 18 augustus 2010 een e-mail met een gecodeerde link naar de vragenlijst en een link om zich af te melden. Degenen die de gecodeerde links nog niet gebruikt hadden, kregen een herinnering op 1 september. De enquête sloot op 13 september. Het resultaat was 396 ingevulde vragenlijsten, een respons van bijna 14 % (zie tabel hieronder; voor aanvullende informatie over de enquête zie Appendix C).

Hoogleraren en UHD's	
totaal aantal e-mailadressen	2869
onbestelbaar retour	58
netto verzonden	**2811**
aantal declined	**396 (14,1%)**
aantal ingevulde vragenlijsten	**392 (13,9%)**

4.2 Peer review datasets als onderdeel peer review publicaties

De eerste sectie van de vragenlijst betrof de peer review van datasets als onderdeel van de peer review van publicaties. Het overgrote deel van de respondenten geeft aan actief te zijn als peer reviewer van publicaties (95,2%). Meer dan de helft is ook lid van de redactie van een peer-reviewed tijdschrift (56,9 %).

A. Peer review van datasets als onderdeel van peer review van publicaties	Technische en Natuurweten-schappen (n=61)	Sociale en Geesteswetenschappen (n=153)	Levenswetenschappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percentage (alle respondenten)	Statistisch significant?
	% (enigszins mee eens)					
1. In mijn vakgebied vragen veel tijdschriften om zo'n replicatiedataset ter beschikking te stellen.	9,8	6,5	17,0	16,1	11,7	nee
2. In mijn vakgebied heeft de peer reviewer normaliter toegang tot de achterliggende dataset bij een publicatie en wordt de dataset meegenomen in de beoordeling van de publicatie.	19,7	7,8	18,4	12,9	14,0	nee
3. Het tegelijkertijd beoordelen van de publicatie en de achterliggende dataset is voor een peer reviewer in mijn vakgebied haalbaar.	39,3	26,8	37,4	22,6	32,4	ja
4. Ik vind het belangrijk dat de achterliggende dataset samen met de publicatie wordt beoordeeld in het peerreviewproces.	47,5	39,9	51,0	45,2	45,7	nee

De tabel presenteert de resultaten per vakgebied: 61 respondenten Technische en Natuurwetenschappen, 153 respondenten Sociale en Geesteswetenschappen en 147 respondenten Levenswetenschappen. De overige respondenten behoorden niet eenduidig tot deze categorieën[10].

Conclusies

Uit deze resultaten zijn de volgende conclusies te trekken:

- De tijdschriften in het vakgebied Levenswetenschappen lopen voorop in het vereisen van terbeschikkingstelling van replicatiedatasets bij het publiceren van een artikel. Gevolgd door tijdschriften in Technische en Natuurwetenschappen en in de Sociale en Geesteswetenschappen.
- Toegang voor de peer reviewers tot deze replicatiedatasets tijdens de peer review van de publicatie blijkt een redelijk normale gang van zaken te zijn volgens de respondenten uit de Natuur- en Levenswetenschappen. Voor de sociale en geesteswetenschappen is dit veel minder normaal.
- Is het wel haalbaar om de dataset mee te nemen in de peer review van het tijdschriftartikel? Over het geheel genomen blijkt dat 32% van de respondenten dit haalbaar vindt, terwijl bijna 44% denkt dat dit niet haalbaar is. De respondenten uit de Sociale en Geesteswetenschappen denken nog sterker dan hun collega's dat dit niet haalbaar is.
- Hoewel aan de haalbaarheid wordt getwijfeld, vindt een duidelijke meerderheid dit wel wenselijk: bijna 46% vindt het belangrijk terwijl ruim 28% het niet (zo) belangrijk vindt. De levenswetenschappers en de natuurwetenschappers lopen hierin voorop.

Haalbaarheid

Waarom vinden veel respondenten het niet haalbaar dat de peer review van de dataset tegelijkertijd met de peer review van het tijdschriftartikel plaatsvindt? De belangrijkste redenen zijn:

- De tijd die dit zal vergen van de peer reviewer (dit argument wordt door de meeste respondenten naar voren gebracht).

[10] 28 respondenten gaven meerdere vakgebieden aan, 3 respondenten verzuimden enig vakgebied aan te geven.

- Datasets kunnen niet altijd toegankelijk worden gesteld vanwege privacy, contractuele of wettelijke redenen.
- Voor kwalitatieve datasets wordt het minder relevant gevonden.
- Enkele respondenten stellen dat de peer reviewer wel naar de methoden dient te kijken, maar dat de dataset zelf niet toegankelijk hoeft te zijn. In dit verband stelde sommigen dat het systeem op vertrouwen gebaseerd moet zijn.
- Enkele respondenten stellen dat sommige datasets door peer reviewers niet geraadpleegd kunnen worden omdat deze specifieke software of apparatuur vereisen.
- Eén respondent stelt voor om hiervoor een apart persoon per tijdschrift aan te stellen.

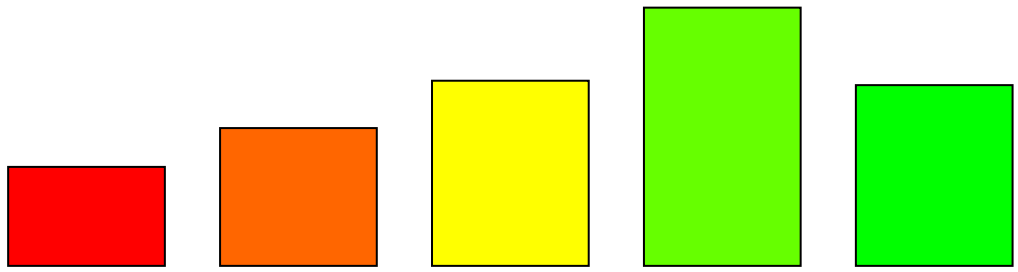
4.3 Datapublicaties: peer-reviewed publicaties over datasets

Het onderwerp datapublicaties werd als volgt geïntroduceerd in de vragenlijst: “Er zijn enkele peer-reviewed tijdschriften voor artikelen gewijd aan datasets. De artikelen betreffen een uitgebreide beschrijving van de dataset gericht op de wijdere betekenis en de kwaliteit ervan. Het idee is dat deze peer-reviewed datapublicaties onderzoekers stimuleren om hun datasets ter beschikking te stellen, omdat ze zodoende een aparte bron van wetenschappelijk prestige kunnen vormen.”

Conclusies

- Ongeveer de helft van de respondenten (50,8 %) ziet een waardevolle rol voor een dergelijk datapublicatietijdschrift in hun vakgebied (zie de staafdiagram hieronder).

5. Voor mijn vakgebied zou een dergelijk tijdschrift met data publicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets. **(n=392)**



- De verschillen tussen de drie vakgebieden zijn klein (zie tabel hieronder).

B. Datapublicaties: peer-reviewed publicaties over datasets	Technische en Natuurweten-schappen (n=61)	Sociale en Geestesweten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
	% (enigszins mee eens)					
5. Voor mijn vakgebied zou een dergelijk tijdschrift met datapublicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets.	47,5	50,3	51,0	58,1	50,8	nee

4.4 Commentaren over kwaliteit door hergebruikers van datasets

In de derde sectie van de vragenlijst kwamen de commentaren van hergebruikers over de kwaliteit van onderzoeksdatasets aan de orde. De resultaten per discipline staan de tabel hieronder.

C. Commentaren over kwaliteit door hergebruikers van datasets	Technische en Natuurweten-schappen (n=61)	Sociale en Geestesweten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
	% (enigszins mee eens)					
6. Voor latere hergebruikers van datasets zouden deze commentaren over kwaliteit waardevol kunnen zijn.	72,1	79,7	87,1	80,6	81,4	nee
7. Als hergebruiker zou ik zeker de moeite nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset die ik heb hergebruikt.	65,6	67,3	78,9	77,4	72,2	nee
8. Als producent van een dataset zou ik dergelijke commentaren over kwaliteit van anderen verwelkomen.	70,5	74,5	82,3	80,6	77,3	nee

Conclusies
De respondenten uit de drie disciplines wijken nauwelijks af wat betreft hun antwoorden. In het algemeen zijn zij positief over het idee van commentaren door hergebruikers over de kwaliteit van datatsets:

- Ruim 80% denkt dat deze commentaren waardevol kunnen zijn voor latere hergebruikers van de datasets.
- Ruim 70% stelt zelf in de positie van hergebruiker de moeite te zullen nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset.
- Bijna 80% zou in de positie van dataproducent dergelijke commentaren over kwaliteit door hergebruikers verwelkomen.

4.5 Citeren van datasets

De enquête peilde vervolgens de mening van de respondenten over het citeren van datasets.

D. Citeren van datasets	Technische en Natuurweten-schappen (n=61)	Sociale en Geesteswe-ten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
	% (enigszins mee eens)					
9. Als hergebruiker zou ik zeker de dataset citeren met deze mogelijkheid	80,3	73,9	74,1	83,9	75,8	nee
10. Als wetenschapper zou ik zo'n citatie mogelijkheid van 'mijn' dataset verwelkomen.	73,8	69,9	74,1	74,2	72,4	nee

Conclusies

De respondenten uit de drie disciplines zijn vrijwel eensluidend in hun antwoorden:

- Meer dan driekwart van de respondenten stelt in de positie van hergebruiker de dataset zeker te citeren, wanneer daarvoor een mogelijkheid zou zijn.
- Meer dan 70% stelt in de positie van dataproducent zo'n citatiemogelijkheid voor de eigen dataset te verwelkomen.

4.6 Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium

Tenslotte kregen de respondenten de twee mogelijkheden voorgelegd van ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium.

E. Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium	Technische en Natuurweten-schappen (n=61)	Sociale en Geesteswe-ten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
	% (enigszins mee eens)					
11. Trainingen op het gebied van datamanagement zouden op mijn vakgebied voorzien in een behoefte van veel vakgenoten.	37,7	45,8	63,3	67,7	52,8	ja
12. Data-audits, waarin de totstandkoming en het beheer van datasets onder de loep worden genomen, zouden in mijn vakgebied een impuls aan verbeteringen in datamanagement kunnen geven.	36,1	52,9	62,6	61,3	54,6	ja

Conclusies

Hierover blijken de respondenten van de verschillende disciplines sterk van mening te verschillen:

- Trainingen op het gebied van datamanagement zouden volgens ruim 63% van de respondenten van de Levenswetenschappen in een behoefte voorzien. De respondenten uit de Technische en Natuurwetenschappen zien deze behoefte veel minder (een kleine 38%), terwijl de respondenten uit de sociale wetenschappen en Geesteswetenschappen met een kleine 46% een tussenpositie innemen.
- Data-audits worden eveneens door respondenten uit de Levenswetenschappen zinvol geacht (ruim 62%). De respondenten van de Technische en Natuurwetenschappen denken hierover minder positief (36 %) en de respondenten uit de Sociale wetenschappen en Geesteswetenschappen nemen ook hier een tussenpositie in (bijna 53%).

4.7 Populariteit van negen opties voor kwaliteitsbevordering van datasets

De hiervoor behandelde secties van de vragenlijst omvatten zes opties voor kwaliteitsbeoordeling van datasets. Uit de interviews en de literatuurstudie is nog een aantal andere mogelijkheden naar voren gekomen:

1. Een verplichte datamanagementparagraaf in onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers.
2. Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode.
3. Een gedragscode voor onderzoekers over datamanagement en het beschikbaar stellen van datasets.

In het totaal zijn er dus negen opties in dit onderzoek naar voren gekomen. Alle negen opties werden voorgelegd aan de respondenten met een tweetal vragen:

1. Van welke van deze opties zou de meeste stimulans voor de kwaliteitsbeoordeling van datasets in uw vakgebied uitgaan?
2. Welke van deze opties zouden bij u op bezwaren stuiten?

De respondenten konden bij elke vraag maximaal drie opties aan vinken. De resultaten zijn hieronder in een tabel weergegeven.

F. Uw visie	Technische en Natuurweten-schappen (n=61)	Sociale en Geestesweten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
	% aangevinkt					
13. De meeste stimulans voor de kwaliteitsbeoordeling van datasets in mijn vakgebied zal waarschijnlijk uitgaan van:						
• Peer review van de dataset als onderdeel van peer review van de publicatie	39,3	36,6	42,2	35,5	39,0	nee
• Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	41,0	34,0	25,9	32,3	31,9	nee
• Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	27,9	27,5	34,7	25,8	30,1	nee
• Citeren van datasets	32,8	37,9	25,2	32,3	31,9	nee
• Aanbieden van datamanagementtrainingen	6,6	17,0	21,1	16,1	16,8	nee
• Instellen van periodieke data-audits	6,6	5,9	24,5	16,1	13,8	ja
• Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	4,9	11,1	19,7	19,4	14,0	ja
• Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	59,0	40,5	29,9	45,2	39,8	ja
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	13,1	22,9	32,7	32,3	25,8	ja
14. Welke van deze opties zouden bij u op bezwaren stuiten?	29,5	25,5	25,2	22,6	25,8	nee
• Peer review van de dataset als onderdeel van peer review van de publicatie	6,6	5,2	8,8	9,7	7,1	nee
• Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	1,6	5,2	8,2	3,2	5,6	nee
• Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	1,6	4,6	1,4	3,2	2,8	nee
• Citeren van datasets	4,9	3,3	1,4	3,2	2,8	nee
• Aanbieden van datamanagementtrainingen	44,3	21,6	19,0	29,0	24,7	ja
• Instellen van periodieke data-audits	39,3	30,7	29,3	25,8	31,1	nee

• Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	14,8	16,3	26,5	32,3	21,2	ja
• Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	14,8	7,2	6,1	6,5	7,9	nee
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	29,5	25,5	25,2	22,6	25,8	nee

Hieronder zijn de resultaten gepresenteerd in een populariteitindex voor de negen opties: het percentage respondenten dat van de optie een stimulans vindt uitgaan minus het percentage respondenten dat bezwaren tegen de optie heeft staat in de derde kolom. Dit wordt gezien als een maat voor de populariteit van de optie. De opties zijn in drie tabellen (per discipline) gerangschikt naar de mate van populariteit.

