Organisatorische aspecten duurzame opslag en beschikbaarstelling onderzoeksdata

Colofon

Organisatorische aspecten duurzame opslag en beschikbaarstelling onderzoeksdata

Dit rapport is geschreven in het kader van het SURFshare programma, door Maurits van der Graaf; Pleiade Management en Consultancy .

SURFshare
Postbus 2290
3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60
E info@surf.nl
W www.surf.nl/surfshare

Auteurs
Maurits van der Graaf
Pleiade Management and Consultancy BV
www.pleiade.nl

Eindredactie

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

Inhoudsopgave

Management samenvatting

- **Deze studie:** Deze studie van SURFfoundation is gericht op het in kaart brengen van de organisatorische aspecten van de opslag en terbeschikkingstelling van onderzoekdatasets ten behoeve van beleidsmakers van onderzoeksinstellingen. De studie is gebaseerd op een reeks interviews en een review van de recente literatuur.
- **Wetenschappelijk onderzoek steeds meer data-driven:** Het volume van de onderzoekdatasets stijgt sterk terwijl het delen van datasets een vlucht neemt en tot nieuwe vormen van wetenschapsbeoefening leidt.
- **Bewaren met het oog op validatie en hergebruik:** Er zijn op dit moment diverse regels en redenen om onderzoekdatasets voor langere tijd te bewaren. De Gedragscode Wetenschapsbeoefening van de VSNU schrijft een bewaartermijn van vijf jaar voor van de onderzoeksdata na publicatie met het oog op validatie. Steeds meer tijdschriften eisen de terbeschikkingstelling van een zogenaamde replicatiedataset, eveneens vooral met het oog op validatie. Daarnaast zijn er wetenschappelijke en niet-wetenschappelijke redenen om onderzoekdatasets te bewaren en ter beschikking te stellen met het oog op hergebruik. Enkele onderzoeksfinanciers hebben daarom verplichtingen ingesteld om onderzoekdatasets zodanig ter beschikking te stellen dat zij ook voor hergebruik geschikt zijn.
- **Organisatievormen voor opslag en terbeschikkingstelling van onderzoeksdata:** Het landschap voor wat betreft de organisatievormen voor opslag en terbeschikkingstelling van onderzoekdatasets vertoont een grote verscheidenheid, is versnipperd en bevat lacunes. Er zijn vier categorieën van organisatievormen te onderscheiden:
    - **Datacentra:** Meestal behorend bij een onderzoeksfaciliteiten en disciplinair en/of thematisch gericht. In het rapport beschreven voorbeelden hiervan zijn de radiotelescoop LOFAR, de aan elkaar gekoppelde datacentra van de instituten van de Nationale Oceanografische Data Commissie en de datacentra van het KNMI. Deze datacentra bevatten de onderzoeksdata vanaf de creatie en in veel gevallen is datasharing in een vroeg stadium normaal en goed geregeld, maar vooral beperkt tot de ‘eigen community’.
    - **Data-archieven:** Data-archieven zijn eveneens in de meeste gevallen gericht op een of meerdere disciplines en zijn gefocust op een wijde beschikbaarstelling en langdurige bewaartermijn. Onderzoekers in hun rol van dataproducent dragen na afloop van hun onderzoek de onderzoekdatasets over. In het rapport beschreven voorbeelden hiervan zijn de data-archieven van DANS, het Max Planck Instituut en het 3TU Datacentrum. Het laatste fungeert als een institutioneel datarepository voor de drie

technische universiteiten. Institutionele datarepositories, nog relatief schaars, vallen eveneens onder deze categorie.
    - **Door onderzoekers zelf georganiseerd:** Onderzoekers organiseren in veel gevallen zelf de opslag en beschikbaarstelling van de door hen geproduceerde onderzoekdatasets. Deze worden dan bijvoorbeeld intern opgeslagen of zijn via de website van de onderzoeksgroep beschikbaar. Wijdere beschikbaarstelling en een langdurige bewaartermijn zijn in deze situaties meestal niet gegarandeerd.
    - **Supplementary data:** Een toenemend aantal wetenschappelijke tijdschriftredacties eisen dat de onderliggende datasets bij een artikel (replicatiedatasets) Open Access ter beschikking worden gesteld. Tijdschriften bieden daar zelf mogelijkheden toe (maar vaak met beperkingen qua omvang en qua dataformats), de meeste uitgevers hebben (nog) geen duidelijke services ontwikkeld om een grote toestroom van datasets van aanzienlijk volume en grote gevarieerdheid aan te kunnen. Het is niet waarschijnlijk dat uitgevers een dergelijke rol op zich zouden kunnen of willen nemen.
- **Ervaringen met institutioneel beleid:** De ervaringen van enkele onderzoeksinstellingen met het opzetten en uitvoeren van een institutioneel beleid op dit gebied werden in kaart gebracht door middel van interviews. Enkele belangrijke opties voor een institutioneel beleid zijn:
    - **Beleidsdocument:** Opstellen van een institutioneel beleidsdocument, waarin de verantwoordelijkheden voor opslag en beheer van datasets geregeld wordt en een beleidslijn ten aanzien van beschikbaarstelling wordt vastgesteld.
    - **Data audits:** Uitvoeren van data audits, waarin de huidige datamanagement praktijken onder de loep worden genomen
    - **Datamanagement:** Het verzorgen van trainingen en opleidingen op het gebied van datamanagement en voorlichting en advies hierover.
    - **Dataopslagfaciliteiten:** Zorgdragen voor adequate dataopslagfaciliteiten voor onderzoekdatasets tijdens de datacreatie en het gebruik ervan.
    - **Registratie van onderzoekdatasets:** Opzetten van een registratiesysteem voor onderzoekdatasets gekoppeld aan het onderzoekinformatiesysteem van de instelling
    - **Institutioneel datarepository:** Een institutioneel datarepository is functioneel voor replicatiedatasets, die onderzoekers dienen te deponeren met het oog op validatie als gevolg van de Code Wetenschapsbeoefening en/of als gevolg van verplichtingen van wetenschappelijke tijdschriften en voor die datasets, waarvan het belangrijk is om deze langdurig te bewaren en ter beschikking te stellen met het oog op hergebruik, maar waarvoor geen disciplinegericht dataarchief bestaat.

- **Stakeholders:** Er worden zeven belangrijke stakeholders voor wat betreft onderzoeksdata onderscheiden, ieder met een verschillende rol en met verschillende taken en verantwoordelijkheden:
    - **Dataproducenten en dataconsumenten:** De wetenschappers die datasets produceren en deze ter beschikkingstellen en de wetenschappers die deze datasets hergebruiken. Voor de dataproducerende wetenschappers bestaan er belangrijke drempels om door hen geproduceerde datasets ter beschikking te stellen, zowel om inhoudelijke redenen als om praktische redenen. Bij het laatste gaat het vooral om het opstellen van adequate metadata en het leveren van adequate documentatie bij de dataset.
    - **Onderzoeksinstellingen:** Onderzoeksinstellingen zijn verantwoordelijk voor het aanbieden van adequate faciliteiten voor datamanagement van hun onderzoekers, waaronder de mogelijkheid voor de opslag en beschikbaarstelling van onderzoekdatasets.
    - **Het data-archief of datacentrum:** Data-archieven of datacentra hebben de taak om onderzoekdatasets voor de langere termijn te bewaren en zo wijd mogelijk ter beschikking te stellen.
    - **De onderzoeksfinanciers, de uitgevers en de aggregators:** onderzoeksfinanciers hebben een belangrijke stem in het bepalen van het beleid ten aanzien van onderzoeksdata. Een aantal onderzoeksfinanciers vereisen Open Access beschikbaarstelling van datasets van onderzoeksprojecten die door hen gefinancierd worden. Uitgevers van wetenschappelijke tijdschriften hebben als taak om te zorgen dat er een koppeling plaatsvindt tussen de wetenschappelijke artikelen en de onderliggende datasets. Tenslotte is de taak van zgn. aggregators om een search & discovery dienst te onderhouden.
- **Visie op data-infrastructuur:** In meerdere landen worden er acties ondernomen om gebruikmakend van de huidige infrastructuur deze aan te vullen om een dekkend en inter-operabel netwerk van datacentra, data-archieven en institutionele datarepositories in te richten. Daarbij wordt een taakverdeling voorzien tussen onderzoeksinstellingen, die registratie van datasets en de meer korte termijn beschikbaarstelling van datasets op zich nemen en een netwerk van data-archieven en datacentra, nationaal en internationaal, meestal langs disciplinaire lijnen georganiseerd, waar de lange termijn archivering en beschikbaarstelling van belangrijke onderzoekdatasets wordt georganiseerd.
- **Aanbevelingen op nationaal niveau:** Aanbevolen wordt om de Gedragscode Wetenschapsbeoefening van de VSNU aan te scherpen wat betreft de verantwoordelijkheden voor opslag, beheer en de beschikbaarstelling van onderzoekdatasets. Tevens wordt een actieve rol voor NWO en andere onderzoeksfinanciers aanbevolen, o.a. door het voorschrijven van een paragraaf over datamanagement in

onderzoeksvoorstellen en het instellen van verplichtingen ten aanzien van de beschikbaarstelling van de resulterende datasets. Tenslotte wordt gepleit voor een nationale regierol voor DANS, het data instituut van de KNAW en NWO, voor alle disciplines.

- **Aanbevelingen op institutioneel niveau:** Aanbevolen wordt om een institutioneel beleid op het gebied van datamanagement en data sharing vast te stellen. Beleidsinstrumenten zijn:
    - Vaststellen van een institutioneel beleid op het gebied van datamanagement en data sharing.
    - Het uitvoeren van data audits om de voor de instelling specifieke lacunes in datamanagement vast te stellen
    - Verbeteren van het datamanagement door het aanbieden van trainingen, het opnemen van datamanagement in het curriculum van master en/of promovendi en het aanbieden van adviesdiensten en voorlichting
    - Het aanbieden van adequate dataopslagfaciliteiten
    - Registratie van datasets in aansluiting op de registratie van andere onderzoeksinformatie
    - Het inrichten van een datarepository (al dan niet samen met andere onderzoeksinstellingen) gericht op het opslaan van replicatie datasets voor validatie doeleinden en op het opslaan van onderzoekdatasets, wanneer de discipline zelf geen faciliteiten daarvoor heeft.
- **Open Access stimuleren:** Voor het stimuleren van Open Access toegang tot onderzoekdatasets worden twee - elkaar niet uitsluitende - beleidscenario's gepresenteerd. Een universitair beleidscenario gericht op het verleiden van wetenschappers tot Open Access beschikbaarstelling zal zich richten op het delen van datasets en daarbij het toegangsbeheer in eerste instantie in handen van de betreffende dataproducenten te laten. In dit scenario zullen de initiatieven om het publiceren van datasets mee te laten tellen in het ‘academic record’ , bijvoorbeeld op basis van citaties van datasets, zoveel mogelijk worden gestimuleerd. Het beleidscenario gericht op verplichten ligt vooral op het terrein van wetenschappelijke tijdschriften en onderzoeksfinanciers.