Technische en Natuurwetenschappen, populariteitsindex (n=61)	**Stimulans**	**Bezwaren**	**Maat populariteit**
Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	59	14,8	44,2
Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	41	6,6	34,4
Citeren van datasets	32,8	1,6	31,2
Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	27,9	1,6	26,3
Peer review van de dataset als onderdeel van peer review van de publicatie	39,3	29,5	9,8
Aanbieden van datamanagementtrainingen	6,6	4,9	1,7
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	13,1	14,8	-1,7
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	4,9	39,3	-34,4
Instellen van periodieke data-audits	6,6	44,3	-37,7

Sociale en Geesteswetenschappen, populariteitsindex (n=153)	**Stimulans**	**Bezwaren**	**Maat populariteit**
Citeren van datasets	37,9	4,6	33,3
Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	34	5,2	28,8
Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	40,5	16,3	24,2
Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	27,5	5,2	22,3
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	22,9	7,2	15,7
Aanbieden van datamanagementtrainingen	17	3,3	13,7
Peer review van de dataset als onderdeel van peer review van de publicatie	36,6	25,5	11,1
Instellen van periodieke data-audits	5,9	21,6	-15,7
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	11,1	30,7	-19,6

Levenswetenschappen, populariteitsindex (n=147)	Stimulans	Bezwaren	Maat populariteit
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	32,7	6,1	26,6
Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	34,7	8,2	26,5
Citeren van datasets	25,2	1,4	23,8
Aanbieden van datamanagementtrainingen	21,1	1,4	19,7
Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	25,9	8,8	17,1
Peer review van de dataset als onderdeel van peer review van de publicatie	42,2	25,2	17
Instellen van periodieke data-audits	24,5	19	5,5
Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	29,9	26,5	3,4
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	19,7	29,3	-9,6

De tabel hieronder presenteert de percentages uit de enquête per discipline.

Populariteitsmaat per discipline	Technische en natuurweten-schappen (n=61)	Sociale en Geesteswe-ten-schappen (n=153)	Levensweten-schappen (n=147)
Opzetten van datapublicaties: peer-reviewed beschrijvingen van datasets	34,4	28,8	17,1
Citeren van datasets	31,2	33,3	23,8
Commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	26,3	22,3	26,5
Open Access-beschikbaarstelling van datasets, desgewenst na een embargoperiode	44,2	24,2	3,4
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	-1,7	15,7	26,6
Peer review van de dataset als onderdeel van peer review van de publicatie	9,8	11,1	17
Aanbieden van datamanagementtrainingen	1,7	13,7	19,7
Instellen van periodieke data-audits	-37,7	-15,7	5,5
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	-34,4	-19,6	-9,6

Conclusies

Er zijn enkele overeenkomsten te noteren tussen de drie disciplines:

- Het citeren van datasets en commentaren over kwaliteit door hergebruikers zijn bij de respondenten van alle drie disciplines populaire opties, terwijl er weinig respondenten zijn die hier bezwaren in zien.
- Het opzetten van datapublicaties is populair bij respondenten uit alle vakgebieden, in het bijzonder bij de respondenten in de Technische en Natuurwetenschappen.
- Peer review van de dataset als onderdeel van peer review van de publicatie wordt door de respondenten van alle drie disciplines in vrijwel gelijke mate als stimulerend gezien voor de kwaliteit van de datasets, maar stuit ook bij hoge percentages van de respondenten op bezwaren.
- Het aanbieden van datamanagementtrainingen scoort redelijk neutraal op de populariteitsindex: een redelijke percentage respondenten uit de Sociale en Geesteswetenschappen en uit de Levenswetenschappen ziet dit als stimulerend, . De respondenten uit de Technische en Natuurwetenschappen zien dit weliswaar minder, maar de percentages wijken niet significant af.

Daarnaast zijn er ook opvallende (statistisch significante) verschillen te noteren tussen de drie disciplines:

- Wat betreft de Open Access-beschikbaarstelling van datasets zijn er grote verschillen: deze optie is populair bij de Technische en Natuurwetenschappen, redelijk populair bij de Sociale en Geesteswetenschappen, maar nauwelijks populair bij de Levenswetenschappen.
- Een gedragscode voor onderzoekers over datamanagement en beschikbaarstelling van datasets is populair bij de Levenswetenschappen, matig populair bij de Sociale en Geesteswetenschappen, maar nauwelijks populair bij de Technische en Natuurwetenschappen.
- Het instellen van periodieke data-audits en een verplichte datamanagementparagraaf in onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers is het minst populair bij alle respondenten. Echter, de respondenten uit de Technische en Natuurwetenschappen en uit de Sociale en Geesteswetenschappen zijn hierover uitgesproken negatief, terwijl de voorstellen bij de respondenten uit de Levenswetenschappen slechts impopulair zijn.

4.8 Achtergrondkenmerken van de respondenten

De achtergrondkenmerken van de respondenten zijn in de tabel hieronder per discipline weergegeven.

G. Enkele vragen over uw professionele achtergrond	Technische en Natuurweten-schappen (n=61)	Sociale en Geesteswe-ten-schappen (n=153)	Levensweten-schappen (n=147)	Interdisciplinair (meer dan 1 Hoofdgebied) (n=31)	Totaal percen-tage (alle respondenten)	Statistisch significant?
15. Mijn huidige functie is:						
• hoogleraar	55,7	59,5	63,9	61,3	60,7	nee
• universitair hoofddocent	42,6	38,6	29,9	32,3	35,5	
• universitair docent	0,0	0,7	2,0	0,0	1,0	
• postdoc medewerker	0,0	0,7	1,4	0,0	0,8	
• promovendus/AIO/OIO	0,0	0,0	1,4	0,0	0,5	
• anders	1,6	0,7	1,4	6,5	1,5	
16. Als wetenschapper vervul ik de volgende rollen: [MEERDERE ANTWOORDEN MOGELIJK]						
• Ik ben lid van een redactie van een peer-reviewed tijdschrift	50,8	57,5	60,5	48,4	56,9	nee
• Ik ben actief als peer reviewer van publicaties	91,8	94,8	98,0	90,3	95,2	nee
17. Wat betreft datasets vervul ik de volgende rollen: (MEERDERE ANTWOORDEN MOGELIJK]						
• Ik ben (mede) producent van datasets	57,4	67,3	78,9	77,4	70,9	ja
• Ik heb wel eens mijn datasets ter beschikking gesteld aan derden	63,9	58,2	61,2	54,8	59,9	nee
• Ik ben hergebruiker van datasets (van andere onderzoekers en/of van grote onderzoeksfaciliteiten)	47,5	51,6	47,6	51,6	49,5	nee

Daaruit blijken de volgende zaken:

- Het overgrote deel van de respondenten is hoogleraar of universitair hoofddocent (96,2%) conform het gebruikte mailingbestand. De lage percentages respondenten met andere functies zijn waarschijnlijk een gevolg van het doorsturen van de uitnodigingsmail aan hen.
- De gevorderde carrière van de respondenten blijkt tevens uit de rollen die zij vervullen in de academische gemeenschap:
  - 56,9% is lid van een redactie van een peer-reviewed tijdschrift
  - 95,2% is actief als peer reviewer van publicaties
- Ook wat betreft onderzoeksdatasets hebben de respondenten ruime ervaring:
  - 70,9% is (mede)producent van datasets
  - 59,9% heeft wel eens zijn datasets ter beschikking gesteld aan derden
  - 49,5% is (ook) hergebruiker van datasets

De volgende paragraaf presenteert de resultaten van crossanalyses op deze drie kenmerken wat betreft onderzoeksdatasets.
- De achtergrondkenmerken van de respondenten uitgesplitst naar discipline verschillen niet statistisch significant op één uitzondering na: het percentage (mede)producent van datasets verschilt tussen de disciplines en is onder de respondenten van de Technische en Natuurwetenschappen lager. Dit gegeven heeft overigens geen invloed op de hierboven vermelde conclusies[11].

4.9 Relatie hergebruikers, dataproducenten en ervaring met het ter beschikking stellen van de eigen datasets

Er zijn drie crossanalyses uitgevoerd:
1. een vergelijking tussen (mede)dataproducenten en niet-dataproducenten
2. een vergelijking tussen hergebruikers van datasets en niet-hergebruikers
3. een vergelijking tussen de respondenten die de eigen dataset(s) wel eens ter beschikking hebben gesteld en de respondenten die dat niet hebben gedaan.

De resultaten van deze vergelijkingen zijn gepresenteerd in appendix C 2.

Conclusies

De voornaamste conclusies zijn:
- Er is een onderling verband tussen hergebruikers en (mede)producenten: 74,7% van de hergebruikers heeft wel eens eigen datasets ter beschikking gesteld aan derden en 75,8 % is (mede)producent van datasets.
- In vergelijking met niet-hergebruikers tonen hergebruikers zich in vrijwel alle gevallen meer enthousiast over de in de enquête gepresenteerde vormen van kwaliteitsbeoordeling en kwaliteitstoetsen. Er is één uitzondering: de hergebruikers tonen zich minder enthousiast over een gedragscode (statistisch significant).
- In vergelijking met niet-dataproducenten betonen dataproducenten zich eveneens in de meeste gevallen meer enthousiast over de in de enquête gepresenteerde vormen van kwaliteitsbeoordeling en kwaliteitstoetsen op de volgende uitzonderingen na: dataproducenten zijn minder enthousiast over Open Access-beschikbaarstelling van datasets (en zien meer bezwaren hierin) en hebben ook meer bezwaren tegen het peer review van de dataset als onderdeel van het peer review van de publicatie.
- De respondenten die hun dataset(s) wel eens ter beschikking gesteld hebben zijn, in vergelijking met de respondenten die dat niet hebben gedaan, juist positiever over Open Access-beschikbaarstelling van datasets. Ook zien deze respondenten (nog) meer in datapublicaties en in het citeren van datasets.

[11] Zoals uit de volgende paragraaf blijkt, zien dataproducenten minder in Open Access- beschikbaarstelling dan niet-dataproducenten. Zou het lagere percentage dataproducenten onder de respondenten van de Technische en Natuurwetenschappen de reden kunnen zijn van het verschil wat betreft de visie op Open Access in vergelijking met andere disciplines? Om dit uit te zoeken is tevens een crossanalyse uitgevoerd, waarbij uitsluitend de dataproducenten per discipline vergeleken werden. Uit deze analyse bleek dat de dataproducenten van de Technische en Natuurwetenschappen eveneens meer zien in Open Access-beschikbaarstelling dan de dataproducenten van de andere disciplines.

5 Conclusies en aanbevelingen

5.1 Conclusies

Wat houdt kwaliteitsbeoordeling van onderzoeksdata in?

- Kwaliteitsbeoordeling van onderzoeksdatasets focust vooral op kwaliteitszorg bij de creatie van de dataset, en op de metadata. De kwaliteitstoetsen op deze aspecten zijn bedoeld om de toegankelijkheid van de dataset voor hergebruikers te verhogen.
- Een kwaliteitsbeoordeling van de wetenschappelijke waarde (scholarly merit) beperkt zich tot het bepalen of de dataset mogelijk zinvol te hergebruiken is door andere wetenschappers. Een directe kwaliteitsbeoordeling op de werkelijke wetenschappelijke waarde vindt niet plaats. De gesprekspartners wijzen dit ook zowel op theoretische als praktische gronden af.

Wat is de huidige situatie van kwaliteitsbeoordeling van datasets?

Er wordt een onderscheid gemaakt in datasets van onderzoeksfaciliteiten en/of overheidsinstituten waar in principe geen overdracht van de dataset plaatsvindt, en datasets die worden gecreëerd door onderzoeksgroepen.

- Bij de onderzoeksfaciliteiten en/of overheidsinstituten blijken een of meer mechanismen operationeel zijn die de kwaliteit van de datasets waarborgen: peer review vooraf over de data-aanvraag, validatie en datacleaning door het datacentrum van de onderzoeksfaciliteiten, feedback van gebruikers en betrokkenheid van peers bij de aansturing van de onderzoeksfaciliteiten. Aanvullende kwaliteitstoetsen lijken in deze gevallen dan ook overbodig en zijn in deze studie buiten beschouwing gelaten.
- Voor de opslag en beschikbaarstelling van de datasets van onderzoeksgroepen zijn een reeks van mogelijkheden:
    - De datasets kunnen bewaard worden in een data-archief. Bij het moment van overdracht vindt in een aantal gevallen een kwaliteitstoets plaats: op de metadata, op de documentatie bij de dataset en een toets of de inhoud van de dataset binnen de scope van het data-archief past.
    - De datasets kunnen als replicatiedataset (of supplementaire data) bij een tijdschriftartikel gepubliceerd worden. De supplementary data horen in principe ‘meegenomen te worden’ in de peer review van de publicatie, maar in de praktijk lijkt dit slechts op beperkte schaal plaats te vinden. Dit geldt ook voor de zogenaamde replicatiedatasets: deze staan open voor de peer reviewers van een publicatie, maar worden volgens de gesprekspartners zelden ook daadwerkelijk geraadpleegd.
    - In tot op heden uitzonderlijke gevallen kan de dataset als datapublicatie worden gepubliceerd in daartoe opgerichte datapublicatie-tijdschriften. De dataset wordt dan uitvoerig beschreven in het artikel en deze publicatie doorloopt de peer review en wordt eventueel geciteerd.
    - De datasets kunnen door de onderzoeksgroep zelf kunnen worden bewaard. Dit is een groot en weinig bekend gebied. Op basis van de interviews lijkt het zinnig om een onderscheid te maken tussen datasets die een of enkele onderzoekers primair voor eigen gebruik hebben opgesteld, en datasets die grotere groepen onderzoekers hebben opgesteld. In het laatste geval zijn vaak kwaliteitsnormen geëxpliciteerd en mechanismes ingesteld om deze te monitoren en te handhaven. Wat betreft de primair voor eigen gebruik opgestelde datasets zal in de praktijk vooral de documentatie vaak tekortschieten (onder andere door het tacit knowledge-probleem). De bredere beschikbaarstelling van beide categorieën datasets zal sowieso in veel gevallen problematisch zijn.

Opties voor en wenselijkheid van kwaliteitsbeoordeling onderzoeksdatasets

- In dit onderzoek is een aantal mogelijkheden naar voren gekomen voor kwaliteitsbeoordeling van onderzoeksdatasets:
    - Kwaliteitsbeoordeling voorafgaand aan de beschikbaarstelling.
        - Als onderdeel van de peer review van het artikel: uit de enquête blijkt dat veel wetenschappers dit weliswaar wenselijk vinden, maar niet haalbaar achten omdat peer reviewers reeds overbelast zijn en dit extra veel tijd zal gaan kosten.