Management summary

- **This study:** this study by SURFfoundation is focused on the organisational aspects of the preservation and access of research datasets and meant for policymakers at research institutes. The study is based on a series of interviews and a review of the recent literature.
- **Research increasingly data-driven:** the volume of research datasets is strongly increasing while datasharing becomes increasingly common and leads to new forms of science.
- **Preservation for validation purposes and reuse**: at this moment there are several rules and reasons to preserve research datasets for the longer-term. The VSNU Code of Conduct for Scientific Practice prescribes the storage of research data after publication for at least 5 years for validation purposes. Increasingly, scientific journals require that the so-called replication dataset is made available in conjunction with the article, also mainly for validation purposes. In addition, there are a scientific and non-scientific reasons to preserve research datasets and make them accessible for purposes of reuse. Therefore, some research funding organisations have made obligatory that research datasets from projects funded by them has to be made available for purposes of reuse.
- **Categories of organisations for preservation and access of research data:** the landscape for the various organizations for preservation and access of research data is fragmented and diverse, while not providing a full coverage for all disciplines. Four categories of organisations can be distinguished**:**
    - **Data centres:** belonging to research facilities, mostly discipline-oriented. In this report, examples of these data centres are described in detail: the radio telescope LOFAR, the data centres of the institutes of the National Oceanographic Data Committee of the Netherlands and the data centres of the Royal Netherlands Meteorological Institute. These data centres contain the research data from creation onwards, in most cases datasharing is normal and well-organised although mainly limited to the own research community.
    - **Data archives:** data archives are also mostly discipline oriented and are focused on giving wider access to research data and the long-term preservation. Data producing researchers transfer the dataset after finishing their study. In this report, the data archives of DANS, the Max Planck Institute and the 3TU Data Centre are described in more detail. The 3TU Data Centre can also been seen as an institutional data repository for the three Dutch technical universities. Institutional data repositories - still rather few in numbers – may fall under this category as well.
    - **Self-organisation:** in many cases, researchers organise the storage and access to research data themselves. These research datasets are for instance stored internally or are accessible via

the website of the research group. A wider accessibility and the long-term preservation are often not guaranteed in these situations.
    - **Supplementary data:** an increasing number of editorial boards of scientific journals require that the underlying datasets (replication datasets) are published via Open Access in conjunction with article. Journals often give opportunities for depositing these datasets (however often with limitations on volume and data formats). Most publishers do not have (yet) developed services to handle large numbers of datasets, especially when voluminous and with a great variety of data formats. It is also unlikely that publishers would want to have such a role.
- **Learning experiences on institutional level:** The learning experiences of a number of research institutes with regard to the development and implementation of an institutional policy on research data were collected by a number of interviews. Some important options for an institutional policy are:
    - **Policy document:** the development of a policy document describing the institutional policy with regard to the responsibilities on the storage, preservation and the access management of research datasets.
    - **Data audits:** data audits to assess the management of the research datasets.
    - **Data management:** data management training for researchers and in the curriculum for Ph.D. students
    - **Data storage facilities:** adequate data storage facilities for research datasets during creation and usage
    - **Registration of research datasets:** a registration system – aligned with other registration procedures for current research information
    - **Institutional data repository:** an institutional data repository can be useful for replication datasets, that researchers have to deposit for validation purposes following the requirements of the Code of Conduct or by scientific journals and for those datasets that are important to preserve and to be made accessible for purposes of reuse, but in a discipline that has no relevant national or international data archive available.
- **Stakeholders:** seven stakeholders with regard to research datasets have been identified, each with a different role and with different tasks and responsibilities:
    - **data producers and data consumers:** scientists that produce datasets and make them accessible and scientists that reuse these datasets. There appear to be a number of important disincentives for data producing scientists. Some of these disincentives are related to scientific practices, some are purely practical: for instance the time involved in producing meta data and documentation with a dataset.

    - **Research institutes:** research institutes are responsible for creating adequate data management facilities for their researchers, including the option to preserve and give access to research datasets.
    - **Data archives or data centres:** data archives or data centres have responsibility to preserve datasets for the longer-term and make them accessible as widely as possible.
    - **Research funding organisations, publishers and aggregators:** research funding organisations have an important role in setting policies with regard to research datasets. A number of research funding organisations require Open Access accessibility of datasets produced by research projects that are funded by them. Publishers of scientific journals have the task to organize links between journal articles and the underlying datasets. Aggregators have the task to set up and maintain a search and discovery service.
- **Data infrastructure vision:** in several countries, plans are developed and implemented in order to create an interoperable network of data centres, data archives and institutional data repositories that covers all scientific disciplines. In these efforts, registration of datasets and the short-term storage of datasets is seen as a task for individual research institutes. A network of data archives and data centres, nationally and internationally organised, should have the tasks of providing long-term preservation and access to important research datasets.
- **Recommendations for the national level:** for the national level in the Netherlands, it is recommended to elaborate the Code of Conduct of the VSNU with regard to the responsibilities on preservation and access of research datasets. In addition, a more active role for the Dutch research funding organisations is recommended by prescribing a data management paragraph in grant proposals and by requiring accessibility to research datasets that results from the research grants. In addition, a national role for DANS for all disciplines is recommended.
- **Recommendations for the institutional level:** it is recommended to develop an institutional policy with regard to data management and datasharing. Possible instruments for such a policy are:
    - an institutional policy document
    - data audits that can assess possible, for that institute specific problems in the data management
    - improving data management by offering training and making it part of the curriculum for master students or viewers the students
    - the offering of adequate data storage facilities
    - registration of datasets in alignment is the registration procedures for other current research information in the institute
    - the setting up of a institutional datarepository (possibly together with other research institutes) focused on replication datasets for validation purposes and on research datasets for purposes of reuse, if that particular discipline has no facilities for this.

- **Promoting Open Access:** two possible - not mutually exclusive - scenarios for promoting Open Access are presented. One scenario for research institutes is primarily focused on making datasharing attractive for scientists. This scenario thus will promote access management controlled by the data producing scientists. In addition, initiatives to make the publishing of datasets visible in the academic record of the scientists (such as citations of datasets) are very important. The other scenario is focused on making Open Access for research datasets obligatory: this is more in the domain of scientific journals and of research funding organisations.

1. Inleiding

Voor de duurzame opslag en beschikbaarstelling van onderzoekdatasets is een aantal belangrijke argumenten:

- Hergebruik van data voorkomt duplicaat onderzoek
- Toegang tot onderzoeksdata maakt het mogelijk om onderzoeksresultaten te valideren
- Wetenschappelijk onderzoek wordt steeds meer data-intensief en 'data-driven': nieuwe vormen van onderzoek ontstaan rond grote datasets (bijvoorbeeld bioinformatica rond genoom datasets) en stellen wetenschappers in staat om tegelijkertijd te werken aan en met dezelfde data.

Om deze redenen is er een toenemende belangstelling voor dit onderwerp (zie ook het tekstkader hiernaast).

SURFfoundation laat daarom binnen het SURFshare programma onderzoek uitvoeren naar verschillende aspecten rond onderzoekdatasets.
In deze studie staat de vraag centraal welke organisatorische aspecten een rol spelen bij duurzame opslag en beschikbaarstelling van onderzoekdatasets. Het rapport is bedoeld voor beleidsmakers van onderzoeksinstellingen als handreiking voor het verder ontwikkelen van het instituutsbeleid op dit gebied.
In het rapport is ernaar gestreefd om voor de hierboven genoemde doelgroep een handzaam overzicht te bieden van de diverse organisatievormen, van de verschillende stakeholders, de ontwikkelingen wat betreft de onderzoekdatasets zelf, de opties voor institutioneel beleid en de rol en taken van

Contouren van een actieprogramma door Nature
In een recent redactioneel commentaar (referentie 1) onderschrijft het wetenschappelijke tijdschrift Nature het belang van delen van onderzoekdatasets en geeft een korte aanzet voor een actieprogramma om dit te realiseren. Gesteld wordt dat er bij de meeste wetenschappelijke disciplines een duidelijk technisch, institutioneel en cultureel kader ontbreekt. Het voorgestelde actieprogramma om dit aan te pakken omvat de volgende onderdelen:

- **Onderzoeksfinancierende organisaties:** Onderzoeksfinanciers moeten inzien dat het bewaren en toegang geven tot onderzoekdatasets een centraal onderdeel is van hun missie en een actief beleid hierop voeren.
- **Metadata:** metadatamanagement en gereedschappen om dit te vergemakkelijken zijn essentieel om datasets te vinden en te hergebruiken.
- **Erkenning voor onderzoekers:** wetenschappers moeten erkend worden voor het creëren en/of bijdragen aan wetenschappelijke datasets.
- **Langdurig gefinancierde datacentra:** de onderzoeksgegevens en de wetenschappelijke gemeenschap dienen samen een netwerk van datacentra te creëren die de datasets bewaren voor de langere termijn. Belangrijk daarbij is een robuuste, voor de langere termijn gegarandeerde financiering voor de datacentra.
- **Educatie over datamanagement:** universiteiten en de wetenschappelijke disciplines dienen een uitgebreid educatie programma te starten over datamanagement. Nature stelt dat datamanagement onderdeel zou moeten uitmaken van elk wetenschappelijk curriculum.

onderzoeksinstellingen in een verder te ontwikkelen nationale infrastructuur. Het rapport wordt afgesloten met een aantal aanbevelingen voor institutioneel beleid op dit gebied, die naar verwachting de basis kunnen vormen voor een – voor iedere onderzoeksinstelling specifiek toe te spitsen – plan van aanpak. Gezien de nauwe samenhang met landelijke ontwikkelingen bevat het rapport ook een aantal aanbevelingen voor het nationale niveau.

Dit onderzoek is uitgevoerd in opdracht van de Werkgroep Juridische en Organisatorische aspecten van onderzoeksdata van het Onderzoeksdata Forum. Dit is een krachtenbundeling van SURFfoundation, KNAW, NWO, DANS, 3TU, Universiteit van Tilburg en de Nationale Coalitie Digitale Duurzaamheid (NCDD) en wordt uitgevoerd in het kader van het SURFshare programma. De werkgroep heeft de kaders geformuleerd voor dit onderzoek. Het onderzoek is uitgevoerd door Maurits van der Graaf van Pleiade Management en Consultancy, onder supervisie van Wilma Mossink van SURFfoundation en begeleiding door de leden van de genoemde werkgroep.

2. Methoden

2.1 Onderzoeksvragen

Het uitgangspunt voor het onderzoek is dat voor sommige groepen wetenschappers het noodzakelijk zal zijn om onderzoeksdata beschikbaar te stellen en langdurig te bewaren. De kernvraag van het onderzoek is hoe dit georganiseerd kan worden. Het onderzoek was daarom gefocust op de volgende onderwerpen:

- **Organisatorische oplossingen rond het opslaan en beschikbaar stellen van onderzoeksdata:**
    - Wat zijn mogelijke organisatievormen? Welke partijen spelen een rol op het terrein van opslag en delen van onderzoeksdata?
    - Welke data komen in aanmerking voor opslag en publicatie?
    - Wat zijn de incentives voor de onderzoekers?
    - Welke regels worden gehanteerd voor toegang en hergebruik van data?
    - Hoe hebben onderzoeksinstellingen die al in een verder stadium zijn de opbouwfase aangepakt?
    - Welke diensten worden aangeboden rond het opslaan van onderzoeksdata en hoe zijn deze georganiseerd?
- **Analyse en advies:**
    - Naast een inventarisatie dient het onderzoek uitdrukkelijk de voor- en nadelen van de diverse organisatorische mogelijkheden te analyseren en te komen tot een advies over in welke omstandigheden welke oplossing het beste is.

2.2 Onderzoeksopzet

Het onderzoek bestaat uit twee onderdelen: een desk research en literatuurstudie, en elf persoonlijke interviews.

Voor de interviews werd een aantal sleutelfiguren aangezocht bij data-archieven en datacentra van grote onderzoeksfaciliteiten in Nederland. Daarnaast werden enige sleutelfiguren aangezocht wat betreft institutioneel beleid op dit gebied, waarvoor ook een aantal internationale instellingen werd benaderd. Omdat ook de wetenschappelijke uitgeverijen een belangrijke rol spelen, werd tevens een vertegenwoordiger van de uitgeverijen geïnterviewd.

De geïnterviewden zijn hieronder weergegeven in tabel.

Respondent	**Instituut**
Paul Taylor	University of Melbourne
Robin Rice	University of Edinburgh
Laine Rees	University of Toronto
Jacquelijn Ringersma	Max Planck instituut Nijmegen
Wim Som de Cerff	KNMI
Peter Doorn	DANS
Renze Brandsma	UB Amsterdam
Jeroen Rombouts	3TU Datacentrum
Taco de Bruin	Nationale Oceanografische Data Commissie/ NIOZ
Hanno Holties	LOFAR
Eefke Smit	STM

Voor het deskresearch werd vooral gebruik gemaakt van de vele rapporten die op dit gebied recentelijk zijn verschenen. Gezien de vele recente ontwikkelingen werd vooral gebruikgemaakt van rapporten van 2008 en later. Vanwege het karakter van het rapport - een handreiking voor beleidsmakers - is uitsluitend direct relevante literatuur geciteerd.

3.Onderzoeksdata

3.1 Soorten onderzoekdatasets

Onderzoekdatasets kunnen verschillende vormen aannemen: wetenschappelijke disciplines kunnen immers enorm van elkaar verschillen. In het tekstkader hiernaast staat een veelgebruikte indeling voor onderzoeksdata. Er wordt onderscheid gemaakt in observationele onderzoeksdata, experimentele onderzoeksdata, data uit computermodellen, afgeleide of samengestelde data en referentie datasets/canonieke datasets.
Onderzoekdatasets kunnen ook op andere manieren ingedeeld worden: Treloar e.a. (referentie 8) hebben het begrip 'datacuratie continuüm' geïntroduceerd: daarin worden datasets beschreven langs de lijn van datacreatie tot archivering van de dataset voor de langere termijn. Ook is het mogelijk om datasets in te delen op basis van het type data producent:

- individuele onderzoekers en of kleine onderzoeksgroepen
- grotere onderzoeksgroepen of combinaties van onderzoeksgroepen
- (door veel onderzoekers gedeelde/gebruikte) onderzoeksfaciliteiten.

Classification of research data

- **Observational:** data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, neuroimages.
- **Experimental:** data from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, toroid magnetic field data.
- **Simulation:** data generated from test models where model and metadata are more important than output data. For example, climate models, economic models.
- **Derived or compiled:** data is reproducible but expensive. For example, text and data mining, compiled database, 3D models.
- **Reference or canonical:** a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals.

From: Stewardship of digital research data: a framework of principles and guidelines, Research Information Network, January 2008(3); taken from the website of the University of Edinburgh

Deze laatste indeling is vooral relevant met het oog op het delen van datasets en de wijze van archivering en beschikbaarstelling[1].

[1] Zie ook Waaijers en van der Graaf, 'Over de kwaliteit van onderzoekdata';SURFfoundation rapport in voorbereiding (referentie 28) en referentie 26.

3.2 Hergebruik en bewaartermijn

In een recent uitgevoerd onderzoek voor SURFfoundation[2] worden de redenen om onderzoekdatasets te bewaren uiteengezet:

- **Bewaarplicht van minimaal vijf jaar voor alle onderzoekdatasets:** Op grond van de Nederlandse gedragscode wetenschapsbeoefening (VSNU 2004) is er een minimale bewaarplicht voor de onderzoeker van de onderzoeksgegevens bij een publicatie van vijf jaar voor validatie doeleinden.
- **Verplichtingen van wetenschappelijke tijdschriften:** Een toenemend aantal wetenschappelijke tijdschriften verplichten de auteurs de bij het artikel behorende onderzoekdataset terbeschikkingstelling. De primaire doelstelling hiervan is validatie/replicatie.
- **Verplichtingen door onderzoeksfinanciers:** Onderzoeksfinanciers verplichten in een aantal gevallen om de onderzoeksdata voor hergebruik te bewaren en ter beschikking te stellen.
- **Wetenschappelijke redenen om de onderzoekdatasets te bewaren met het oog op hergebruik:** Mogelijke redenen kunnen zijn het belang van de data (potentiële waarde voor hergebruik, internationale of nationale positionering en kwaliteit, oorspronkelijkheid, grote, schaal, productiekosten of innovatief karakter van het onderzoek; uniciteit van de data [de data bevatten niet herhaalbare waarnemingen] en het belang van de data voor historisch onderzoek).
- **Niet-wetenschappelijke redenen:** Er kunnen ook niet-wetenschappelijke redenen zijn om een onderzoekdataset te bewaren: als voorbeeld wordt genoemd cultureel erfgoed.

[2] Selection of Research Data, SURFfoundation, a report by DANS and 3TU Data Centre, in press (referentie 27)

3.3 Trends

Zoals in de inleiding al vermeld tekent zich een aantal ontwikkelingen af die extra aandacht voor de organisatorische aspecten rond opslag en beschikbaarstelling van onderzoekdatasets noodzakelijk maken. De belangrijkste ontwikkelingen zijn:

- **De omvang van de onderzoekdatasets stijgt enorm:** in een groot survey onder onderzoekers (n=1389; referentie 14) geven onderzoekers van alle disciplines aan dat door hen geproduceerde onderzoekdatasets in vijf jaar sterk in omvang zullen toenemen: zo verwacht 36% een omvang van de door hen geproduceerde datasets over 5 jaar van 1 gigabyte tot 1 terabyte, 20% een omvang van 1 terabyte tot 1 petabyte en 5% van 1 tot 10 petabyte.
- **Tijdschriften eisen vaker beschikbaarstelling replicatiedatasets**: Steeds meer wetenschappelijke tijdschriften eisen dat de onderliggende dataset bij het tijdschriftartikel (ook wel replicatiedataset genoemd[3]) ter beschikking wordt gesteld bij het artikel[4].
- **Hergebruik:** De beschikbaarstelling van grote onderzoekdatasets maken nieuwe vormen van onderzoek mogelijk. Het gaat bijvoorbeeld om meta-analyses (waarbij meerdere datasets worden samengevoegd en opnieuw geanalyseerd). Deze nieuwe vormen van wetenschapsbeoefening worden wel data-exploratie of het Fourth Paradigm of Science[5] genoemd.
- **Open Access:** Er komt steeds meer nadruk op het Open Access beschikbaarstellen van onderzoekdatasets in lijn met de Open Access beweging voor publicaties.

[3] Een replicatiedataset kan een subset zijn van de gehele onderzoekdataset: de onderzoekers beschrijven resultaten van een deel van de onderzoekdataset, terwijl andere delen van de onderzoekdataset in andere tijdschriftartikelen beschreven worden.
[4] Uit een survey onder wetenschappers van alle disciplines melden ca. 10% van de respondenten dat dit reeds usance is bij veel tijdschriften in hun vakgebied (Waaijers en van der Graaf, in voorbereiding; referentie 28)
[5] Naar het boek 'The Fourth Paradigm: Data-intensive Scientific Discovery, T. Hey, S. Tansley, K. Tolle (eds.) 2009

4. Overzicht organisatievormen voor opslag en beschikbaarstelling van onderzoekdatasets

4.1 Inleiding

In appendix A zijn de resultaten van de interviews weergegeven in de vorm van case reports. Uit deze interviews en uit de literatuur [6] blijkt een grote verscheidenheid van organisatievormen in een versnipperd landschap van dataopslag en data beschikbaarstelling. Hieronder volgt daarom een indeling van organisatievormen, opgesteld op basis van de interviews, met het doel om het huidige landschap van organisatievormen te beschrijven en te bespreken [7]:

- Datacollecties georganiseerd in datacentra rond disciplinaire of thematische onderzoeksactiviteiten
- Data-archieven en repositories, gericht op hergebruik van onderzoekdatasets
- Door onderzoekers zelf georganiseerde opslag en beschikbaarstelling
- Datacollecties georganiseerd rond wetenschappelijke tijdschriften

Deze vier categorieën worden hieronder nader toegelicht en besproken.

4.2 Datacollecties georganiseerd in datacentra rond disciplinaire of thematische onderzoeksactiviteiten

In een aantal gevallen zijn opslag en beschikbaarstelling van onderzoekdatasets georganiseerd rond onderzoeksfaciliteiten. De organisatievormen hiervan kunnen enigszins verschillen, maar in grote lijnen gelden dezelfde kenmerken:

- De datacentra zijn op een specifieke discipline en/of vakgebied georiënteerd.
- Het hele continuüm van datacreatie tot langdurig bewaren (datacuratie) wordt binnen hetzelfde datacentrum georganiseerd: er is géén moment van overdracht.
- Het delen van datasets (datasharing) vindt in veel gevallen in een vroeg stadium al plaats, is ingebakken in de cultuur van de betreffende wetenschappelijke gemeenschap en in de regels van het datacentrum.
- De datasets worden in principe voor de langere termijn bewaard, hoewel dit in veel gevallen als een inspanningsverplichting wordt gezien.

[6] Zie met name het onderzoek door Research Information Network (referentie 4) en een inventarisatie van Europese data-initiatieven (referentie 17).