- In de peer review van de datapublicatie (artikel met beschrijving dataset): uit de enquête blijkt dat veel wetenschappers het starten van tijdschriften voor dergelijke datapublicaties op hun vakgebied toejuichen.

o Kwaliteitsbeoordeling na de beschikbaarstelling.
- Commentaren over kwaliteit door hergebruikers van datasets: dit betreft in feite een soort peer review achteraf. Hergebruikende onderzoekers krijgen de vraag om commentaren over de kwaliteit van de gehele datasets of van onderdelen bij de dataset te voegen. Deze commentaren/recensies/annotaties worden ter beschikking gesteld aan andere potentiële hergebruikers. Uit de enquête blijkt dat velen dit als een wenselijke optie zien.
- Citeren van datasets: de hergebruikende wetenschappers kunnen de dataset in hun publicaties citeren (enkele initiatieven hebben als doel ervoor te zorgen dat datasets als zodanig citeerbaar zijn). Het aantal citaties kan dan als maat fungeren voor de kwaliteit van de dataset. Ook dit zien veel wetenschappers als een wenselijke optie.

Wenselijkheid van opties voor verbetering van de kwaliteit van datasets
De enquête heeft negen opties voor verbetering van de kwaliteit van datasets voorgelegd aan de respondenten. Zij konden aangeven van welke drie opties zij de meeste stimulans zien uitgaan voor hun vakgebied en bij welke drie opties zij de meeste bezwaren hebben.

De tabel hieronder geeft de resultaten van de populariteit van de verschillende opties in vereenvoudigde vorm weer, uitgesplitst naar het vakgebied van de respondenten (Technische en Natuurwetenschappen, Sociale en Geesteswetenschappen en Levenswetenschappen. De percentages zijn afgerond op 10 punten en ieder tiental aangeduid met + resp. een –.

	Technische en natuurweten-schappen	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	++++	++	0
Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	0	++	+++
Peer review van de dataset als onderdeel van peer review van de publicatie	+	+	++
Aanbieden van datamanagementtrainingen	0	+	++
Instellen van periodieke data-audits	––––	––	+
Verplichte datamanagementparagraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	–––	––	–

Drie opties zijn populair bij alle disciplines, terwijl weinig respondenten hierin bezwaren zien:
- Het opzetten van tijdschriften voor datapublicaties
- Het citeren van datasets
- Commentaren over kwaliteit door hergebruikers

Twee opties zijn het minst populair (dit geldt overigens in mindere mate voor de respondenten uit de Levenswetenschappen):
- Een verplichte datamanagementparagraaf in onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers
- Het instellen van periodieke data-audits

Een aantal opties is (vrijwel) uitsluitend populair bij een of twee vakgebieden:

- Open Access-beschikbaarstelling van datasets na een eventuele embargoperiode is populair bij respondenten van de Technische en Natuurwetenschappen, redelijk populair bij de Sociale en Geesteswetenschappen. Deze optie scoort echter laag bij de respondenten uit de Levenswetenschappen.
- Het vaststellen van een gedragscode voor onderzoekers over datamanagement en beschikbaarstelling van datasets is populair bij de respondenten uit de Levenswetenschappen, maar scoort laag bij de Technische en Natuurwetenschappen. De respondenten uit de Sociale en Geesteswetenschappen nemen hier een tussenpositie in.

Twee opties bevinden zich tussen populair en impopulair in:

- Peer review van de dataset als onderdeel van peer review van de publicatie wordt door de respondenten van alle drie disciplines in vrijwel gelijke mate als stimulerend gezien voor de kwaliteit van de datasets. Maar deze optie stuit ook bij hoge percentages van de respondenten op bezwaren.
- Het aanbieden van datamanagementtrainingen wordt door een redelijke percentages respondenten uit de Sociale en Geesteswetenschappen en uit de Levenswetenschappen als stimulerend gezien, terwijl weinig respondenten uit deze disciplines hierin bezwaren zien. De respondenten uit de Technische en Natuurwetenschappen wijken hierin enigszins af. Slechts een laag percentage ziet een stimulerende werking hiervan uitgaan, terwijl een vergelijkbaar percentage hierin bezwaren ziet. Overigens is het geen significant verschil.

5.1.1 Verschillen tussen disciplines

Open Access van data scoort hoog bij de Technische en Natuurwetenschappen terwijl deze discipline weinig behoefte heeft aan een gedragscode, data-audits en datamanagementtrainingen. De Levenswetenschappen reageren complementair. De Sociale en Geesteswetenschappen bevinden zich tussen beide in. De analyse van de opstellers van het rapport is dat deze verschillen met name voortkomen uit de worsteling van levenswetenschappers (en in mindere mate van sociale en geesteswetenschappers) met gedragsregels voor het beschikbaar stellen van onderzoeksdata. Het gaat dan waarschijnlijk met name om privacy en ethiek. Deze overwegingen spelen nauwelijks een rol bij de Technische en Natuurwetenschappen waar datasets zelden over mensen gaan. Een gedragscode en een data-audit kunnen houvast bieden bij het omgaan met privacy en ethiek.

5.2 Aanbevelingen: een per vakgebied verschillende aanpak

Op grond van de gebleken verschillen tussen de disciplines bevelen de opstellers van het rapport aan per discipline een verschillende aanpak te kiezen om de kwaliteit van datasets te bevorderen.

5.2.1 Voor het gehele wetenschapsgebied

Drie opties zijn relevant voor alle drie vakgebieden en dienen voor het hele wetenschapsgebied gestimuleerd worden:

- Het mogelijk maken van citeren van datasets. Standaarden over de technische en inhoudelijke afspraken hierover kunnen mede worden ontwikkeld in het SURFshare-programma of het vervolg daarop.
- De mogelijkheid om hergebruikers commentaren over kwaliteit te laten toevoegen aan datasets. Dit is met name van belang voor data-archieven. De in Nederland bekende data-archieven van DANS, MPI en 3TUDatacentrum kunnen dit gezamenlijk technisch ontwikkelen. SURFfoundation kan de inhoudelijke afspraken en regels hierover ontwikkelen in samenspraak met de VSNU (zie ook gedragscode hieronder).
- Het oprichten van tijdschriften met datapublicaties. Wetenschappelijke verenigingen kunnen dit stimuleren, door het zelf te doen of uitgevers hiervoor te vragen.

5.2.2 Voor de Technische en Natuurwetenschappen en de Sociale en Geesteswetenschappen

Eén optie is in eerste instantie vooral relevant voor de Technische en Natuurwetenschappen en de Sociale en Geesteswetenschappen:

- Het bevorderen van Open Access-beschikbaarstelling van datasets, eventueel na een embargoperiode. Dit ligt met name op de weg van onderzoeksfinanciers. Onze aanbeveling is om hierover regels op te stellen voor de Technische en Natuurwetenschappen. Deze regels kunnen mogelijk in aangepaste vorm ook acceptabel zijn voor wetenschappers uit de Sociale en Geesteswetenschappen. Bij de Levenswetenschappen zullen regels over Open Access-toegankelijkheid van datasets op grote weerstanden stuiten. Deze weerstanden - waarschijnlijk gebaseerd op overwegingen van privacy en ethiek - dienen eerst weggenomen te worden door afspraken en regels.

5.2.3 Voor de Levenswetenschappen en de Sociale en Geesteswetenschappen

Eén optie is in eerste instantie vooral relevant voor de Levenswetenschappen en de Sociale en Geesteswetenschappen:

- Het vaststellen van een gedragscode over datamanagement en het beschikbaar stellen van datasets. Dit kan SURFfoundation in samenwerking met onder andere de VSNU en NWO ter hand nemen. De gedragscode moet aandacht besteden aan de weerstanden van (met name) levenswetenschappers tegen Open Access-beschikbaarstelling van datasets. Zij kunnen aansluiting zoeken bij het initiatief van de NFUMC (#25). We verwachten dat Open Access-beschikbaarstelling van data meer acceptabel wordt voor levenswetenschappers als er duidelijke regels komen over hoe om te gaan met privacy, ethiek en competitie door andere wetenschappers (embargo, verplicht citeren en dergelijke).
  In deze gedragscodes moeten ook gedragsregels opgenomen worden voor het hergebruik van onderzoeksdatasets van derden (onder andere ten aanzien van een co-auteurschap), het citeren van een dataset, het raadplegen van de dataproducent bij andere onderzoeksvragen, en het toevoegen van commentaren over kwaliteit aan de oorspronkelijke dataset.

5.2.4 In andere vorm implementeren

Twee opties stuiten op grote weerstanden en we beleven dan ook aan deze niet in de besproken vorm te implementeren:

- Een verplichte datamanagementparagraaf in onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers. Waarom respondenten deze optie zo onwenselijk vinden, is niet helemaal duidelijk: enkele opmerkingen van de respondenten wijzen in de richting van bureaucratie en rompslomp. Mogelijk is de intentie van deze optie verkeerd overgekomen. Het idee was om de opsteller van het onderzoeksvoorstel twee of drie eenvoudige vragen te laten beantwoorden om het bewust omgaan met datasets te bevorderen. Enkele voorbeeldvragen zijn: Dient de onderzoeksdataset bewaard te worden na afloop van het onderzoek? Indien ja, hoe stelt u deze dan beschikbaar?
  We geven NWO en andere onderzoeksfinanciers in overweging om in afstemming met de wetenschappers van de betreffende disciplines een dergelijke ‘lichte’ aanpak te ontwikkelen. In dit verband lijkt het nuttig om er op te wijzen dat het reeds ingevoerde NWO-DANS data-contract voor de Sociale en Geesteswetenschappen (referentie 26) algemeen aanvaard is.
- Periodieke data-audits. Ook dit stuit op grote weerstanden bij veel respondenten van alle disciplines, hoewel minder bij levenswetenschappen. De aangegeven redenen voor deze weerstand betreffen eveneens bureaucratie en rompslomp. Het implementeren van data-audits zou met name op het terrein van de individuele onderzoeksinstellingen kunnen liggen. In onze optiek worden data-audits alleen geaccepteerd als zij wetenschappers duidelijk ondersteunen bij het omgaan met privacy en ethiek rond de beschikbaarstelling van datasets.

5.2.5 Voor de wetenschappelijke tijdschriften

Het is opmerkelijk dat volgens de respondenten een belangrijk percentage van de tijdschriften in hun vakgebied de beschikbaarstelling van replicatiedatasets en/of supplementary data bij het publiceren van een artikel al vereist. Een stelling hierover wordt door 17% van de respondenten uit de Levenswetenschappen onderschreven, door 9,8% van de respondenten uit de Technische en Natuurwetenschappen en door 6,5% van de respondenten uit de Sociale en Geesteswetenschappen. Een logische vervolgstap is dat de dataset betrokken wordt in de peer review van het artikel. Een hoog percentage van de respondenten acht dit ook belangrijk, maar ziet dit als niet haalbaar. Het peer review systeem is reeds overbelast en deze extra belasting stuit op grote weerstanden.

Een eventuele aanbeveling hierover moet gericht zijn op de wetenschappelijke tijdschriften. Het lijkt erop dat de trend dat steeds meer tijdschriften beschikbaarstelling van de onderliggende datasets bij de artikelen vereisen zich doorzet. De verwachting is dat dit uiteindelijk ook gevolgen heeft voor het peer review proces. Maar gezien de overbelasting van peer review en het gebrek aan ervaring met het opnemen van datasets verwachten we dat het nog geruime tijd zal duren voordat dit gewoonte zal zijn.

5.2.6 Voor individuele onderzoeksinstellingen (Levenswetenschappen en Sociale en Geesteswetenschappen)

De optie van het aanbieden van datamanagementtrainingen ligt met name op het terrein van de individuele onderzoeksinstellingen. Volgens dit onderzoek voorzien redelijke percentages van wetenschappers uit de Levenswetenschappen en Sociale en Geesteswetenschappen dat dergelijke trainingen in een behoefte voorzien. De respondenten uit de Technische en Natuurwetenschappen ontvangen deze optie nogal lauw. Volgens onze analyse komt dit verschil (alhoewel niet statistisch significant) voort uit de verschillende aard van de data: datamanagement in de ‘harde’ wetenschappen betreft vaak het kalibreren van de apparatuur en dergelijke, terwijl datamanagement in de ‘zachtere’ wetenschappen meer focust op de methoden van datacollectie en in geval van ‘human subjects’ op regels van ‘informed consent’ en privacy. Een goed aanbod van trainingen op het gebied van datamanagement lijkt essentieel. Maar waarschijnlijk is dit met name relevant voor wetenschappers uit de Levenswetenschappen en uit de Sociale en Geesteswetenschappen.

Appendix A - Bibliografie en bloemlezing

A1. Bibliografie

Hedendaags wetenschappelijk onderzoek resulteert niet alleen in publicaties, maar in toenemende mate ook in onderzoeksdata: gegevensverzamelingen waarop het onderzoek gebaseerd is, maar die na of naast de publicatie een zelfstandig leven leiden. Voor hergebruik moeten deze verzamelingen, net als publicaties, vindbaar en toegankelijk zijn. Ook kwaliteit speelt bij beide soorten onderzoeksresultaten een belangrijke rol. Waar echter voor wetenschappelijke artikelen dit aspect in de loop der tijd – niet altijd onomstreden – is geoperationaliseerd via peer review en citatie-indexen, staat het voor onderzoeksdata nog in de kinderschoenen. Literatuur over kwaliteit van onderzoeksdata is van recente datum.

Publicaties (gesorteerd in retrochronologische volgorde).