[7] Er worden door de grote IT bedrijven cloud computer services aangeboden, die onder andere mogelijkheden bieden voor dataopslag. Voorbeelden zijn Microsoft Windows Azure platform en the Google App Engine. In de meeste gevallen zijn hier wel kosten aan verbonden. Bij de aanvang van het onderzoek werd verondersteld dat deze ontwikkeling mogelijk relevant zou kunnen zijn voor de beschikbaarstelling van onderzoekdatasets. Uit de literatuurstudie en de interviews zijn geen aanwijzingen gevonden dat deze dienst op dit moment gebruikt worden. Ook lijkt het erop dat deze commerciële 'cloud computer'-diensten mogelijk de eigen servers van de diverse instellingen zouden kunnen vervangen, maar dat de organisatie van de beschikbaarstelling van onderzoekdatasets in handen van de onderzoeksinstellingen blijft.

Hieronder volgt een beschrijving van drie organisatievormen binnen deze categorie:

- **Datacentra gekoppeld aan grote onderzoeksfaciliteiten**: bijvoorbeeld in de Astronomie wordt een belangrijk deel van het onderzoek gedaan door middel van grote telescopen, waaraan datacentra zijn gekoppeld. De werkwijze is als volgt: een onderzoeker of onderzoeksgroep vraagt een bepaalde waarnemingstijd aan met een observatie instrument. De aanvragers krijgen dan voor een bepaalde periode (bij LOFAR bijvoorbeeld 1 jaar) exclusieve toegang tot de data om deze te analyseren en erover te publiceren. Na dit jaar worden de data dan voor andere onderzoekers beschikbaar gesteld. Een vergelijkbare werkwijze geldt bij het door NWO gesubsidieerde LISS panel van Centerdata: daar worden maandelijks 5000 Nederlandse huishoudens bevraagd. Wetenschappers van verschillende disciplines (sociale en economische wetenschappen, maar in toenemende mate ook biomedische wetenschappen) laten onderzoek verrichten onder deze huishoudens. Zij krijgen dan toegang tot de resulterende dataset. In het geval van het LISS panel wordt de dataset ook direct publiek beschikbaar gesteld.
- **Netwerk van databases met centrale index:** de Nationale Oceanografische Data Commissie is een voorbeeld van een netwerk van databases met onderzoeksdata, aan elkaar gekoppeld en toegankelijk gemaakt door een centrale index. Het betreft een samenwerkingsverband van acht Nederlandse onderzoeksinstituten met het doel om oceanografische gegevens aan een zo breed mogelijke gebruikersgroep beschikbaar te stellen. Naast de centrale index van alle gegevens, is er een toegang via Internet tot de databases door een volledig gedistribueerd systeem. De NODC is tevens een onderdeel van een Europese infrastructuur van een netwerk van datacentra, SeaDataNet. Het Nationaal Georegister kent een vergelijkbare opzet: hierin worden onderzoeksdata op het gebied van geo-informatie doorzoekbaar gemaakt [8].
- **Datacollecties van overheidsinstituten:** op de vakgebieden Sociale wetenschappen en Volksgezondheid worden er veel datasets opgezet en beheerd door grote nationale (en soms internationale) onderzoeksinstituten. Het gaat dan vaak om cohortstudies of uitgebreide longitudinale enquêteonderzoeken. Toegang tot deze datasets is bepaald door de regelgeving van de overheid. Voor datasets op het gebied van Volksgezondheid kan dit beperkt zijn vanwege privacy redenen (referentie 22), maar meestal is toegang wel mogelijk: een voorbeeld hiervan is het KNMI, dat haar datasets met meteorologische meetgegevens ter beschikking stelt aan derden.

[8] Andere initiatieven - met name gericht op het delen van IT faciliteiten – zijn o.a. Big Grid (http://www.biggrid.nl/) en het Nederlands Bioinformatica centrum (www.nbic.nl).

4.3 Data-archieven en repositories, gericht op opslag en hergebruik van onderzoekdatasets

Een tweede categorie is een aantal organisatievormen die specifiek gericht zijn op opslag en hergebruik van onderzoekdatasets. Deze organisatievormen hebben de volgende kenmerken gemeen:

- De onderzoekdatasets worden verkregen na overdracht door de dataproducent (meestal een onderzoeker of groep onderzoekers). Bij dit overdracht moment speelt de problematiek van het opstellen van metadata voor de dataset en de documentatie bij de dataset een belangrijke rol.
- De dataproducent houdt in veel gevallen het toegangbeheer in eigen hand: het dataarchief/repository biedt een aantal mogelijkheden aan, variërend van Open Access tot toegang uitsluitend na toestemming van de dataproducent.
- De benodigde menskracht voor ondersteuning van de dataproducenten bij toelevering en opstellen van de metadata en documentatie is aanzienlijk.
- Data-archieven en datarepositories zijn (mogelijk met uitzondering van institutionele datarepositories) gericht op het bewaren van de onderzoekdatasets voor de lange termijn. Dit vereist specifieke IT systemen en een op den duur naar verwachting aanzienlijk inspanning op het gebied van datacuratie.

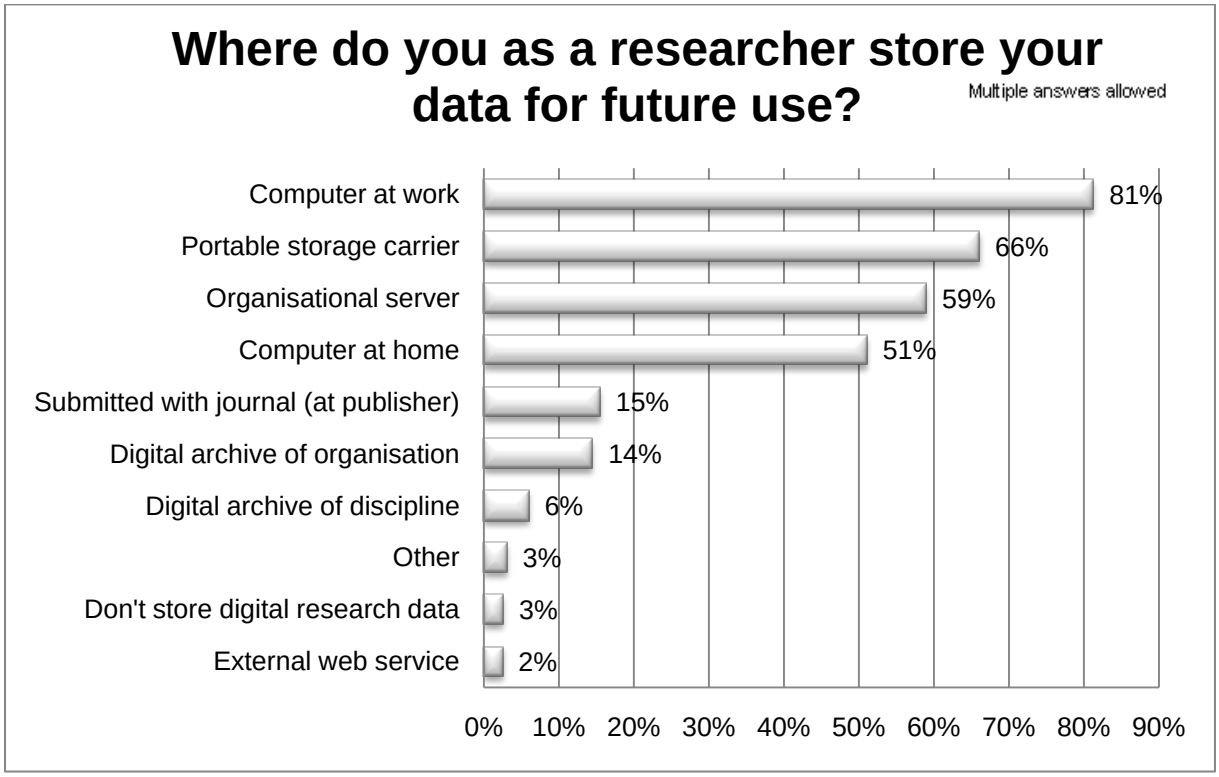
De volgende organisatievormen kunnen binnen deze categorie onderscheiden worden:

- **Discipline-gerichte data archieven, nationaal of internationaal:** in Nederland is in 2005 het inmiddels welbekende DANS opgericht: een nationaal dataarchief voor onderzoekers in de maatschappij- en gedragswetenschappen en de geesteswetenschappen (referentie 11). In deze wetenschappen gaat het om een grote variëteit aan soorten databestanden: catalogi, geannoteerde tekstcorpora , woordenboeken en numerieke en statistische gegevens, spraak-, beeld-en videobestanden en GIS en CAD bestanden.
  Een ander voorbeeld is het Max Planck instituut te Nijmegen, dat een dataarchief op het gebied van linguïstische datasets onderhoudt. Linguïstische wetenschappers wereldwijd kunnen door hen geproduceerde datasets deponeren in dit dataarchief.
  Tenslotte is er nog de categorie (internationale) databanken die zich richten op één soort referentiedatasets. Een voorbeeld hiervan is Genbank. De meeste tijdschriften in het vakgebied *Genomics* vereisen dat de datasets waarover het artikel gaat, zijn gedeponeerd in Genbank (van het National Institute of Health) of vergelijkbare publieke databanken. In dit vakgebied is het de norm om datasets te publiceren en met elkaar te delen. Onderzoekers vermelden door hen gepubliceerde datasets normaal gesproken ook op hun cv.

- **Institutionele datarepositories:** veel universiteiten hebben ook een digitaal repository opgezet, meestal gericht op publicaties. Sommige van deze repositories nemen ook onderzoekdatasets op. Uit een recente Nederlandse inventarisatie (referentie 5) blijkt dat één van de 15 onderzochte (veelal universitaire) repositories onderzoekdatasets opneemt (en drie repositories beelden en vijf video's), uit een Europese inventarisatie (referentie 6) blijkt dat 8,4% van de 178 repositories datasets opnemen (en 15% beelden en 12% video's). in hoofdstuk 5 worden de institutionele datarepositories nader besproken.

4.4 Door onderzoekers zelf georganiseerde opslag en beschikbaarstelling

In het eerder genoemde survey onder wetenschappers (referentie 14) is ook gevraagd hoe men de eigen onderzoeksdata bewaart. De resultaten zijn weergegeven in de grafiek hieronder. Veel datasets blijken bewaard te worden op een computer op het werk of thuis, een draagbaar opslagmedium of een server van de onderzoeksinstelling. Sommigen stellen door hen geproduceerde onderzoekdatasets zelf beschikbaar via de eigen website: Volgens een survey onder Britse wetenschappers (n=176; referentie 2) komt dit vaak voor: ruim 50% van de wetenschappers stelt op deze manier datasets beschikbaar.



4.5 Datacollecties georganiseerd rond wetenschappelijke tijdschriften

Tenslotte worden er onderzoekdatasets ter beschikking gesteld bij digitale wetenschappelijke tijdschriften. Veel uitgevers bieden de mogelijkheid om de datasets bij het betreffende artikel te deponeren, maar vaak met beperkingen met betrekking tot de omvang en het formaat. In de praktijk blijkt het meestal te gaan om kleinere onderzoekdatasets (replicatie datasets) in formats die uitsluitend geschikt zijn voor validatie en niet voor hergebruik. Ook stellen de uitgevers bij monde van STM in de zgn. Brussels Declaration[9] dat onderzoeksdata buiten het copyright van uitgevers vallen en dat het niet binnen het business model past van uitgevers om datasets op grote schaal op te slaan en te archiveren (referentie 23).

Er zijn twee organisatievormen in deze categorie:

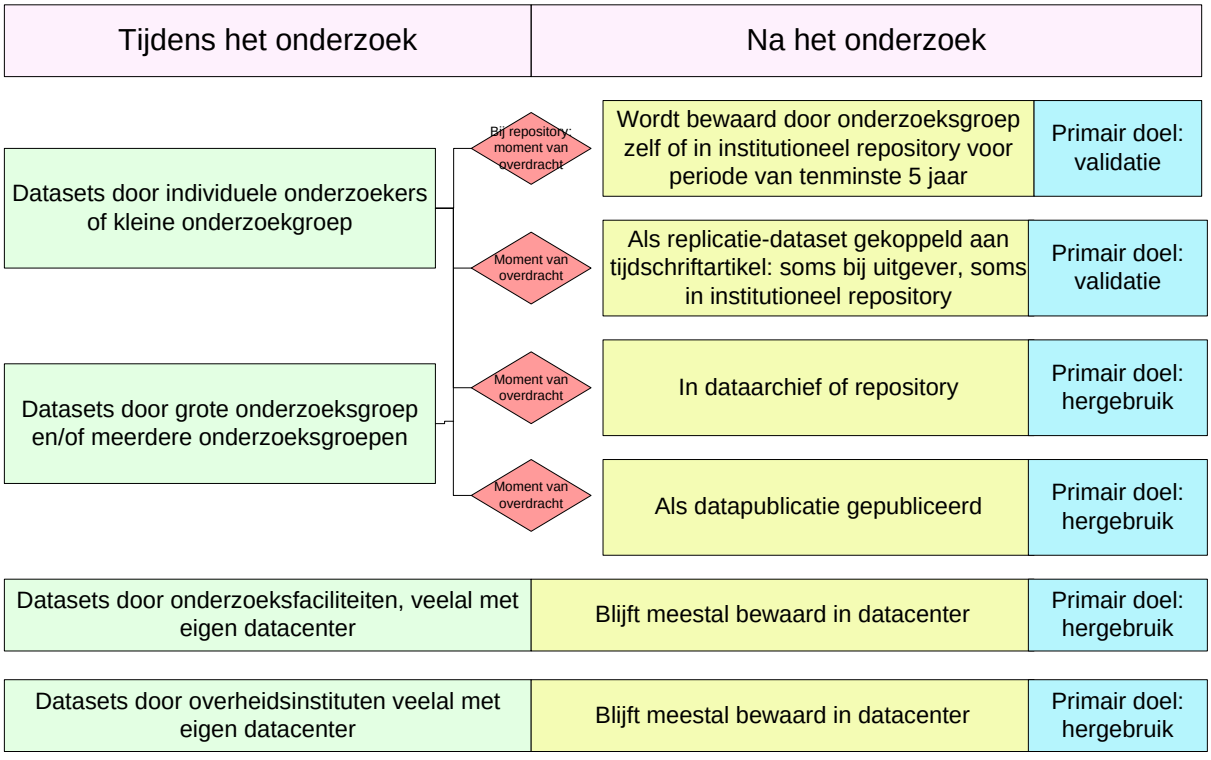
- **Supplementary data bij tijdschriften:** in steeds meer vakgebieden of onderdelen van een vakgebied wordt door de peer-reviewed tijdschrift vereist dat de onderliggende datasets eveneens gepubliceerd worden (zie hierboven). De uitgever kan daarvoor dan een mogelijkheid bieden, maar het is aannemelijk dat uitgevers vooral bij volumineuze datasets of datasets in afwijkende formats in toenemende mate zullen verzoeken om een link of referentie naar een centraal gedeponeerde dataset.
- **Tijdschrift om datasets te publiceren:** een bijzonder geval is dat er een tijdschrift is om datasets zelfstandig te publiceren buiten de context van een onderzoeksartikel: Acta Crystallographica E. Het is een Open Acces tijdschrift waarin datasets als waren het zelf artikelen worden gepubliceerd, inclusief peer review en die ook als zodanig kunnen worden geciteerd[10].

[9] http://www.stm-assoc.org/public_affairs_brussels_declaration.php

[10] Een andere datapublicatie is Earth System Science Data: dit tijdschrift publiceert wel het artikel over de dataset maar archiveert de dataset zelf niet (i.t.t. Acta Crystallographica E)

4.6 Overzicht landschap organisatievormen in relatie met soorten datasets

In de onderstaande figuur wordt het landschap van organisatievormen globaal geschetst in relatie met de indeling van datasets naar dataproducenten. In de figuur wordt nog eens benadrukt dat er voor de datasets van de datacentra van onderzoeksfaciliteiten en overheidsinstituten geen overdracht plaatsvindt. De datasets zijn beschikbaar voor hergebruik vanaf een vroeg stadium na hun creatie. Voor datasets door individuele onderzoekers of kleine onderzoeksgroepen dan wel voor datasets van grotere onderzoeksgroepen en/of meerdere onderzoeksgroepen (ook wel genoemd ‘Small Science’, referentie 26) vindt een dergelijk moment van de overdracht wel in veel gevallen plaats: of aan een institutioneel repository, of aan een supplementary dataset depot bij een wetenschappelijk tijdschrift - in de meeste gevallen gericht op het faciliteren van validatie - of aan een dataarchief/repository of publieke databank gericht op hergebruik.



5. Opties voor institutioneel beleid

5.1 Inleiding

In appendix B zijn de resultaten van de interviews met onderzoeksinstituten weergegeven. Uit deze interviews zijn de volgende opties voor institutioneel beleid af te leiden:

- opstellen van een institutioneel beleidsdocument
- uitvoeren van data audits
- datamanagement: het verzorgen van trainingen en opleidingen op het gebied van datamanagement en voorlichting en advies hierover
- dataopslag faciliteiten
- registratie van onderzoekdatasets
- het implementeren van een institutioneel datarepository

Hieronder worden deze opties voor institutioneel beleid nader toegelicht en besproken.

5.2 Opstellen van een institutioneel beleid

De ervaringen van de Universiteit van Melbourne en de Universiteit van Edinburgh maken duidelijk dat het opstellen en vaststellen van een institutioneel beleid met betrekking tot de dataopslag en beschikbaarstelling cruciaal is voor het verbeteren van de praktijk ervan. Een institutioneel beleidsdocument dient zich vooral te richten op een duidelijke vaststelling van verantwoordelijkheden voor de opslag en beheer van onderzoekdatasets, die gecreëerd worden binnen de instelling, en een beleidslijn wat betreft de beschikbaarstelling ervan. Een dergelijke beleidsdocument op instituutsniveau zal vanzelfsprekend aan dienen te sluiten op landelijke beleidslijnen. In Australië heeft een aanscherping van de landelijke gedragscode direct geleid tot aanscherpingen op instituutsniveau.

5.3 Data audits

De universiteiten van Melbourne en Edinburgh hebben data audits laten uitvoeren, waarin het huidige datamanagement van onderzoekdatasets onder de loep werd genomen. De resultaten van deze data audits in Edinburgh zijn beschreven (referentie 24). Men heeft vijf audits uitgevoerd bij een afdeling voor fysiologie, theologie, economie en sociale historie, astronomie en hersenonderzoek. Er werden belangrijke problemen vastgesteld wat betreft de dataopslag, datamanagement en de toegankelijkheid van de data (zie textbox hieronder). Men vermeldt overigens ook dat er veel weerstand bestond bij de onderzoekers mee te werken aan de data audit. Een dergelijke weerstand wordt ook gesignaleerd in een enquête onder Nederlandse wetenschappers (Waaijers en van der Graaf 2010; referentie 28).

Storage provision:
In most cases there was insufficient storage space available on the servers.
Many datasets (mostly arising from small-scale projects) were stored by researchers themselves in a more or less ad hoc manner on personal external storage devices with little chance of effective retrieval. As a result, these datasets were not managed effectively or made readily-accessible to other researchers. Regular back-up facility was only available for the data that is stored on the school server.
Data value and retention period:
Majority of the participants' data were very valuable and it would be difficult to regenerate the data in case of a loss. In some cases, particularly interviews and surveys, it would be impossible to regenerate the data. The data were generally retained up to or more then 10 years.
Lack of a formal data management plan: All the participants –only with a few exceptions - who involved in the interviews and completed the survey acknowledged that they did not have a formal data management plan. Some research staff mentioned that there is pressure from funding bodies, especially, to ensure that data, once created, is properly managed and stewarded. Staff also indicated that they would prefer the issue of data management was raised at the beginning of the research process rather than at the end of the project.
Lack of guidelines and standardised procedures in creating and storing data: In most cases access to data stored on the server was not straightforward as the data were not catalogued or there was little manual cataloguing. Users were expected to search for the required data themselves, or with guidance from the relevant research staff. Searching for the data was difficult as most of the data was undocumented and there was not a well defined folder structure. Metadata was sparse at best. When such metadata items existed, they were either minimally populated or were the default files generated automatically by the data's host proprietary application (and were hence incomplete).

A clear message that came through was the urgent need to develop greater awareness and understanding of data management within the university as well as guidance on best practice.
uit: Edinburgh Data Audit Implementation Project (referentie 24)

5.4 Datamanagement

Universiteiten met een actief beleid op dit gebied stellen het verbeteren van datamanagement door hun onderzoekers voorop. Een van de respondenten benadrukte dat het een relatief nieuw vakgebied is, waarin onderzoekers niet of nauwelijks in zijn opgeleid. Daarom worden er trainingen aangeboden op het gebied van datamanagement aan senior medewerkers en wordt datamanagement onderdeel gemaakt van opleidingen voor Masterstudenten of promovendi. Daarnaast blijkt er voorlichting nodig te zijn over wat datamanagement precies inhoudt en waarom het belangrijk is. Tenslotte geven meerdere respondenten aan dat onderzoekers in bepaalde gevallen advies en begeleiding nodig hebben en dat deze aangeboden kan worden door experts op dit gebied: de universiteit van Melbourne heeft reeds drie 'data librarians' in dienst en verwacht dit verder uit te breiden met nog eens 2-3 medewerkers.

5.5 Dataopslag faciliteiten

Soms blijken de dataopslagfaciliteiten die de onderzoeksinstelling aanbiedt onvoldoende te zijn voor de onderzoekers. Een respondent waarschuwt met name voor een doorberekening door centrale IT afdelingen voor opslagcapaciteit. Dit kan de onderzoeker om financiële redenen dwingen

tot een suboptimale dataopslag van de onderzoekdatasets. Mede daarom wordt het beleid op het gebied van onderzoeksdata binnen de Universiteit van Melbourne vormgegeven door een nauwe samenwerking tussen de bibliotheek, de IT afdeling en het onderzoeksbureau.