1. 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
   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
2. Open to All? Case studies of openness in research. A joint RIN/NESTA report. September 2010
   http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies
3. Selection of Research Data. Heiko Tjalsma (Ed.). DANS and 3TU Data Centre. July 2010. In press.
4. Data Sharing Policy: version 1.1 (June 2010 update). Biotechnology and Biological Sciences Research Council UK.
   http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf
5. Quality assurance and assessment of scholarly research. RIN report. May 2010.
   www.rin.ac.uk/quality-assurance
6. Rechtspraak en digitale rechtsbronnen: nieuwe kansen, nieuwe plichten. Marc van Opijnen. Rechtstreeks 1/2010.
   http://www.rechtspraak.nl/NR/rdonlyres/6F244371-265F-4348-B7BD-22EB0C892811/0/rechtstreeks20101.pdf
7. Perceived Documentation Quality of Social Science Data. Jinfang Niu. 2009.
   http://deepblue.lib.umich.edu/bitstream/2027.42/63871/1/niujf_1.pdf
8. Guide to Social Science Data Preparation and Archiving. Best Practice Throughout the Data Life Cycle. 4th edition. 2009. Inter-University Consortium for Political and Social Research.
   http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf
9. The Publication of Research Data: Researcher Attitudes and Behaviour. Aaron Griffiths, Research Information Network The International Journal of Digital Curation Issue 1, Volume 4 | 2009.
   http://www.ijdc.net/index.php/ijdc/article/viewFile/101/76
10. Managing and sharing data, a best practice guide for researchers, 2nd edition. UK Data Archive. 18 september 2009.
    http://www.esds.ac.uk/news/publications/managingsharing.pdf
11. e-IRG and ESFRI, Report on Data Management. Datamanagement Task Force. December 2009.
    http://ec.europa.eu/research/infrastructures/pdf/esfri/publications/esfri_e_irg_report_data_management_december_2009_en.pdf
12. Enige artikelen in Nature over beschikbaarheid van onderzoeksdata.
    a. News Feature, 9 September 2009. Data sharing: Empty archives.
       http://www.nature.com/nature/journal/v461/n7261/pdf/461145a.pdf
    b. Opinion, 9 September 2009. Prepublication data sharing.
       http://www.nature.com/nature/journal/v461/n7261/full/461168a.html
    c. Opinion, 9 September 2009. Post-publication sharing of data and tools
       http://www.nature.com/nature/journal/v461/n7261/full/461171a.html
    d. Editorial, 10 September 2009.
       http://www.nature.com/nature/journal/v461/n7261/pdf/461145a.pdf

13. Waardevolle Data & Diensten. Eindrapport SURFshare project. 14 juli 2009. http://www.surffoundation.nl/SiteCollectionDocuments/Eindrapportage_WDenD_v10.pdf
14. Australian National Data Service (ANDS) interim business plan, 2008/9. http://ands.org.au/andsinterimbusinessplan-final.pdf
15. Peer Review: benefits, perceptions and alternatives. PRC Summary Papers 4. 2008. http://www.publishingresearch.net/documents/PRCPeerReviewSummaryReport-final-e-version.pdf
16. To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. RIN report; main report. June 2008. http://eprints.ecs.soton.ac.uk/16742/1/Published_report_-_main_-_final.pdf
17. Stewardship of digital research data: a framework of principles and guidelines. Responsibilities of research institutions and funders, data managers, learned societies and publishers. RIN report. January 2008 http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines
18. The Genographic Project Public Participation Mitochondrial DNA Database. Behar, D.M, Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., Quintana-Murci, L., Tyler-Smith, C., Spencer Wells, R. PLoS Genet 3 (6). 29 June 2007. http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0030104
19. Dealing with Data: Roles, Rights, Responsibilities and Relationships, Consultancy Report. Dr Liz Lyon, UKOLN, University of Bath. 19 June 2007. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf
20. OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. http://www.oecd.org/dataoecd/9/61/38500813.pdf
21. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. International Council for Science. 2004. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

Web sites (bezocht in augustus en september 2010)

22. Acta Crystallographica Section E: Structure Reports Online. http://journals.iucr.org/e/journalhomepage.html
23. Earth System Science Data. The Data Publishing Journal. http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html
24. Appraisal Criteria in Detail. Inter-University Consortium for Political and Social Research. http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/appraisal.jsp
25. Data Seal of Approval. DANS. http://www.datasealofapproval.org/
26. NWO-DANS datacontract. http://www.dans.knaw.nl/sites/default/files/file/NWO-DANS_Datacontract.pdf
27. UK Data Archive. http://store.data-archive.ac.uk/store/
28. World Ocean Database 2009. NOAA Atlas NESDIS 66. ftp://ftp.nodc.noaa.gov/pub/WOD09/DOC/wod09_intro.pdf
29. RCSB Protein Data Bank Deposition Portal. http://deposit.rcsb.org/
30. Data Documentation Initiative http://www.ddialliance.org/
31. European Research Council. http://en.wikipedia.org/wiki/European_Research_Council
32. The String of Pearls Initiative. Netherlands Federation of University Medical Centres. http://www.string-of-pearls.org/

A2. Bevindingen

Generiek
Alle gelezen publicaties onderschrijven het belang van het aspect kwaliteit bij onderzoeksdata. Vrijwel altijd heeft dit betrekking op de kwaliteit van de metadata en de documentatie; soms ook op de inhoudelijke kwaliteit. Een generieke definitie van het begrip kwaliteit is niet aangetroffen. In het algemeen wordt verdere studie aanbevolen; het onderwerp is nog in een pioniersfase.

Specifiek
(De nummers verwijzen naar de bovenstaande referentielijst)

- In het RIN rapport (#16) wordt een eerste stap gezet naar een operationalisering van het aspect kwaliteit door dit te koppelen aan verschillende fasen uit de levenscyclus van onderzoeksdata: de creatiefase van de data, het duurzaam bewaren en toegankelijk maken van de data, het gebruik van de data. In onze studie is deze driedeling als basis voor ons onderzoek aangehouden.
- Het Interim Business Plan 2008/9 van de Australian National Data Services (ANDS) (#14) kondigt aan een annotatieservice op te zetten voor de research data van de Australian Life Atlas als instrument om de kwaliteit van de data te verbeteren.
- Het proefschrift van Jinfang Niu (#7) analyseert uitgebreid het belang van een goede documentatie voor het hergebuik van onderzoeksdata in de sociale wetenschappen.
- DANS heeft in 2008 een Datakeurmerk geintroduceerd voor de sociale en geesteswetenschappen dat internationale erkenning heeft gekregen (#25). Om voor het Keurmerk in aanmerking te komen moeten data-producenten, data-repositories en data-consumenten minimaal aan bepaalde kwaliteitseisen voldoen.
- Toegesneden inhoudelijke kwaliteitscontroles worden aangetroffen bij enige grote verzamelbestanden zoals het Genographic Project (#18), de RCSB Protein Data Bank (#29) en de World Ocean Database 2009 (#28).
- De publicatie van Marc van Opijnen (#6) richt zich op de structurering van (juridische) informatie en slaat een brug tussen het onderwerp van onze studie (kwaliteit van onderzoeksdata) en een centraal thema in het SURFshare programma (verrijkte publicaties), met als conclusie dat niet technologie de belangrijkste uitdaging is, maar professionaliteit.

Onderstaand een bloemlezing van relevante passages uit de rapporten.

A3. Bloemlezing

In de gelezen rapporten werden de hiernavolgende passages over kwaliteit van onderzoeksdata aangetroffen.

1. **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**

   p. 5
   From the shortlist of actions
   “ 3. Develop and use new ways to measure data value, and reward those who contribute it. If we are to encourage broader use, and re-use, of scientific data we need more, better ways to measure its impact and quality. We urge the European Commission to lead the study of how to create meaningful metrics, in collaboration with the ‘power users’ in industry and academia, and in cooperation with international bodies. **”**

   p. 34
   “ Propose reliable metrics to assess the quality and impact of datasets. All agencies should recognise high quality data publication in career advancement. **”**

   p. 36

“ Provide ways for data producers to benefit from publishing their data. **”**

2. Open to All? Case studies of openness in research. A joint RIN/NESTA report. September 2010

p.45
“Imposing a delay on access to data and other materials is thus a feature of the practice of many of the groups. There is also a common realisation across the groups that making materials openly accessible requires effort to ensure that they are properly presented, with adequate metadata, proper calibration, high-quality documentation, and appropriate tools. Only if they are easy to understand and to use can materials be said to be truly open. But all this requires effort and resources, and has to find a place among competing priorities: “So it’s both a good thing and also a kind of trap”.**”**

p. 47
“Researchers are concerned, as both creators and users of ‘research objects’ of many different kinds, that effective quality checks are in place. They also retain a strong belief in peer review. But there is an increasingly widespread view that in its current form it cannot cope with the vast volume and range of materials that are being produced. Hence the growing interest in ‘post-publication’ and machine-readable forms of peer review and tests of usability.**”**

3. Selection of Research Data. Heiko Tjalsma (Ed.). Forthcoming report by DANS and 3TU Data Centre. July 2010

p. 11
“*Application of criteria by repositories*
Looking from the perspective of data repositories, it is not possible to draw general conclusions from the limited number of interviews being held for this exploration. It is however striking to observe that most data repositories at the moment do not seem to have strong selection and explicit criteria or to have these criteria at all. The American data archive ICPSR might be an exception having a clearly formulated appraisal policy. Of course data repositories restrict themselves to the limits of their data acquisition profiles and most data repositories select on quality and completeness of the data and metadata. However: no explicit criteria are mentioned. How is quality determined, is scientific quality meant or technical quality, like the presence of metadata or data formats suitable for archiving?**”**

5. Quality assurance and assessment of scholarly research. RIN report. 28 May 2010.

p. 12 -14
“5.5 Data sharing
[...]
Providing access to the data underlying publications can itself constitute an important quality check, although it may also present a challenge to peer reviewers. In some disciplines, reviewers check data thoroughly, and are capable of unearthing flaws or inconsistencies. In other cases, checking is less thorough, partly because reviewers may not be able to judge the data satisfactorily, partly because datasets may be too large to review in their entirety, and partly because the data may be too complex to be judged in this way. Reviewers may check that the data are present and of the format and type that the work warrants, and leave it at that. Overall the approach is uneven. Nevertheless, some publishers now require authors to deposit data in an authorised repository and to make it available to others; and once findings are formally published, other researchers may well interrogate the underlying data as a quality check both on the data themselves and on the findings as published.

Making large datasets available through online repositories may also involve quality checks and assurance. Many data centres and databanks apply stringent rules to ensure that datasets meet quality standards and are accessible (through the effective use of metadata), and usable (by providing the information and possibly the software tools necessary for others to re-use the data).

Where a dataset has significant scholarly merit but is lacking in other respects, data centres will often work with data creators to ensure that the dataset is discoverable, accessible, and usable. Some repositories also implement checks on the quality of the data themselves. For example, the Genographic Project, a database created to study the genetic signatures of ancient human migrations, has quality checks integrated into the database design[6] ; and researchers wishing to contribute to the RCSB Protein Data Bank are given a chance to check and improve the quality of their data through online tools made available prior to submission.[7]

However, not all databases have such quality assurance mechanisms associated with them, and it may not always be practical to implement them. In such cases, greater responsibility passes to potential users, who must decide whether the data are of high quality and usable within the context of their own specific research projects. Good metadata are essential for both purposes.

Good metadata, and standards for the identification, linking and citation of data are also essential if researchers who create data are to have effective incentives to share them with others. At present, relatively few researchers receive recognition or rewards for sharing their data; and very few submit datasets they have created for formal assessment in the RAE or other assessment exercises.

7. Perceived Documentation Quality of Social Science Data. Jinfang Niu. 2009.

p. 10
"Summary
For the past several decades, more and more data have been shared or at least are required to be shared with the public. It is widely acknowledged that research data sharing will save funding and avoid repeated data collecting effort, facilitate open science and deter scientific fraud. The benefit of data sharing can only be reaped through the secondary use of data. However, there is a lack of research about secondary use of social science data. To enable secondary data use, data and knowledge about data need to be transferred from data producers to secondary users. Documentation is metadata of data. It is an important knowledge transfer channel between data producers and secondary users.
However, even though it is often reported that inadequate documentation is a barrier in secondary data analysis, documentation of social science data has rarely been the focus of existing studies. This study is trying to address the void of studies on secondary users and documentation. It will study user perceived documentation quality. The goal is to identify impacting factors of documentation quality, and find out how perceived documentation quality affects secondary data use. Findings from this study will inform decisions about how to improve documentation quality and facilitate secondary data use.**"**

9. The Publication of Research Data: Researcher Attitudes and Behaviour. Aaron Griffiths, Research Information Network The International Journal of Digital Curation Issue 1, Volume 4 | 2009

p. 54
"Quality Assurance
Research communities are to a large extent self-regulating in respect of data quality assurance. Most researchers interviewed by the authors of the RIN report replied that they generally take other researchers' outputs on trust in terms of data quality and integrity, and there is little evidence of dissatisfaction with this state of affairs.
The scholarly merit of data is assessed by the peer community by comment, re-use, and building upon data outputs. Peer review may involve checking supporting data in a more or less detailed way. In many cases, reviewers may not be able to judge the data satisfactorily, and especially in scientific disciplines the datasets may be too large or complex to review manually or in their entirety. Reviewers may check that the data are present and in the format and of the type that the work warrants, and leave it at that.

Variability in the quality of peers' assessment of the content of the datasets that underpin publications is one of the key reasons why many researchers interviewed by the RIN "do not discount the idea of instituting a formal process for assessing the quality of datasets". However, the report continues, "no one can see it working effectively in practice" (RIN, 2008, p. 49). There are concerns that it would be difficult to find reviewers with sufficient expertise in highly specialised fields to understand the data, let alone appraise them. Reviewing datasets would also have costs and add time to the research process at a point in the project life cycle where researchers want to be writing papers for publication.
But, as the report notes, this lack of grassroots demand does not mean that research funders might not wish to see a more rigorous and consistent quality assurance process for datasets, particularly if they, along with other organisations, are investing heavily in the infrastructure required to support their publication. As Borgman notes, "As data reuse becomes more common, the pressure on reviewers to assess and certify data will only increase" (Borgman, 2007, p. 135).
The report notes that there is more to quality assessment than just the consideration of the scholarly merit of a dataset. If the process of data sharing is to become more effective and useful, much more consideration needs to be given to making datasets accessible (through the effective use of metadata) and usable (by providing the information and possibly software tools necessary for others to re-use the data). Whether the creators of datasets should be encouraged to gain such skills through education, persuasion, grant conditions or other means is an issue for research councils, other funders and the data centres to consider. An alternative approach would be to train and recognise the value of data scientists – whether from a research or information background – whose role would be to work alongside researchers, helping them devise and achieve the goals of effective data management plans. This suggests the need for further consideration by all stakeholders on what approaches to the formal assessment of datasets are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum. **"**

10. Managing and sharing data, a best practice guide for researchers. UK Data Archive. Second edition, 18 september 2009.

p. 10-11
"Data quality control
Quality control of data is an integral part of all research and takes place at various stages, during data collection, data entry or digitisation, and data checking. It is important to assign clear roles and responsibilities for data quality assurance at all stages of research.

During data collection, researchers must ensure that the data recorded reflect the actual facts, responses, observations or events, for example:

- if data are collected with instruments:
  - calibration of instruments is essential to check the precision, bias and/or scale of measurement
  - data are validated by checking for equipment as well as digitisation errors
  - data may be verified by checking the truth of the record with an expert or by taking multiple measurements, observations or samples
- standardised methods and protocols can be used for capturing observations, alongside recording forms with clear instructions
- computer-assisted interview software can be used to standardise interviews, verify response consistency, route and customise questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses

The quality of data collection methods used has a significant bearing on data quality. Documenting in detail how data are collected increases their quality.

When data are digitised, transcribed, entered in a database or spreadsheet or coded, quality is ensured by adhering to standardised and consistent procedures for data entry with clear instructions. This may include setting up validation rules or input masks in data entry software;

using data entry screens; using controlled vocabularies, code lists and choice lists to minimise manual data entry; detailed labelling of variable and record names to avoid confusion; or designing a purpose-built database structure to organise data and data files.

During data checking, data are edited, cleaned, verified, cross-checked and validated. Checking typically involves both automated and manual procedures. This may include: double-checking coding of observations or responses and out-of-range values; checking data completeness; verifying random samples of the digital data against the original data; double entry of data; statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values; or peer review."

11. e-IRG and ESFRI, Report on Data Management. Datamanagement Task Force. December 2009

p. 61 - 68
"*3. QUALITY OF DATA RESOURCES*
3.1. INTRODUCTION
The quality of research is usually measured in terms of scientific output. The quality of scientific output, especially in paper journals, has traditionally been assured through the system of peer review. For electronic publications, new ways of reviewing articles in digital repositories have been sought and implemented since the mid 1990s [1]. Applications for projects are also usually peer reviewed. According to National Institutes of Health (NIH) in the US, the increasing breadth, complexity, and interdisciplinary nature of modern research has necessitated a more formal review of the NIH peer review system. A recent report identifies the most significant challenges and proposes recommendations that would enhance the peer review system [2].

The databases underlying scientific publications are only rarely reviewed, although an increasing number of journals require the submission of such data sets in (publicly accessible) data repositories. Such journals have been labelled as "DAP Journals" [3] or journals with a Data Availability Policy. Data archives, which have been set up since the 1960s, have always used the potential for checking of possible errors in data collections as a motive for their (continued) existence. Such digital data archives are the main advocates of quality assurance for research data. Quality control by data archives is usually achieved by painstaking and labour-intensive checks on the data, carried out by data archive staff. Quality checks carried out include:

- checking the format of the data files
- checking whether a complete code book is available for coded data
- checking the anonymity of personal data; data are de-identified by expunging names, addresses, etc.
- checking for missing values and overall completeness / data integrity
- checking for consistency

Moreover, the description of the data sets by adding metadata (for archiving and retrieval) is often carried out by data archive staff, with some exceptions. Few archives use a self-deposit system, in which the depositors (i.e. the researchers who have produced the data) add the metadata describing the data they submit. In these cases, the data archives still perform (marginal) quality checks on the metadata and data deposited.

Some data archives have plans for introducing a form of review by users (which are normally peers in research data archives) similar to product or service reviews that are common in many web shops. Such review systems have to our knowledge not yet been implemented.

Internationally, there are several initiatives for setting criteria to certify digital repositories, such as TRAC [4], DRAMBORA [5] and the *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* developed by NESTOR[6]. All concentrate on quality of digital archives as organisations and (technical) service providers, not necessarily on the quality of their contents.

The content of research data repositories is always dependent upon what is submitted by the depositors, the researchers, who are primarily responsible for the quality of their work.

The OECD has published a set of thirteen principles and guidelines for access to research data from public funding [7], several of which are linked to data quality, and one is explicitly labelled "quality". The guidelines concern: *A. Openness, B. Flexibility, C. Transparency, D. Legal conformity,E. Protection of intellectual property, F. Formal responsibility, G. Professionalism, H. Interoperability, I. Quality, J. Security, K. Efficiency, L. Accountability, M. Sustainability.* In the guidelines improved access itself is seen as benefiting the advancement of research, boosting its quality (p. 8). The guideline on quality per se is reproduced in Table 2.1.
Note: italics and bold added by the author.

In a recent report by the Strategic Committee on Information and Data (SCID) of the International Council for Science (ICSU), advice is given on the future organisation and direction of activities in relation to scientific data and information. Although the term "data quality" appears 17 times in the text, it is not explicitly stated what "quality" means or how it can be guaranteed. The report makes very clear, however, that data organisations are to play an important role in ensuring the quality of and access to research data, as shown in Table 2.2.
Note: bold and italics added by the author.

3.2. SHARING DATA AND QUALITY ASSURANCE

In a recent report commissioned by the Research Information Network (RIN), one chapter is devoted to "Quality assurance in the data creation process" [8]. With regard to creating, publishing and sharing datasets the RIN report identifies three key purposes:

- datasets must meet the purpose of fulfilling the goals of the data creators' original work;
- datasets must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers;
- datasets should be discoverable, accessible and re-usable by others.

The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organisations, should pay particular attention to ensuring compliance with explicit quality standards. Where such standards do not yet exist, institutions and research associations should engage with their research community on their development. Although all areas of research can benefit from improved data quality, some require much more stringent standards than others.
For this reason alone, **universal data quality standards are not practical**. Standards should be developed in consultation with researchers to ensure that the **level of quality and precision meets the needs of the various disciplines**. More specifically:

- Data access arrangements should describe good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable **quality control by peer review and other means of safeguarding quality** and authenticity.
- The **origin of sources should be documented** and specified in a verifiable way. Such documentation should be readily available to all who intend to use the data and incorporated into the metadata accompanying the data sets. Developing such metadata is important for enabling scientists to understand the exact implications of the data sets.
- Whenever possible, access to data sets should be **linked with access to the original research materials**, and copied data sets should be linked with originals, as this facilitates validation of the data and identification of errors within data sets.
- Research institutions and professional associations should develop appropriate practices with respect to the **citations of data** and the recording of citations in indexes, as these are mportant indicators of data quality.

Table 2.1: The quality guideline from the OECD principles and guidelines (2007)

Following an earlier priority area assessment exercise in this area, ICSU's declared strategic goal is: to facilitate a new, coordinated global approach to scientific data and information that ensures **equitable**

access to quality data and information for research, education and informed decision-making [30].
The ICSU plans to create a new World Data System (as an ICSU Interdisciplinary Body), incorporating the existing World Data Centers (WDC) and the Federation of Astronomical and Geophysical Data analysis Services (FAGS) as well as other state-of-the-art data centres and services. This new structure or system must be designed to ensure the longterm stewardship and provision of **quality-assessed data** and data services to the international science community and other stakeholders.
ICSU has an important responsibility on behalf of the global scientific community for promoting the optimal stewardship and policy development for scientific data and information.
Three major trends in data and information management are dramatically changing science. The first is the major step change in the sheer volume and diversity of data suitable for science. Many fields from geo-demographics to particle physics are witnessing dramatic increases in data and information volumes. The second is the availability of new information and communication technologies, such as Grid computing or Sensor Web, which means that very ambitious modelling and data processing are within the scope of an increasing number of scientists. The third is the increasing need for **scientific datasets to be properly identified, qualityassured, tracked and accredited** (for example, through assignment of digital object
identifiers or DOIs). This requires professional data management and, in some areas, may involve review and publication of datasets. Publication and accreditation can also act as an important incentive for primary data producers to make their data available.
There is a need for global federations of professional state-of-the-art data management institutions, working together and exchanging practices. Such federations can **provide quality assurance** and promote data publishing, providing the backbone for the development of a global virtual library for scientific data.

Table 2.2: Excerpts from the Final Report to the ICSU Committee on Scientific Planning and Review [30]

Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated. The scientific or scholarly value of datasets that are not accessible for re-use by others can obviously not be assessed by independent peers.

The RIN report distinguishes data sets created by machines (such as telescopes, spectrometers, gene sequencers) from those created in other ways (such as social surveys, databases created by manual input, source editions of texts, etc.). This distinction roughly (though by no means completely) coincides with the distinction between the sciences and the humanities. Machines that create data often have inbuilt data validation mechanisms. Manual checking is usually added, and, in those disciplines where data are collected by other means, manual verification may involve very detailed work.

Although there is no information on how many data sets are checked by others than the researcher herself, it is in many cases taken for granted that when a paper is accepted for publication after peer review the underlying data will pass the quality standard as well. Peer review may involve checks of the supporting data. In some disciplines, reviewers do checks on data. In other cases, checking is superficial or absent, because the data are too complex or voluminous to be judged satisfactorily. Most researchers take other researchers' outputs on trust in terms of data quality and integrity. Moreover, there are no apparent signs of dissatisfaction with this state of affairs.

Summary
17. Most researchers believe that data creators are best-placed to judge the quality of their own datasets,

and they generally take other researchers' outputs on trust in terms of data quality and integrity.
18. There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability.
19. Data centres apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata. But many researchers lack the skills lack the skills to meet those standards without substantial help from specialists.
Recommendation
9. Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum.

Table 2.3: Summary and recommendation with respect to data quality in the RIN report

The accessibility of data for re-use renders checking possible at a later time. Experimental, machine-produced data can in principle be re-created if the whole experiment is done again. In practice, such validity-checking procedures are rarely carried out. Nevertheless, new experiments with better measuring equipment, or the secondary analysis of survey data or text corpora, re-appraisal of digital scholarly editions and so on, may result in the discovery of earlier flaws, and in extreme cases in the exposure (and shaming) of mistakes or even fraud.

Although it is useful to distinguish between the scientific/scholarly content of data and the technical merits that facilitate re-use, it is questionable whether the two can be separately reviewed, as recommended in a report commissioned by the Arts and Humanities Research Council [9]. British researchers generally support the idea of instituting a formal process for assessing the quality of datasets, although they have concerns as to whether it will work effectively in practice. Among these are the difficulty to find reviewers who are willing and who have the expertise to understand and appraise the data [10]. Another concern involves the costs (in terms of money and time) of a formalised data review process.

It is not likely that the pressure to improve the quality assurance process for datasets will come from researchers, although they generally seem to favour a more thorough assessment. Research funders, who are investing heavily in the data infrastructure, are in a better position to take the initiative to introduce a formal assessment process. This would also imply that data creation itself should be rewarded with scholarly merit and scientific credits [11]. Funding agencies do not always require a formal data plan, although for certain subsidies in some countries, such a requirement exists. For instance, researchers receiving a grant from the Dutch funding organisation NWO for a data creation project are required to sign a "data contract" with DANS [12], in which they are obliged to comply with the "Data Seal of Approval".

Whilst datasets that are deposited at data centres must conform to certain quality standards, there is no such imperative yet for researchers who look after their own datasets. According to the RIN study, some researchers believe this to be outside the boundaries of their research function, while others lack the skills (and/or time) to publish their data such that it can be discovered, accessed and re-used by the scholarly community.

3.3. ASSESSING THE QUALITY OF RESEARCH DATA

Recently a method for assessing the quality of research data, called the "Data Seal of Approval" (DSA) has been developed by DANS in The Netherlands. The ambition of the DSA is to ensure that research data of a guaranteed quality can be found, recognised and used in a reliable way [13]. An international board consisting of data archivists and researchers from various disciplines and countries has taken on the responsibility for the guarding of the DSA and its application in practice. An assessment procedure has been developed and tested and is currently undergoing wider implementation and roll out. The quality guidelines formulated in the DSA are of interest to researchers and institutions that create digital research files, to organisations that manage, archive, curate and disseminate research files, and to users of research data.

This DSA contains a total of 16 guidelines for the application and verification of quality aspects with regard to creation, storage and (re)use of digital research data (see Appendix A). Although originally developed for application to the social sciences and humanities, the draft text of the DSA has also met with positive responses from the natural sciences. Slight modifications have been made to make the DSA more generically applicable. The international board will take additional requests for adaptations into consideration if these are required for certain research fields. However, the DSA guidelines seem well capable of serving the quality requirements of many disciplines, and even of areas outside of science (e.g. cultural heritage, public archives).

The criteria for assigning the seal of approval to data are in accordance with, and fit in with, national and international guidelines for digital data archiving such as the *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* [15] as developed by NESTOR, the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) [14] published by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), and the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* of the Research Library Group (RLG). The *Foundations of Modern Language Resource Archives* of the Max Planck institution [14] and the *Principles and guidelines for the stewardship of digital research data* published by the Research Information Network [15] have also been taken into consideration. The DSA guidelines mcan be seen as a minimum set distilled from the above proposals.

Digital research data must meet five quality criteria:
- The research data can be found on the Internet.
- The research data are accessible, while taking into account ruling legislation with regard to personal information and intellectual property of the data.
- The research data are available in a usable data format.
- The research data are reliable.
- The research data can be referred to.

The DSA guidelines focus on three operators: the *data producer*, the *data repository* and the *data consumer*:
- The *data producer* is responsible for the quality of the digital research data.
- The *data repository* is responsible for the quality of storage and availability of the data: data management.
- The *data consumer* is responsible for the quality of use of the digital research data.

Data producers

The quality of digital research data is determined by:
- Their intrinsic scientific quality;
- The format in which the research data and supporting information are stored;
- The documentation (metadata or contextual information) regarding the research data.

Scientific quality criteria indicate to what degree the research data are of interest to the business of science. The assessment by experts, colleagues in the field, is the main decisive factor for the scientific quality of research data. Three questions must be answered to be able to provide an assessment.

1. Are the research data based on original work performed by the data producer (the researcher or institution that makes the research available) and does the data producer have a solid reputation? This question can be answered by proviing information regarding the researcher and/or research group and by providing references to publications pertaining to these particular research data.
2. Was data creation carried out in accordance with prevailing criteria in the research discipline? Answering this question requires information on the methods and research techniques used, including those for data collection, digitisation or other means of data creation.
3. Are the research data useful for certain types of research and suitable for reuse? The answer requires information regarding the data format, content and structure. The data

producer therefore provides sufficient information to enable fellow scientists to assess the research data.

Data format: The elements that form a digital research object are organised according to the rules for a particular data format. Various data formats exist for digital objects. With all formats, there is a risk of obsolescence. This creates the possibility that the data object may become unusable. For storage of data objects, so-called preferred formats are therefore used. Preferred formats are formats designated by a data repository for which it guarantees that they can be converted into data formats that will remain readable and usable. Usually, the preferred formats are de facto standards employed by particular research communities.

Documentation: The data producer provides the research data with contextual information (metadata). There is a distinction between descriptive, structural and administrative metadata.
These must be provided in accordance with the applicable guidelines of the data repository.