5.6 Dataset registratie

Bij de Universiteit van Melbourne vindt een registratie plaats van onderzoekdatasets. Deze is gekoppeld aan het onderzoek informatiesysteem van de universiteit (vergelijkbaar met het METIS van de Nederlandse universiteiten). Een dergelijke registratie lijkt essentieel: uit hoofdstuk vier van dit rapport is immers gebleken dat een aantal disciplines data-archieven en/of datacentra hebben voor de opslag en terbeschikkingstelling van onderzoekdatasets. Dit is echter voor veel andere disciplines niet het geval: deze onderzoekers zijn dus voor de opslag en beschikbaarstelling van door hen geproduceerde onderzoekdatasets afhankelijk van een institutioneel datarepository.

5.7 Institutioneel datarepository

Rol bibliotheken bij opzetten institutionele datarepositories

Al eerder werd vermeld dat uit inventarisaties onder de Nederlandse en Europese academische repositories blijkt dat een klein percentage ervan al onderzoekdatasets opneemt en grotere percentages beelden en video-opnames, vaak ook gerelateerd aan onderzoek. De Amerikaanse Association of Research Libraries stelt in een rapport (referentie 15) dat bibliotheken van onderzoeksinstellingen een belangrijke rol kunnen spelen bij het opzetten en onderhouden van repositories voor verschillende soorten van onderzoeksoutput. Zij voorzien een inter-operabel ‘patchwork of repositories operating at national, disciplinary and institutional levels’. In een ander rapport van de ARL (referentie 18) wordt eveneens een rol voorzien voor bibliotheken van onderzoeksinstellingen in E-science en datacuratie. Ook volgens een redactioneel commentaar van het tijdschrift Nature (referentie 1) zijn universiteitsbibliotheken voor de hand liggende kandidaten om de benodigde institutionele repositories voor datasets te organiseren.

Minimale omvang datarepository

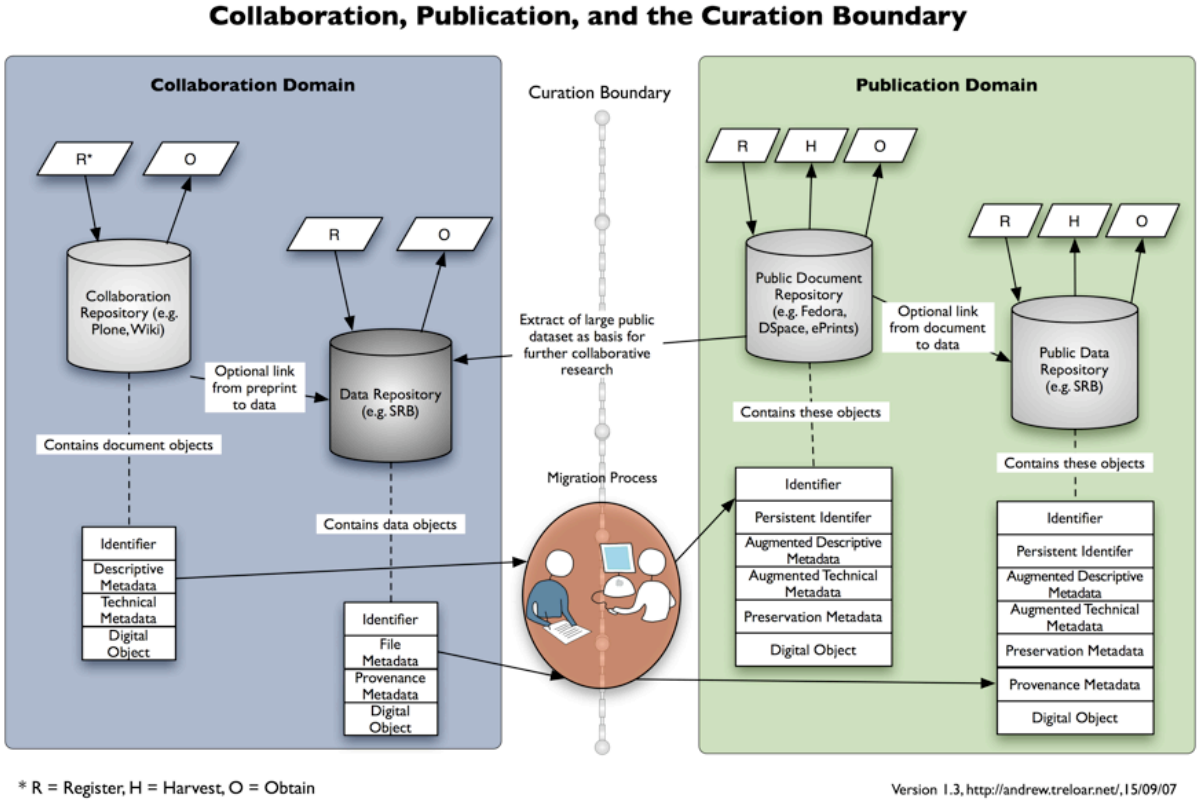
De Blue Ribbon Taskforce (referentie 13) stelt dat gecentraliseerde diensten met name efficiënt zijn wanneer de datasets hoge niveaus van expertise in datacuratie en archivering vereisen. Hoe groot zou een organisatie rond een dataarchief of repository voor datasets minimaal moeten zijn? In de UKRDS feasibility study (referentie 7) wordt op basis van een uitgebreide analyse gesteld dat een minimale bemensing van een datarepository 2,5 FTE bedraagt. Op grond daarvan concluderen zij dat een gecentraliseerde service financieel voordeliger is dan dat elke

universiteit - met name de kleinere universiteiten - een eigen datacentrum zou opzetten.

Leerervaring van de universiteit van Monash

Treloar e.a. (referentie 8) beschrijven de ervaringen met dataopslag van de Universiteit van Monash. Zij stellen in de eerste plaats dat zij streven naar één repository voor de universiteit, zowel voor publicaties als voor data, maar dat zij teruggekomen zijn op dit idee: de verschillen tussen de objecten, de verschillen in beheer en toegang tussen de publicaties en datasets blijken hiervoor in de praktijk te groot.
In de figuur hieronder wordt de nu geïmplementeerde infrastructuur weergegeven:

- er worden samenwerkingsomgevingen voor text en data ingericht
- men ziet een scheidslijn tussen het samenwerkingsdomein en het publicatie domein.
- in het publicatie domein ziet men twee repositories: een repository voor documenten en een repository voor data.



Is een institutioneel datarepository nodig?

Uit de case reports blijkt dat de toelevering van datasets aan institutionele repositories in de praktijk vaak moeizaam verloopt: dit wordt gerapporteerd door zowel de Universiteit van Edinburgh als van Toronto. Ook de eisen aan een minimale omvang van 2,5 fte om een dergelijk datarepository op te zetten en te onderhouden en de disciplinegerichtheid

van de benodigde expertise van dataformats en metadata zullen een rol spelen bij beleidsbeslissingen van een onderzoeksinstelling hierover. Een belangrijke rol lijkt in een instelling daarom weggelegd voor de eerdergenoemde dataset registratie, waardoor het aantal onderzoekdatasets binnen de instelling inzichtelijk wordt. Mogelijk is een gezamenlijk datarepository met andere universiteiten organisatorisch een betere oplossing. De inrichting van het 3TUDatacentrum is een goed voorbeeld hiervan.
Volgens de resultaten van deze studie is het echter duidelijk dat drie soorten datasets redenen zijn voor onderzoeksinstellingen om een eigen institutioneel datarepository in te richten:

- replicatiedatasets, die aan de bewaartermijn voor de Gedragscode van Wetenschapsbeoefening moeten voldoen
- replicatiedatasets, die wetenschappers ter beschikking stellen als gevolg van verplichtingen ten opzichte van de wetenschappelijke tijdschriften.
- die onderzoekdatasets, waarvan het belangrijk is om deze ter beschikking te stellen en langdurig te bewaren met het oog op hergebruik, maar waarvoor geen disciplinegerichte oplossing (in de vorm van een dataarchief of databank) bestaat.

6. Visie op data infrastructuur

6.1 Inleiding

Wat zijn de stakeholders met betrekking tot onderzoekdatasets en wat zijn hun respectievelijke rollen en verantwoordelijkheden? In een door velen aangehaald rapport (referentie 9) worden de volgende stakeholders genoemd:

- onderzoekers in hun rol van dataproducenten
- onderzoekers in hun rol van dataconsumenten
- het instituut waaraan de wetenschapper verbonden is
- het datacentrum
- de onderzoeksfinancier
- de uitgever
- de search & discovery service (aggregator).

In dit hoofdstuk wordt de samenhang van de verschillende taken en verantwoordelijkheden stakeholders besproken met het oog op een nationaal en internationaal vorm te geven data infrastructuur.

6.2 Dataproducenten

Wat zijn de beweegredenen voor dataproducerende wetenschappers om datasets ter beschikking te stellen? In de interviews met de data-archieven en datacentra zijn hierover de volgende aspecten naar voren gekomen:

- **Overtuiging dat het een goede zaak is:** een belangrijk deel van de wetenschappers hangen de Open Access gedachte aan. Bovendien blijkt het in de cultuur van een aantal disciplines reeds volstrekt normaal te zijn om datasets te delen en te hergebruiken.
- **Regels en/of kwaliteitsnormen:**
    - **Onderzoeksfinanciering:** een belangrijke beweegreden voor wetenschappers tot deponeren en beschikbaar stellen is wanneer onderzoeksfinanciers dit verplicht stellen.
    - **Gedragscodes en kwaliteitsnormen:** Bij archeologie is in de afgelopen jaren veel aanwas geweest bij DANS vanwege de verplichting tot het deponeren via de kwaliteitsnormen Nederlandse archeologie. In Australië gaat een sterk verplichtende werking uit van de daar ingestelde Code of Conduct voor wetenschappers, mede omdat deze door de onderzoeksfinanciers aldaar als uitgangspunt wordt gehanteerd.
    - **Lage drempel:** Meerdere geïnterviewde universiteiten (Universiteit van Edinburgh; Universiteit van Melbourne) streven ernaar om in de bestaande onderzoekregistratie procedures voor

wetenschappers voor het registreren cq. deponeren van onderzoekdatasets op te nemen. Voor de Nederlandse situatie betekent dit dat universitaire onderzoekers, die reeds jaarlijks hun publicatiegegevens in Metis zetten en hun publicaties in het institutionele publicatie repository deponeren tevens de mogelijkheid geboden zou worden om in één moeite door door hen geproduceerde datasets te registreren en/of te deponeren in een daartoe ingericht datarepository.