- Descriptive metadata are data necessary to be able to find research data and that add transparency to their meaning (definition and value) and importance. Examples of descriptive metadata are the data elements of the Dublin Core Element Set [16], with fields such as creator, type, and date.
- Structural metadata indicate how different components of a set of associated data relate to one another. These metadata are needed to be able to process the research data. When data are coded, the codebook will be a component of the structural metadata.
- Administrative metadata are required to enable permanent access to the research data. This concerns the description of intellectual property, conditions for use and access, and so-called preservation metadata needed for durable archiving of the research data.

Data repositories
The data repository is responsible for access and preservation of digital research data in the long term. Two factors determine the quality of the data repository:

- The quality of the organisational framework in which the data repository is incorporated (organisation and processes);
- The quality of the technical infrastructure of the data repository.

Organisation and processes: Organisations that play a role in digital archiving and are establishing a Trusted Digital Repository (TDR) minimally possess a sound financial, organisational and legal basis for the long term. Depending on the task assigned to an organisation, a TDR may distinguish itself qualitatively by carrying out research and by cooperating with other organisations in the realm of data archiving and data infrastructure. The outcomes of such research are shared, both nationally and internationally. In addition, these organisations will also share physical infrastructures, software and other knowledge among each other, where possible.

Technical Infrastructure: The technical infrastructure constitutes the foundation of a Trusted Digital Repository. The OAIS reference model, an ISO standard, is the *de facto* standard for using digital archiving terminology and defining the functions that a data repository fulfils [17].

Data consumers
The quality of the use of research data is determined by the degree to which the data can be used without limitation for scientific research by the various target groups, while complying with certain rules of conduct. The open and free use of research data takes place within the legal frameworks and the policy guidelines as determined by the relevant (national) authorities.

The “OECD Principles and Guidelines for Access to Research Data from Public Funding” [18] provide policy guidelines regarding access to research data, which are accepted by the governments of the OECD countries. The principles of “Open Access” are moreover described in the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [19],

which has been signed by over 250 scientific organisations in more than 30 countries (end of 2008).

In the Netherlands, the *Code of Conduct for Research* [20] is of importance for the use of research data. This Code of Conduct utilises five core concepts that are essential for the quality of scientific research: due care, reliability, verifiability, objectivity, and independence. An additional *Code of Conduct for the use of personal information in scientific research* [21] focuses on responsible handling of privacy-sensitive data and describes the legal frameworks for scientific work with such data. These codes of conduct comply with the Dutch Personal Data Protection
Act (WBP). This law provides the frameworks within which personal information may be used in The Netherlands. The Dutch Data Protection Board (CBP) monitors compliance with legislation that regulates the use of personal information.**"**

13. Waardevolle Data & Diensten. Eindrapport SURFshare project. 14 juli 2009.

p. 14
" Kwaliteit
Het datacentrum moet een hoge kwaliteit van de door het centrum beschikbaar gestelde onderzoeksdata nastreven. De ervaring leert dat de kwaliteit van de onderzoeksdata toeneemt door de eisen die het datacentrum stelt aan de consistentie. De verantwoordelijkheid voor de kwaliteit ligt echter bij de onderzoeker. Om de kwaliteit te borgen moeten de gegevens altijd gescreend zijn voordat ze worden opgenomen in het datacentrum. De controle op de onderzoeksdata neemt veel tijd in beslag en het is niet altijd mogelijk de data automatisch te screenen. De ervaring leert echter dat deze kwaliteitscontrole wel waardevol is.

Doordat de onderzoeksdata openbaar zijn, ontstaat tevens een soort sociale druk om te zorgen voor de juiste kwaliteit. Een datacentrum moet o.a. vanwege zichtbaarheid van de datasets streven naar het citeerbaar maken daarvan. Indien referenties naar de dataset mogelijk zijn, zal dit enerzijds een impuls kunnen geven aan kwaliteitszorg door dataproducenten maar anderzijds
de drempel voor het deponeren van data misschien verhogen. De meeste dataproducenten die tijdens dit onderzoek aan het woord gekomen zijn zagen hierin echter geen probleem.
Na afronding van het onderzoek kunnen fouten in de meetdata ontdekt worden door andere onderzoekers. Het moet mogelijk zijn om de geïdentificeerde fouten kenbaar te maken en te verbeteren, zonder het brondocument aan te passen.
Beloning van de onderzoekers die hun data beschikbaar maken zou uiteraard een enorme impuls geven aan het vormen van datacollecties.
Een bepaald minimum kwaliteitsniveau van de onderzoeksdata lijkt een voorwaarde voor hergebruik en hoe hoger de kwaliteit hoe vaker de gegevens zullen worden gebruikt ten opzichte van 'concurrerende' (vergelijkbare) datasets.

Aanbeveling

- Het datacentrum moet streven naar een zo hoog mogelijke kwaliteit van de opgenomen datasets (zie ook 'Cultuur').
- De onderzoeker is verantwoordelijk voor de kwaliteit van de datasets.
- De in het datacentrum opgeslagen datasets moeten citeerbaar zijn.
- Fouten moeten kunnen worden aangepast door (collega) onderzoekers zonder het bronbestand te veranderen (bijv. door middel van versiebeheer)."

p. 26
" 3.3. Expertbijeenkomst; toetsing van de bevindingen
Om de bevindingen uit de casestudies te toetsen organiseerde de projectgroep een bijeenkomst met onderzoekers van de drie technische universiteiten (TU Delft, TU Eindhoven en Universiteit Twente). [...] Deze paragraaf beschrijft de aandachtspunten voor het datacentrum die tijdens de bijeenkomst naar voren kwamen.
De deelnemers benoemden zes thema's waar het datacentrum zijn aandacht op moet richten voor succesvol gebruik:

1. Kwaliteit van de data: De kwaliteit van de onderzoeksdata is voor de deelnemers van groot belang. Goede beschrijving van het onderzoek zorgt ervoor dat het herhaalbaar is, terwijl een consistente werkwijze bij databewerking binnen collecties bepalend is voor de mogelijkheden tot hergebruik. Beide aspecten dragen bij aan de valorisatie van kennis.**"**

14. AUSTRALIAN NATIONAL DATA SERVICE (ANDS) INTERIM BUSINESS PLAN, 2008/9.

p. 6
" ANDS will initially focus on content recruitment into stores and federation across stores so as to achieve a wide coverage of data quickly at an agreed level of quality; in later years the emphasis will shift towards quality improvement.**"**

p. 19
" General content recruitment
[...]
work with DCC, NSF and SURFnet to collaboratively develop tools to help improve quantity and quality of repository content **"**

p. 45
" 11.4 Data Integration and Annotation Services in Biodiversity (DIAS-B)
[...]
The quality and consistency of the ALA (Atlas of Living Australia, L.W.) data is crucial for its use. There is a need for an authenticated annotation service that will allow users or automated data analysis tools to provide information to users and feedback to data providers by annotating data records and resource metadata with comments on data quality and suggested corrections.

Description of the proposed service solution and how it meets that need
Data quality services:
- Annotation service allowing human and machine users to store and retrieve annotations relating to any data record within the ALA to record possible errors.
- Reporting service that alerts data provider/owners of possible quality issue. **"**

p. 46
" Improvement of data quality via use of annotation services, measured by:
- Number and range of annotations in annotation database
- Number of responses from data providers
- Direct use of services (UI and web services) for providing annotations (other than through the ALA portal UI and ALA data validation tools services)
- Direct use of services (UI and web services) for accessing annotations
- Number of data records for which annotations have led to corrections in source data.

Improved level of user experience, measured by:
[...]
- An online survey tool to allow users to document their experience in using the ALA infrastructure. This survey tool will be continuously available as a data capture method. This survey will explicitly determine success in using data quality annotation.

[...]

There are two key benefits being sought. [...] Secondly, we wish to understand how well an annotation service supports improvement of data quality – it will be measured using the survey described above. **"**

15. Peer Review: benefits, perceptions and alternatives. PRC Summary Papers 4. 2008.

p. 21
“ *Reviewing authors' data*
As science utilises more automated experimental equipment and otherwise moves towards a more data-centric “e-science” model, the amount of data that supports (and could potentially be linked to) the average scientific paper increases. The question arises as to whether this data should itself be subject to peer review. There are clearly a number of practical issues: do reviewers have the time to do this? Is the data sufficiently standardised, and do the software tools exist to handle it? Are authors even prepared to share their data with reviewers?
A majority of reviewers (63%) and editors (68%) said that it was desirable in principle to review authors’ data. Perhaps surprisingly, a majority of reviewers (albeit a small one, 51%) said that they would be prepared to review authors’ data themselves, compared to only 19% who disagreed. This was despite 40% of reviewers (and 45% of editors) saying that it was unrealistic to expect peer reviewers to review authors’ data. ”

16. To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). Main Report. June 2008.

p. 9
“ Key findings.
Quality assurance
17. Most researchers believe that data creators are best-placed to judge the quality of their own
datasets, and they generally take other researchers’ outputs on trust in terms of data quality and integrity.
18. There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability.
19. Data centres apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata. But many researchers lack the skills to meet those standards without substantial help from specialists. ”

p. 48-50
“ 5. Quality assurance
5.1 Quality assurance in the data creation process
The term “quality” is conventionally associated with the notion of being “fit for purpose”. With regard to creating, publishing and sharing datasets we identified three key purposes: first, the datasets must meet the purpose of fulfilling the goals of’ the data creators’ original work; second, they must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers; third, they should ideally be discoverable, accessible and re-usable by others. Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated.

Researchers are interested in producing the best data they can in order to answer the research questions they are posing, but also to provide a sound basis for producing papers that pass the scrutiny of their peers and are published in reputable journals – this being one of the key drivers of career progression. Most researchers also believe that data creators are best placed to judge the worth of their own datasets and that, on the whole, these judgments fairly reflect their scholarly value.

The requirement to uphold reasonable quality assurance standards is partly met in the sciences by machine efficacy: most machines that create data (such as telescopes, spectrometers, gene sequencers) have inbuilt data checking and verification steps. Manual checking is usually added, and in those disciplines where data are collected by other means manual verification may involve very detailed work. There is pride in turning out datasets of good quality and

shame in exposure as a creator of flawed or incorrect data. Research communities are thus to a large extent self-regulating in respect of data quality assurance. As a result, most researchers reported to us that they generally take other researchers' outputs on trust in terms of data quality and integrity, and we received no reports of dissatisfaction with this state of affairs.

5.2 Data management planning

In all disciplines, larger projects that are receiving or expecting to receive substantial levels of funding engage in a data planning process. A formal data plan is included in grant applications – written by the data manager if the team has one – and this covers the kinds of data that will be created and how; how they will be manipulated; where they will be stored; and how they will be made available for sharing.

Smaller projects typically do not have this degree of formality or concern about data management. Researchers will acknowledge that their funder has a policy or guidelines on data management, and if this includes a requirement to produce a data plan then they will write something in their grant application, usually at the last instance and quite often with little care. There is, therefore, a big difference between the professional and detailed approach taken to data management by some researchers and the cursory approach taken by others. In summary, data management planning is at present highly variable in quality.

5.3 Quality assessment of datasets

At present, the scholarly merit of data is assessed by the peer community by comment, re-use, and building upon data outputs. This is done at two stages: first, when an article or monograph is peer-reviewed by other researchers in the field prior to acceptance of the work for publication; and second, when the community accesses and uses the published work.

Peer review may involve checking supporting data in a more or less detailed way. In some disciplines, reviewers check data extremely thoroughly and are capable of unearthing flaws or inconsistencies at this point. In other cases, checking is less than thorough, partly because reviewers may not be able to judge the data satisfactorily, partly because datasets may be too large to review in their entirety, and partly because the data may be too complex to be judged in this way. Reviewers may check that the data are present and in the format and of the type that the work warrants, and leave it at that. Overall the approach is uneven. There is a concern also that even if peers have the skills to review the scholarly content, they may not be able to judge the technical aspects of a dataset that facilitate usability.

It might be possible to have a two-stage review process focusing on content and technical merit separately, though many argue that in a digital environment the two are interdependent. A eport commissioned by the Arts and Humanities Research Council[17] and published in 2006 made a number of recommendations which included consideration of a two-stage process at grant application stage, focusing separately on scholarly content and technical issues, along with an open post-completion review where paid reviewers' comments would be attributable and data creators have a right of reply.

Variability in and concern about the quality of peers' assessment of the content of the datasets that underpin publications is one of the key reasons why many researchers to whom we spoke do not discount the idea of instituting a formal process for assessing the quality of datasets. Some researchers are mildly enthusiastic but – and it is a big but – no-one can see it working effectively in practice. Several concerns were expressed:

- that it would be difficult to find reviewers with sufficient expertise in highly-specialised fields to understand the data, let alone appraise it;
- that the pool of researchers willing to take the time to review journal articles is diminishing, and that hard-pressed reviewers would be even more unlikely to want to take on further work in assessing datasets[18];
- the costs of the review process, who would pay, and whether the money might not be better spent on research; and
- that having to have a dataset reviewed would add time to the research process at a point in the project life cycle where researchers want to be writing papers for publication.

In summary, there is some sympathy with the concept of expert assessments of the quality of datasets, but researchers don't see how it might work in practice and, given that they are not unhappy with the present situation, there is no grass-roots pressure to introduce a formal assessment process. That is not to say that, in time, research funders themselves might wish to see a more rigorous and consistent quality assurance process for datasets, particularly if they, along with other organisations, are investing heavily in the infrastructure required to support their publication.
There is, however, more to quality assessment than just the consideration of the scholarly merit of a dataset. If the process of data sharing is to become more effective and useful, much more consideration needs to be given to making datasets accessible (through the effective use of metadata) and usable (by providing the information and possibly software tools necessary for others to re-use the data without the help of the data creator). Data centres and databanks come into their own in this regard, applying stringent rules and checks that ensure that datasets deposited meet quality standards, both of the structure and format of data themselves and of the metadata. Where a dataset has significant scholarly merit but is lacking other respects, data centres will normally work with data creators to ensure that the dataset is discoverable, accessible, and usable.

Whilst datasets that are accepted by data centres must conform to such standards, there is no such imperative for researchers who look after their own datasets. They may believe this to be outside the boundaries of their research function but, perhaps more importantly, our study indicates that many researchers do not have the skills to publish their data such that it can be discovered, accessed and re-used by the scholarly community at large and beyond. Whether the creators of datasets should be encouraged to gain such skills through education, persuasion, grant conditions or other means is an issue for Research Councils, other funders and the data centres to consider. An alternative approach would be to train and recognise the value of data scientists – whether from a research or information background – whose role would be to work alongside researchers, helping them devise and achieve the goals of effective data management plans.

Recommendation
Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum. "

Peer review and evaluation of digital resources for the arts and humanities, Arts and Humanities Research Council, September 2006

http://www.britac.ac.uk/reports/peer-review/index.html
18 This echoes concerns reported in a new study of researchers' attitudes to peer review where 40% of reviewers and 45% of journal editors said it was unrealistic to expect peer reviewers to review authors' data. Ware, M (2008), Peer Review: benefits, perceptions and alternatives. http://www.publishingresearch.net/documents/PRCPeerReviewSummaryReport-final-e-version.pdf "

19. Dealing with Data: Roles, Rights, Responsibilities and Relationships, Consultancy Report. Dr Liz Lyon, UKOLN, University of Bath. 19th June 2007

p.6
" REC 19. All relevant stakeholders should commission a study to evaluate the re-purposing of data-sets, to identify the significant properties which facilitate re-use, and to develop and promote good practice guidelines and effective quality assurance mechanisms. "

p. 50

" One further area of uncertainty, in terms of having adequate evidence to inform service development, is around the degree of re-use and re-purposing of data sets. How much data are used again in recombination or re-analysis? What features of a dataset and its associated metadata facilitate re-use? Simple discovery of its existence? A comprehensive set of metadata describing its properties? Some sort of quality rating or peer-review imprimatur? There is a perception that at present, most data are not re-purposed (with notable exceptions such as astronomical survey data which is routinely mined to extract further information). Clearly a researcher should acknowledge the source of their data (if they are not the creator), provide provenance information to assure its quality, and if appropriate, add annotations and tags to add value to the data holdings. Whilst in certain domains such as genomics, these processes are well-established, in other disciplines good practice is less commonplace."

20. OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. June 2008.

p. 19

" *Quality*

The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organisations should pay particular attention to ensuring compliance with explicit quality standards. Where such standards do not yet exist, institutions and research associations should engage with their research community on their development. Although all areas of research can benefit from improved data quality, some require much more stringent standards than others. For this reason alone, universal data quality standards are not practical. Standards should be developed in consultation with researchers to ensure that the level of quality and precision meets the needs of the various disciplines.

More specifically,

- Data access arrangements should describe good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding quality and authenticity.
- The origin of sources should be documented and specified in a verifiable way. Such documentation should be readily available to all who intend to use the data and incorporated into the metadata accompanying the data sets. Developing such metadata is important for enabling scientists to understand the exact implications of the data sets.
- Whenever possible, access to data sets should be linked with access to the original research materials, and copied data sets should be linked with originals, as this facilitates validation of the data and identification of errors within data sets.
- Research institutions and professional associations should develop appropriate practices with respect to the citations of data and the recording of citations in indexes, as these are important indicators of data quality. "

24. Appraisal Criteria in Detail. Inter-University Consortium for Political and Social Research. Web site visited 21 July 2010

"Quality

- ICPSR strongly prefers data collections that have comprehensive technical documentation providing ample information on sampling procedures, weighting, recoding rules, skip patterns, constructed variables, and data collection procedures.
- ICPSR prefers data in the most complete and original form, with the exception of data extracts specifically intended for instructional purposes.
- Lower quality data will be considered for inclusion if the data have unique historical value."

Appendix B - Overzicht geïnterviewden

Er zijn in het totaal 16 interviews gehouden met 17 respondenten, waarvan drie persoonlijke interviews en de overige telefonisch. Elf interviews vonden met respondenten plaats die direct betrokken zijn bij data-archieven, datacentra of bij onderzoeksfaciliteiten die datasets produceren. Twee interviews betroffen respondenten van uitgeverszijde en drie interviews onderzoekers. De interviews vonden plaats in de periode van mei tot en met augustus 2010. Van alle interviews werd een verslag gemaakt en dit werd toegezonden aan de respondenten voor eventuele aanvullingen of correcties.

Geïnterviewden

- Prof. Dr. David Carlson, chief editor, Earth System Science Data (ESSD).
- Prof. Dr. Myron Guttman, voormalig directeur, ICPSR (thans NSF).
- Drs. Eefke Smit, director for standards and technology, STM.
- Dr. Hanno Holties, system engineer, LOFAR.
- Dr. Harm Slijper, senior onderzoeker Erasmus MC, onderzoekscoordinator Xpert Clinic.
- Dr. IJsbrand Jan Aalbersberg, vice president content innovation, Elsevier Science.
- Drs. Ing. Jacquelijn Ringersma, archive manager, Max Planck Instituut Nijmegen.
- Prof. Dr. Kees Mandemakers, directeur, Stichting Historische Steekproef Nederlandse bevolking (IISG).
- Prof.Dr. Kevin Schurer, directeur, UKDA.
- Prof. Dr. Lex Bouter, rector magnificus, Vrije Universiteit.
- Prof. Dr. Marcel Das, directeur, CenterData (LISS panel).
- Dr. Peter Doorn, directeur, DANS.
- Dr. Ross Wilkinson, executive director, and Dr. Andrew Treleor, director of technology, Australian National Data Service (ANDS).
- Dr. Taco de Bruin, voorzitter, Nationale Oceanografische Data Commissie/NIOZ.
- Ir. Wim Som de Cerff, projectleider KNMI Datacentrum, KNMI.

Appendix C - Aanvullende gegevens enquête

C1. Methoden

Op basis van de interviews werd een conceptvragenlijst opgesteld. Deze vragenlijst werd besproken met Wilma Mossink en Gerard van Westrienen van SURFfoundation en met Rene van Horik en Peter Doorn van DANS. Voor de verzending van de uitnodigingen tot deelname aan de enquête stelde DANS twee adressenbestanden beschikbaar:

- een adressenbestand van hoogleraren en UHD's - representatief voor de Nederlandse universitaire gemeenschap.
- een adressenbestand van geregistreerden in het EASY-systeem van DANS: voor het overgrote deel zijn dit hergebruikers van datasets (en sommige gevallen dataproducenten die hun dataset via DANS ter beschikking hebben gesteld). In het rapport zal deze dataset als DANS gebruikers wordt aangemerkt.

Voor de online enquête werd gebruikgemaakt van Pleiadesurvey (zie verder: http://www.pleiade.nl/ict.php?lng_id=NL)
De respondenten werden uitgenodigd tot deelname door middel van een e-mail met een gecodeerde link naar de vragenlijst. De e-mail bevat eveneens een link om aan te geven dat de respondent van deelname afzag. De eerste e-mail werd verzonden op 18 augustus 2010, een herinnering werd verzonden aan hen die de gecodeerde links nog niet gebruikt hadden op 1 september. De enquête werd afgesloten op 13 september.

De gegevens over de respons zijn in de tabel hieronder weergegeven: van de 2811 uitgenodigde hoogleraren en UHD's heeft 13,9% de vragenlijst ingevuld en 14,1% van deelname afgezien. Van de 3280 uitgenodigde DANS gebruikers heeft 8,8% de vragenlijst ingevuld en 16,7 % van deelname afgezien.

Hoogleraren en UHD's		
Totaal aantal e-mailadressen	2869	
onbestelbaar retour	58	
out of office replies eerste mailing	410	
out of office replies herinnering	103	
netto verzonden	2811	
aantal declined	396	(14,1%)
aantal ingevulde vragenlijsten	392	(13,9%)
DANS gebruikers		
Totaal aantal e-mailadressen	3394	
onbestelbaar retour	123	
out of office replies eerste mailing	331	
out of office replies herinnering	114	
netto verzonden	3280	
aantal declined	548	(16,7%)
aantal ingevulde vragenlijsten	287	(8,8%)

In het hoofdrapport zijn uitsluitend de resultaten van de enquête onder de representatieve steekproef van hoogleraren en UHD's gepresenteerd. In paragraaf C3 worden de resultaten van de enquête onder DANS gebruikers gepresenteerd.

C2. Resultaten crossanalyses wel/ niet hergebruiker; wel/niet dataproducent en wel/niet eigen dataset wel eens ter beschikking gesteld

Hieronder de tabellen waarin de resultaten van deze crossanalyses worden gepresenteerd.

Hergebruikers versus niet-hergebruikers	Hergebruikers (n=194)	Niet-hergebruiKers (n=198)	Totaal percentage (alle respondenten; n=392)	Statistische significantie (Chi-kwadraat-toets)
	% (enigszins) mee eens			
1. In mijn vakgebied vragen veel tijdschriften om zo'n replicatie-dataset ter beschikking te stellen.	13,9	9,6	11,7	
2. In mijn vakgebied heeft de peer reviewer normaliter toegang tot de achterliggende dataset bij een publicatie en wordt de dataset meegenomen in de beoordeling van de publicatie.	17,5	10,6	14,0	
3. Het tegelijkertijd beoordelen van de publicatie en de achterliggende dataset is voor een peer reviewer in mijn vakgebied haalbaar.	35,1	29,8	32,4	p=4%
4. Ik vind het belangrijk dat de achterliggende dataset samen met de publicatie wordt beoordeeld in het peer review proces.	56,2	35,4	45,7	p=0%
B. Datapublicaties: peer-reviewed publicaties over datasets				
5. Voor mijn vakgebied zou een dergelijk tijdschrift met data publicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets.	60,8	40,9	50,8	p=0%
C. Commentaren over kwaliteit door hergebruikers van datasets				
6. Voor latere hergebruikers van datasets zouden deze commentaren over kwaliteit waardevol kunnen zijn.	89,7	73,2	81,4	p=0%
7. Als hergebruiker zou ik zeker de moeite nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset die ik heb hergebruikt.	79,4	65,2	72,2	p=1%
8. Als producent van een dataset zou ik dergelijke commentaren over kwaliteit van anderen verwelkomen.	83,5	71,2	77,3	p=1%
D. Citeren van datasets				
9. Als hergebruiker zou ik zeker de dataset citeren met deze mogelijkheid.	84,5	67,2	75,8	p=0%
10. Als wetenschapper zou ik zo'n citatie-mogelijkheid van 'mijn' dataset verwelkomen.	82,5	62,6	72,4	p=0%
E. Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium				
11. Trainingen op het gebied van datamanagement zouden op mijn vakgebied voorzien in een behoefte van veel vakgenoten.	58,2	47,5	52,8	p=1%
12. Data audits, waarin de totstandkoming en het beheer van datasets onder de loep worden genomen, zouden in mijn vakgebied een impuls aan verbeteringen in datamanagement kunnen geven.	62,9	46,5	54,6	p=0%
F. Uw visie	% aangevinkt	% aangevinkt		% aangevinkt
13. De meeste stimulans voor de kwaliteitsbeoordeling van datasets in mijn vakgebied zal waarschijnlijk uitgaan van:				
• peer review van de dataset als onderdeel van peer review van de publicatie	43,3	34,8	39,0	
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	36,1	27,8	31,9	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	34,5	25,8	30,1	
• citeren van datasets	35,6	28,3	31,9	
• aanbieden van datamanagement trainingen	14,4	19,2	16,8	
• instellen van periodieke dataaudits	13,9	13,6	13,8	

• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	15,5	12,6	14,0	
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	47,9	31,8	39,8	p=0%
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	20,1	31,3	25,8	p=1%
14. Welke van deze opties zouden bij u op bezwaren stuiten?				
• peer review van de dataset als onderdeel van peer review van de publicatie	22,7	28,8	25,8	
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	5,7	8,6	7,1	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	4,6	6,6	5,6	
• citeren van datasets	2,6	3,0	2,8	
• aanbieden van datamanagement trainingen	4,1	1,5	2,8	
• instellen van periodieke dataaudits	22,2	27,3	24,7	
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	26,8	35,4	31,1	
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	18,6	23,7	21,2	
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	9,3	6,6	7,9	
G. Enkele vragen over uw professionele achtergrond				
15. Mijn huidige functie is:				
• hoogleraar	60,3	61,1	60,7	
• universitair hoofddocent	37,6	33,3	35,5	
• universitair docent	1,0	1,0	1,0	
• postdoc medewerker	0,5	1,0	0,8	
• promovendus/AIO/OIO	0,0	1,0	0,5	
• anders	0,5	2,5	1,5	
16. Als wetenschapper vervul ik de volgende rollen: [MEERDERE ANTWOORDEN MOGELIJK]				
• Ik ben lid van een redactie van een peer-reviewed tijdschrift	58,8	55,1	56,9	
• Ik ben actief als peer reviewer van publicaties	98,5	91,9	95,2	p=0%
17. Wat betreft datasets vervul ik de volgende rollen: (MEERDERE ANTWOORDEN MOGELIJK]				
• Ik ben (mede-) producent van datasets	75,8	66,2	70,9	p=4%
• Ik heb wel eens mijn datasets ter beschikking gesteld aan derden	74,7	45,5	59,9	p=0%

(Mede-)producent van datasets	(Mede-) Producent van Datasets (n=278)	Niet (n=114)	Totaal percentage (alle respondenten; n=392)	Statistische significantie (Chi-kwadraat-toets)
	% (enigszins) mee eens			
1. In mijn vakgebied vragen veel tijdschriften om zo'n replicatie-dataset ter beschikking te stellen.	13,3	7,9	11,7	
2. In mijn vakgebied heeft de peer reviewer normaliter toegang tot de achterliggende dataset bij een publicatie en wordt de dataset meegenomen in de beoordeling van de publicatie.	14,4	13,2	14,0	p=4%
3. Het tegelijkertijd beoordelen van de publicatie en de achterliggende dataset is voor een peer reviewer in mijn vakgebied haalbaar.	33,5	29,8	32,4	
4. Ik vind het belangrijk dat de achterliggende dataset samen met de publicatie wordt beoordeeld in het peer review proces.	46,8	43,0	45,7	
B. Datapublicaties: peer-reviewed publicaties over datasets				
5. Voor mijn vakgebied zou een dergelijk tijdschrift met data publicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets.	54,0	43,0	50,8	
C. Commentaren over kwaliteit door hergebruikers van datasets				
6. Voor latere hergebruikers van datasets zouden deze commentaren over kwaliteit waardevol kunnen zijn.	82,7	78,1	81,4	
7. Als hergebruiker zou ik zeker de moeite nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset die ik heb hergebruikt.	75,5	64,0	72,2	
8. Als producent van een dataset zou ik dergelijke commentaren over kwaliteit van anderen verwelkomen.	80,9	68,4	77,3	p=1%
D. Citeren van datasets				
9. Als hergebruiker zou ik zeker de dataset citeren met deze mogelijkheid.	78,8	68,4	75,8	
10. Als wetenschapper zou ik zo'n citatie-mogelijkheid van 'mijn' dataset verwelkomen.	76,6	62,3	72,4	p=1%
E. Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium				
11. Trainingen op het gebied van datamanagement zouden op mijn vakgebied voorzien in een behoefte van veel vakgenoten.	54,7	48,2	52,8	
12. Data audits, waarin de totstandkoming en het beheer van datasets onder de loep worden genomen, zouden in mijn vakgebied een impuls aan verbeteringen in datamanagement kunnen geven.	58,6	44,7	54,6	p=1%
F. Uw visie	% aangevinkt	% aangevinkt		% aangevinkt
13. De meeste stimulans voor de kwaliteitsbeoordeling van datasets in mijn vakgebied zal waarschijnlijk uitgaan van:				
• peer review van de dataset als onderdeel van peer review van de publicatie	39,6	37,7	39,0	
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	30,2	36,0	31,9	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	30,2	29,8	30,1	
• citeren van datasets	36,0	21,9	31,9	p=1%
• aanbieden van datamanagement trainingen	17,3	15,8	16,8	
• instellen van periodieke data-audits	14,7	11,4	13,8	
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	16,2	8,8	14,0	p=5%
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	37,4	45,6	39,8	

• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	29,1	17,5	25,8	p=2%
14. Welke van deze opties zouden bij u op bezwaren stuiten?				
• peer review van de dataset als onderdeel van peer review van de publicatie	29,5	16,7	25,8	p=1%
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	8,6	3,5	7,1	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	5,0	7,0	5,6	
• citeren van datasets	2,5	3,5	2,8	
• aanbieden van datamanagement trainingen	3,2	1,8	2,8	
• instellen van periodieke dataaudits	23,4	28,1	24,7	
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	28,1	38,6	31,1	p=4%
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	24,5	13,2	21,2	p=1%
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	7,6	8,8	7,9	
G. Enkele vragen over uw professionele achtergrond				
15. Mijn huidige functie is:				
• hoogleraar	61,9	57,9	60,7	
• universitair hoofddocent	33,5	40,4	35,5	
• universitair docent	1,4	0,0	1,0	
• postdoc medewerker	1,1	0,0	0,8	
• promovendus/AIO/OIO	0,7	0,0	0,5	
• anders	1,4	1,8	1,5	
16. Als wetenschapper vervul ik de volgende rollen: [MEERDERE ANTWOORDEN MOGELIJK]				
• Ik ben lid van een redactie van een peer-reviewed tijdschrift	61,5	45,6	56,9	
• Ik ben actief als peer reviewer van publicaties	96,0	93,0	95,2	
17. Wat betreft datasets vervul ik de volgende rollen: (MEERDERE ANTWOORDEN MOGELIJK]				
• Ik heb wel eens mijn datasets ter beschikking gesteld aan derden	70,1	35,1	59,9	p=0%
• Ik ben hergebruiker van datasets (van andere onderzoekers en/of van grote onderzoeksfaciliteiten)				

Dataset wel eens ter beschikking gesteld versus niet	Dataset wel eens ter beschikking gesteld (n=235)	Niet- (n=157)	Totaal percentage (alle respondenten; n=392)	Statistische significantie (Chi-kwadraat-toets)
	% (enigszins) mee eens			
1. In mijn vakgebied vragen veel tijdschriften om zo'n replicatie-dataset ter beschikking te stellen.	14,5	7,6	11,7	
2. In mijn vakgebied heeft de peer reviewer normaliter toegang tot de achterliggende dataset bij een publicatie en wordt de dataset meegenomen in de beoordeling van de publicatie.	17,9	8,3	14,0	p=1%
3. Het tegelijkertijd beoordelen van de publicatie en de achterliggende dataset is voor een peer reviewer in mijn vakgebied haalbaar.	31,9	33,1	32,4	
4. Ik vind het belangrijk dat de achterliggende dataset samen met de publicatie wordt beoordeeld in het peer review proces.	47,2	43,3	45,7	
B. Datapublicaties: peer-reviewed publicaties over datasets				
5. Voor mijn vakgebied zou een dergelijk tijdschrift met data publicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets.	57,0	41,4	50,8	p=1%
C. Commentaren over kwaliteit door hergebruikers van datasets				
6. Voor latere hergebruikers van datasets zouden deze commentaren over kwaliteit waardevol kunnen zijn.	85,1	75,8	81,4	
7. Als hergebruiker zou ik zeker de moeite nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset die ik heb hergebruikt.	75,3	67,5	72,2	
8. Als producent van een dataset zou ik dergelijke commentaren over kwaliteit van anderen verwelkomen.	79,6	73,9	77,3	
D. Citeren van datasets				
9. Als hergebruiker zou ik zeker de dataset citeren met deze mogelijkheid.	78,7	71,3	75,8	
10. Als wetenschapper zou ik zo'n citatie-mogelijkheid van 'mijn' dataset verwelkomen.	77,9	64,3	72,4	
E. Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium				
11. Trainingen op het gebied van datamanagement zouden op mijn vakgebied voorzien in een behoefte van veel vakgenoten.	56,2	47,8	52,8	
12. Data audits, waarin de totstandkoming en het beheer van datasets onder de loep worden genomen, zouden in mijn vakgebied een impuls aan verbeteringen in datamanagement kunnen geven.	58,3	49,0	54,6	
F. Uw visie	% aangevinkt	% aangevinkt		% aangevinkt
13. De meeste stimulans voor de kwaliteitsbeoordeling van datasets in mijn vakgebied zal waarschijnlijk uitgaan van:				
• peer review van de dataset als onderdeel van peer review van de publicatie	39,1	38,9	39,0	
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	32,3	31,2	31,9	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	29,8	30,6	30,1	
• citeren van datasets	37,4	23,6	31,9	p=0%
• aanbieden van datamanagement trainingen	16,6	17,2	16,8	
• instellen van periodieke dataaudits	12,3	15,9	13,8	
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	14,9	12,7	14,0	
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	44,3	33,1	39,8	p=3%
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	28,9	21,0	25,8	

14. Welke van deze opties zouden bij u op bezwaren stuiten?				
• peer review van de dataset als onderdeel van peer review van de publicatie	30,6	18,5	25,8	
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	8,1	5,7	7,1	
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	6,0	5,1	5,6	
• citeren van datasets	1,7	4,5	2,8	
• aanbieden van datamanagement trainingen	3,8	1,3	2,8	
• instellen van periodieke dataaudits	26,4	22,3	24,7	
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	29,8	33,1	31,1	
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	20,4	22,3	21,2	
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	8,5	7,0	7,9	
G. Enkele vragen over uw professionele achtergrond				
15. Mijn huidige functie is:				
• hoogleraar	68,1	49,7	60,7	p=0%
• universitair hoofddocent	29,4	44,6	35,5	
• universitair docent	1,7	0,0	1,0	
• postdoc medewerker	0,0	1,9	0,8	
• promovendus/AIO/OIO	0,0	1,3	0,5	
• anders	0,9	2,5	1,5	
16. Als wetenschapper vervul ik de volgende rollen: [MEERDERE ANTWOORDEN MOGELIJK]				
Ik ben lid van een redactie van een peer-reviewed tijdschrift	68,1	40,1	56,9	p=0%
Ik ben actief als peer reviewer van publicaties	97,9	91,1	95,2	p=0%
17. Wat betreft datasets vervul ik de volgende rollen: (MEERDERE ANTWOORDEN MOGELIJK]				
Ik ben (mede-) producent van datasets	83,0	52,9	70,9	p=0%
Ik ben hergebruiker van datasets (van andere onderzoekers en/of van grote onderzoeksfaciliteiten)				p=0%

C3. Enquête resultaten onder DANS gebruikers

De resultaten staan weergegeven in de onderstaande tabel in een vergelijking met de resultaten van de enquête onder de hoogleraren/UHD's. Een zeer belangrijk verschil is de samenstelling van de beide bestanden: het mailing bestand van de EASY geregistreerden (=DANS gebruikers) blijkt voor een belangrijk deel uit niet-wetenschappers te bestaan: 59,2% geeft aan een ander beroep te hebben (hoogleraren/UHD's: 1,5 %). De overige 41,8% bestaat voor een belangrijk deel uit promovendi en postdoc medewerkers. Het percentage hoogleraren en UHD's is zeer klein (5,2% versus 96,2 %). Vandaar dat ook een lager percentage van de EASY geregistreerden lid is van de redactie van een peer reviewed tijdschrift (8,4 %) en ongeveer een derde actief is als peer reviewer van publicaties (34,1%). Wel is een groot deel van de EASY geregistreerden hergebruiker van datasets: 79,4 % versus 49,5% hoogleraren/UHD's.

Deze andere samenstelling van het bestand van EASY geregistreerden maakt dat de resultaten voor een deel minder relevant zijn voor wat betreft de kwaliteitsbeoordeling van wetenschappelijke onderzoek datasets. Het bestand is voor een belangrijk deel te karakteriseren als hergebruikers van datasets, uit wetenschappelijke en niet wetenschappelijke hoek, in de sociale en geesteswetenschappen. Daarom is het met name interessant op welke punten de visie van deze groep afwijkt van een de groep van hoogleraren/UHD's. Het meest opvallende verschil zit in de Open Access beschikbaarstelling van datasets: de hergebruikers van DANS zien daarvan vaker een goede stimulans uitgaan (54,7 % versus 39,8%, hoogleraren/UHD's) en zien ook minder bezwaren hiertegen: 10,8% versus 21,2%. Verder is het opvallend dat de vragen over de commentaren over kwaliteit door hergebruikers vrijwel hetzelfde zijn ingevuld door beide groepen.

Resultaten DANS gebruikers en Hoogleraren/UHD's vergeleken	**Hoogleraren/ UHD's**	**EASY-geregistreerden**	**Significantie**
	% (enigszins) mee eens		
1. In mijn vakgebied vragen veel tijdschriften om zo'n replicatie-dataset ter beschikking te stellen.	11,7	15	p=0
2. In mijn vakgebied heeft de peer reviewer normaliter toegang tot de achterliggende dataset bij een publicatie en wordt de dataset meegenomen in de beoordeling van de publicatie.	14	13,6	p=0
3. Het tegelijkertijd beoordelen van de publicatie en de achterliggende dataset is voor een peer reviewer in mijn vakgebied haalbaar.	32,4	35,2	ns
4. Ik vind het belangrijk dat de achterliggende dataset samen met de publicatie wordt beoordeeld in het peer review proces.	45,7	58,9	p=0
B. Datapublicaties: peer-reviewed publicaties over datasets			
5. Voor mijn vakgebied zou een dergelijk tijdschrift met data publicaties een waardevolle rol kunnen vervullen in het zichtbaar maken van belangrijke datasets.	50,8	67,6	p=0
C. Commentaren over kwaliteit door hergebruikers van datasets			
6. Voor latere hergebruikers van datasets zouden deze commentaren over kwaliteit waardevol kunnen zijn.	81,4	85	ns
7. Als hergebruiker zou ik zeker de moeite nemen om dergelijke commentaren over kwaliteit toe te voegen aan de dataset die ik heb hergebruikt.	72,2	76	ns
8. Als producent van een dataset zou ik dergelijke commentaren over kwaliteit van anderen verwelkomen.	77,3	79,4	ns
D. Citeren van datasets			
9. Als hergebruiker zou ik zeker de dataset citeren met deze mogelijkheid.	75,8	82,2	ns
10. Als wetenschapper zou ik zo'n citatie-mogelijkheid van 'mijn' dataset verwelkomen.	72,4	80,8	p=0
E. Ondersteuning bij kwaliteitszorg van datasets in een vroeg stadium			
11. Trainingen op het gebied van datamanagement zouden op mijn vakgebied voorzien in een behoefte van veel vakgenoten.	52,8	63,8	p=0

12. Data audits, waarin de totstandkoming en het beheer van datasets onder de loep worden genomen, zouden in mijn vakgebied een impuls aan verbeteringen in datamanagement kunnen geven.	54,6	63,8	p=0
F. Uw visie	% aangevinkt		
13. De meeste stimulans voor de kwaliteitsbeoordeling van datasets in mijn vakgebied zal waarschijnlijk uitgaan van:			
• peer review van de dataset als onderdeel van peer review van de publicatie	39	30	p=1
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	31,9	33,8	ns
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	30,1	35,9	ns
• citeren van datasets	31,9	34,8	ns
• aanbieden van datamanagement trainingen	16,8	19,9	ns
• instellen van periodieke dataaudits	13,8	7,7	p=1
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	14	15,3	ns
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	39,8	54,7	p=0
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	25,8	27,2	ns
14. Welke van deze opties zouden bij u op bezwaren stuiten?			
• peer review van de dataset als onderdeel van peer review van de publicatie	25,8	19,2	p=4
• opzetten van datapublicaties: peer reviewed beschrijvingen van datasets	7,1	5,2	ns
• commentaren over kwaliteit door hergebruikers, te publiceren bij de dataset	5,6	7,7	ns
• citeren van datasets	2,8	2,8	ns
• aanbieden van datamanagement trainingen	2,8	5,2	ns
• instellen van periodieke data-audits	24,7	21,3	ns
• verplichte datamanagement paragraaf in onderzoeksvoorstellen in te dienen bij onderzoeksfinanciers	31,1	20,6	p=0
• Open Access beschikbaarstelling van datasets, desgewenst na een embargoperiode	21,2	10,8	p=0
• Een gedragscode voor onderzoekers vaststellen over datamanagement en het beschikbaar stellen van datasets	7,9	16,4	p=0
G. Enkele vragen over uw professionele achtergrond			
15. Mijn huidige functie is:			
• hoogleraar	60,7	2,4	p=0
• universitair hoofddocent	35,5	2,8	p=0
• universitair docent	1	11,8	p=0
• postdoc medewerker	0,8	5,9	p=0
• promovendus/AIO/OIO	0,5	17,8	p=0
• anders	1,5	59,2	p=0
16. Als wetenschapper vervul ik de volgende rollen:			
• Ik ben lid van een redactie van een peer-reviewed tijdschrift	56,9	8,4	p=0
• Ik ben actief als peer reviewer van publicaties	95,2	34,1	p=0
17. Wat betreft datasets vervul ik de volgende rollen:			
• Ik ben (mede-) producent van datasets	70,9	56,8	p=0
• Ik heb wel eens mijn datasets ter beschikking gesteld aan derden	59,9	47	p=0
• Ik ben hergebruiker van datasets (van andere onderzoekers en/of van grote onderzoeksfaciliteiten)	49,5	79,4	p=0