- **Hergebruik en erkenning:**
  - **Co-auteurschap**: in een aantal gevallen wordt bepleit dat hergebruikende wetenschappers de data producerende wetenschapper uitnodigen voor een co-auteurschap op de uit het hergebruik voortkomende publicaties.
  - **Citaties:** Piwowar en anderen (referentie 21) tonen aan dat het ter beschikking stellen van onderliggende datasets de citatiescores van de publicaties ten goede komt. Niettemin wordt er allerwegen gepleit om het direct citeren van datasets mogelijk te maken. Een van de mogelijke hulpmiddelen daarbij is DataCite: dit is een registratiesysteem dat het mogelijk maakt voor onderzoekers om datasets te laten registreren en persistente identifiers (in dit geval DOI) daaraan toe te kennen, zodat onderzoekdatasets als onafhankelijke, citeerbare wetenschappelijke objecten kunnen worden behandeld (referentie 16). Het MPI dataarchief hanteert een ander systeem – Handle - terwijl anderen weer het URN systeem hanteren. Volgens een ter zake deskundige geïnterviewde in dit onderzoek is het geen probleem dat er meerdere registratiesysteem zijn. Deze zouden later gemakkelijk aan elkaar gekoppeld kunnen worden via mapping, zolang de toegekende identifiers maar uniek zijn.

Als redenen om datasets niet te delen worden de volgende categorieën argumenten gebruikt:

- **Eerst zelf publiceren:** veel wetenschappers stellen dat het om hun onderzoeksidee gaat en dat hun werk erin zit. Daarom willen zij en/of een onderzoeksgroep in veel gevallen er eerst zelf over publiceren. In de praktijk hanteren om deze redenen veel datacenters daarom een embargoperiode. Voor een bepaalde periode - meestal een jaar - behoudt de onderzoeksgroep exclusieve toegang tot de dataset. Data-archieven laat in veel gevallen de toegang door de deponerende onderzoekers zelf beheren: zij kunnen dan zelf een embargoperiode instellen.

- **Privacy/ regelgeving:** wetenschappers zijn vaak huiverig om door hen geproduceerde datasets ter beschikkingstelling in verband met de privacy van (gegevens van) deelnemers aan het onderzoek. Ook kan er beperkende regelgeving wat betreft toegang zijn: veel van de eerder genoemde archeologische gegevens in DANS zijn uitsluitend voor archeologen beschikbaar en niet voor een breed publiek. Tenslotte kan het zijn dat in de dataset weer data van anderen zijn opgenomen, waardoor er onduidelijkheid is bij de dataproducerende wetenschapper over wat er (juridisch) mogelijk is m.b.t. beschikbaarstelling van de dataset.
- **Verkeerde interpretaties:** Een aantal wetenschappers voert het argument aan dat door hen geproduceerde datasets zo ingewikkeld zijn, dat er een grote kans is dat andere onderzoekers deze verkeerd begrijpen en/of verkeerd interpreteren. Ook wordt er aangevoerd dat bijvoorbeeld commerciële partijen, wier belangen geschaad worden door de resultaten van het onderzoek, naar ongerechtigheden in de dataset zouden kunnen zoeken en hiermee in de publiciteit zouden kunnen treden ten einde twijfel over de resultaten van het onderzoek te zaaien. Dit is het ‘doubt is our product’ argument, waarmee de tegenpartij het leven van de betreffende wetenschappers zuur zou kunnen maken .
- **Documenteren kost teveel tijd**: Dit is een praktisch bezwaar, dat in veel gevallen naar voren wordt gebracht. Hier zijn twee aspecten aan: Het documenteren van het onderzoek zelf heeft soms gebrekkig plaatsgevonden en is achteraf een lastige en tijdrovende klus. Ten tweede wordt het aanmaken van metadata door veel wetenschappers als lastig ervaren: door de geïnterviewde data-archieven wordt overigens gesteld dat dit een bezigheid van enkele uren is.
- **Angst om op fouten betrapt te worden**: Dit argument (en de angst dat anderen op betere, andere en/of conflicterende conclusies uitkomen) wordt vanzelfsprekend zelden genoemd, maar speelt volgens een geïnterviewde wel degelijk een rol bij een aantal wetenschappers.

Het is van belang vast te stellen dat de bovengenoemde bezwaren tegen het toegankelijk maken van onderzoekdatasets voor veel dataproducerende wetenschappers zwaar wegen met als gevolg dat velen door hen geproduceerde datasets niet Open Access beschikbaar stellen. Uit het grote, recente survey onder wetenschappers (referentie 14) blijkt dat slechts 25% dit wel doet.

6.3 Dataconsumenten

Wat zijn de redenen voor wetenschappers om datasets te hergebruiken? Een geïnterviewde onderscheidende volgende redenen om datasets te raadplegen:

- **Voor een specifiek onderzoek/secundaire analyse:** Een bepaalde gegevensreeks bestaat al en kan de onderzoeker gebruiken om zijn/haar onderzoeksvraag te beantwoorden. Onder secundaire analyse wordt verstaan het opnieuw analyseren van gegevens met een andere vraagstelling dan waarvoor ze oorspronkelijk zijn verzameld.
- **Bij het opstellen van een onderzoeksvoorstel**: De opsteller kijkt of zijn/haar voorstel nieuw is en/of deze gegevens niet al bestaan. Dit leidt vaak tot fine-tuning van het voorstel. Eén respondent pleitte ervoor om dit soort 'nieuwheidsonderzoek' verplicht te stellen bij onderzoeksaanvragen bij NWO.
- **Wetenschapsgebieden die gegevens van veel verschillende disciplines gebruiken:** een aantal wetenschapsgebieden gebruiken gegevens van veel verschillende disciplines of doen aan grote eenheden onderzoek, veelal op hoge aggregatieniveaus onderzoek doen. Als voorbeelden kunnen gelden de bioinformatica (gebruikmakend van genenbanken), de klimatologie en de epidemiologie (referentie 22). Deze vakgebieden kunnen niet of nauwelijks functioneren zonder toegang tot datasets.
- **Model input:** Ook voor computermodellen is het noodzakelijk om datasets te raadplegen voor de inputgegevens voor het (simulatie-) model.

In een in 2009 verschenen proefschrift aan de Universiteit van Michigan (referentie 25) worden de stappen voor een dataconsument in de sociale wetenschappen beschreven:

1. Zoeken en selecteren van relevante datasets en het verkrijgen van de data
2. Controleren van de data: de onderzoeker checkt of de variabelen in de documentatie overeenkomen met de variabelen in de dataset en kijkt naar de wijze van steekproefname en hoe non-response en ontbrekende data zijn behandeld.
3. Data manipulatie: de onderzoeker construeert nieuwe variabelen gebaseerd op de bestaande variabelen in de dataset, hercodeert de data en/of voegt datasets samen.
4. Data-analyse: de hergebruikende onderzoeker voert data-analyse technieken uit op de dataset.

6.3 Overige stakeholders

De taken en verantwoordelijkheden van alle stakeholders staan in de tabel hieronder weergegeven[11]. Duidelijk is dat in deze taakverdeling een netwerk wordt voorzien van institutionele repositories, die datasets voor de kortere termijn bewaren en datacentra die hiertoe geselecteerde datasets voor de langere termijn bewaren.

Rol	Rechten	Verantwoordelijkheden
Wetenschapper: creatie en gebruik van de data	• Op het eerste gebruik. • Op citatie / erkenning • Op erkenning van Intellectual Property Rights[12]. • Ontvangen van datamanagement training en advies.	• Datamanagement voor de duur van het project . • Voldoen aan standaarden voor Good Practice. • Voldoen aan de voorwaarden en vereisten van onderzoeksfinanciers en het eigen instituut en het respecteren van Intellectual Property Rights van derden. • Geschikt maken van datasets voor hergebruik door derden.
Gebruiker: gebruik van dataset van derde partij	• Om datasets te hergebruiken (niet exclusieve licentie). • Toegang krijgen tot metadata van goede kwaliteit met het oog op de bruikbaarheid .	• Voldoen aan de voorwaarden van de licenties. • Erkenning van de data producenten en het datacentrum . • Datamanagement van de afgeleide datasets.
Instituut: opslag en beschikbaarstelling data	• Aangeboden krijgen van een kopie van de dataset .	• Opstellen van intern beleid voor datamanagement. • Datamanagement op de kortere termijn. • Voldoen aan de standaarden voor Good Practice. • Geven van training and advies aan wetenschappers. • Promotie van de diensten van het repository.
Datacentrum: opslag en beschikbaarstelling data	• Aangeboden krijgen van een kopie van de dataset. • Selecteren van datasets met waarde op de lange termijn	• Datamanagement voor de lange termijn • Voldoen aan de standaarden voor Good Practice. • Geven van training over deponeren. • Promotie van de diensten van het repository. • Waarborgen van de rechten van de data producenten. • Aanbieden van tools om hergebruik te faciliteren.

[11] Vertaald door opsteller rapport

[12] Dit is niet onomstreden. Zie ook: De juridische status van ruwe data: wegwijzer voor de onderzoekspraktijk, SURFdirect, juli 2009 en het RGO rapport (referentie 22).

Rol	Rechten	Verantwoordelijkheden
Financier: bepalen van beleid	• Implementatie van datamanagement beleid. • Eisen stellen aan gefinancierde onderzoekers om aan hun verplichtingen te voldoen .	• Oog houden op overheidsbeleid en behoeften van de verschillende stakeholders. • Deelnemen in coördinatie van strategieën. • Ontwikkel beleid in samenspraak met stakeholders. • Deelnemen in beleidscoördinatie en gezamenlijke planning. • Monitoren en handhaven van het datamanagement beleid. • Mogelijk maken van post-project long-term datamanagement. • Lobby voor datacuratie • Ondersteun opbouw capaciteit en competenties van data curators.
Uitgever: bewaken van integriteit van wetenschappelijke documenten	• Vereisen dat gegevens beschikbaar zijn om de publicatie te ondersteunen. • Vereisen dat datasets zijn gedeponeerd in long-term repositories voorafgaande aan publicatie	• Betrekken van stakeholders bij de ontwikkeling van publicatie standaarden. • Verzorgen link naar datasets bij publicaties. • Monitoren en handhaven van standaarden.
Aggregator: onderhouden search & discovery service[13]	• 'Federated' discovery en toegang mogelijk maken • aangeboden krijgen van metadata van de datasets in de datacentra	• Betrekken van stakeholders bij de opbouw van een federated metadatarepository • Registratie van de toeleverende data producenten • Promotie van de discovery service • Faciliteren toegang tot datasets

[13] Wat betreft de search en discovery service is het van belang te vermelden dat in Nederland dit reeds operationeel is: NARCIS geeft toegang tot Open Access onderzoekspublicaties, proefschriften, beschrijvingen van onderzoekprojecten, onderzoekorganisaties en onderzoekers en onderzoekdatasets: begin mei 2010 zijn er bijna 15.000 datasets doorzoekbaar (het overgrote deel afkomstig van DANS).

6.4 Taken op nationaal niveau

De Amerikaanse National Science Foundation heeft in 2008 een programma (referentie 12) uitgeschreven van meer dan honderd miljoen dollar om een nieuw soort organisaties te creëren die (vrij vertaald en ingekort):

(1) Bibliotheek- en archiefwetenschappen, IT infrastructuur, computer- en informatiewetenschappen en disciplinegerichte expertise weten te integreren ten einde een betrouwbare digitale bewaarfunctie te ontwikkelen
(2) Voortdurend anticiperen op en zich aanpassen aan de veranderingen in technologie en in de behoeften en verwachtingen van de gebruikers.

Ook in andere landen is men druk bezig met het opzetten van datacenters, onder andere ANDS in Australië en de UK Research Data service (UKRDS) in Groot-Brittannië.

Voor de Australische Nationale Data Service (ANDS) worden in het businessplan (referentie 10) de volgende taken genoemd:

- het ontwikkelen van beleidskaders
- het opzetten en leveren van diensten, zoals search & discovery services, registers voor collecties van datasets en dergelijke
- het stimuleren en op gang brengen van het bewaren en beschikbaar stellen van onderzoekdatasets (‘seeding the commons’)
- capaciteitsopbouw: opzetten van trainingprogramma's voor datamanagers en certificering daarvan.

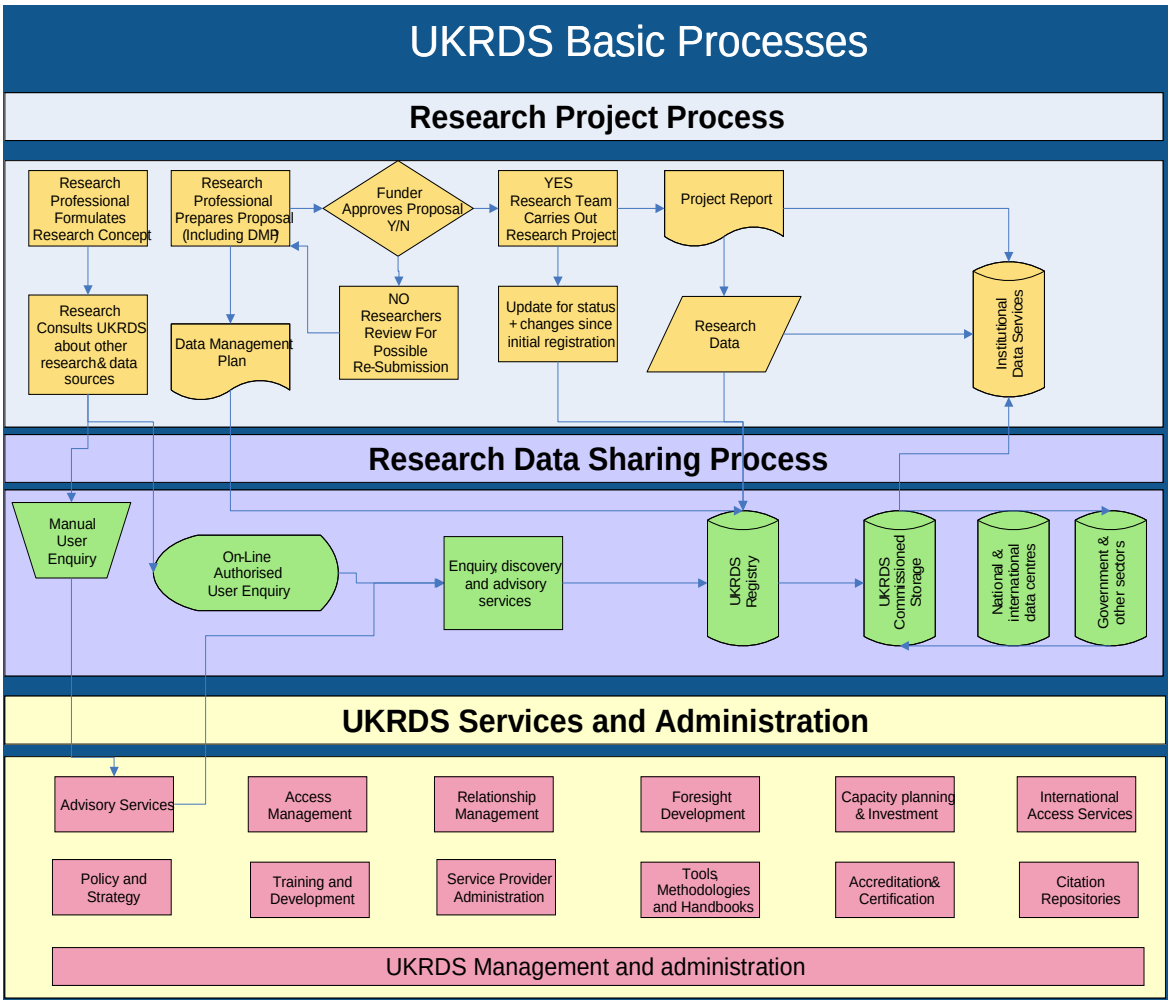
De ANDS heeft dus geen data-archief en neemt vooral de rol van regisseur en aggregator op zich.

Voor het UKRDS is een en ander in detail uitgewerkt in onderstaand schema. Er worden drie lagen in het schema onderscheiden:

1. Een laag die de start en de voortgang van het individuele onderzoeksproject beschrijft. Men beschrijft daarin o.a. dat bij de start van een onderzoeksproject wordt nagegaan of er andere, relevante datasets aanwezig zijn. Tevens wordt een datamanagementplan opgesteld, dat wordt geregistreerd bij het UKRDS.
2. Een laag die het delen van de onderzoeksdata beschrijft: de uit het onderzoeksproject voortkomende dataset wordt eveneens geregistreerd bij het UKRDS en opgeslagen in het institutionele datarepository. De datasets kunnen (ook) opgeslagen worden in een van de door de UKRDS op disciplinaire basis georganiseerde datacentra voor de langere termijn.
3. Een laag waarin de nationale regietaken van het UKRDS worden beschreven voor het regelen van toegang tot de datasets (technisch en juridisch), zoals:

- het organiseren van internationale toegang tot datasets en naar datasets
- het anticiperen op nieuwe ontwikkelingen op gebied van techniek en verwachtingen van gebruikers
- het ontwikkelen en vastleggen van gereedschappen, methodes en handboeken voor datacuratie
- het bieden van faciliteiten voor repositories en uitgevers om het citeren van datasets te faciliteren.

In de UK ziet men dus duidelijk een rol voor het instituut op het gebied van datamanagement en op het gebied van registreren en bewaren van datasets. De UKRDS vervult daarin met name een regierol.



7. Conclusies en aanbevelingen voor een strategie m.b.t. de opslag en toegankelijkheid van onderzoekdatasets op nationaal en institutioneel niveau

7.1 Samenvatting en conclusies

- **Wetenschappelijk onderzoek steeds meer data-driven:** Het volume van de onderzoekdatasets stijgt sterk terwijl het delen van datasets een vlucht neemt en tot nieuwe vormen van wetenschapsbeoefening leidt.
- **Bewaren met het oog op validatie en hergebruik:** Er zijn op dit moment diverse regels en redenen om onderzoekdatasets voor langere tijd te bewaren. De Gedragscode Wetenschapsbeoefening schrijft een bewaartermijn van vijf jaar voor van de onderzoeksdata na publicatie met het oog op validatie. Steeds meer tijdschriften eisen de terbeschikkingstelling van een zogenaamde replicatiedataset, eveneens vooral met het oog op validatie. Daarnaast zijn er wetenschappelijke en niet-wetenschappelijke redenen om onderzoekdatasets te bewaren en ter beschikking te stellen met het oog op hergebruik. Enkele onderzoeksfinanciers hebben daarom reeds verplichtingen ingesteld om onderzoekdatasets zodanig ter beschikking te stellen dat zij ook voor hergebruik geschikt zijn.
- **Organisatievormen voor opslag en terbeschikkingstelling van onderzoeksdata:** Het landschap voor wat betreft de organisatievormen voor opslag en terbeschikkingstelling van onderzoekdatasets vertoont een grote verscheidenheid, is versnipperd en bevat lacunes. Er zijn vier categorieën te onderscheiden:
    - **Datacentra:** Meestal behorend bij een onderzoeksfaciliteiten en disciplinair en/of thematisch gericht. In het rapport beschreven voorbeelden hiervan zijn de radiotelescoop LOFAR, de aan elkaar gekoppelde datacentra van de instituten van de Nationale Oceanografische Data Commissie en de datacentra van het KNMI. Deze datacentra bevatten de onderzoeksdata vanaf de creatie en in veel gevallen is datasharing in een vroeg stadium normaal en goed geregeld, maar vooral beperkt tot de ‘eigen community’. In principe worden de datasets voor de langere termijn bewaard.
    - **Data-archieven:** Data-archieven zijn eveneens in de meeste gevallen gericht op een of meerdere disciplines en zijn gefocust op een wijde beschikbaarstelling en langdurige bewaartermijn. In het rapport beschreven voorbeelden hiervan zijn de data-archieven van DANS, het Max Planck Instituut en het 3TU Datacentrum. Onderzoekers in hun rol van dataproducent dragen na afloop van hun onderzoek de onderzoekdatasets over. In veel gevallen blijft het toegangsbeheer in handen van de dataproducent. Het goed beschrijven van onderzoekdatasets ten

behoeve van het dataarchief (metadata, onderliggende documentatie) blijkt vaak een bottleneck voor de data producent. De activiteiten van het dataarchief ter ondersteuning van de dataproducenten en dataconsumenten en de datacuratie blijken veel menskracht te vereisen evenals een specifieke, op de disciplinegerichte expertise.
    - **Door onderzoekers zelf georganiseerd:** Onderzoekers organiseren in veel gevallen zelf de opslag en beschikbaarstelling van door hen geproduceerde onderzoekdatasets. Deze worden dan bijvoorbeeld intern opgeslagen of zijn via de website van de onderzoeksgroep beschikbaar. Wijdere beschikbaarstelling en een langdurige bewaartermijn zijn in deze situaties dan meestal niet gegarandeerd.
    - **Supplementary data:** Een toenemend aantal wetenschappelijke tijdschriftredacties eisen dat de onderliggende datasets bij een artikel (replicatiedatasets) Open Access ter beschikking worden gesteld. Tijdschriften bieden daar zelf mogelijkheden toe (maar vaak met beperkingen qua omvang en qua dataformats), de meeste uitgevers hebben (nog) geen duidelijke services ontwikkeld om een grote toestroom van datasets van aanzienlijk volume en grote gevarieerdheid aan te kunnen. Het is niet waarschijnlijk dat uitgevers een dergelijke rol op zich zouden kunnen of willen nemen.
- **Ervaringen met institutioneel beleid:** De ervaringen van enkele onderzoeksinstellingen met het opzetten en uitvoeren van een institutioneel beleid op dit gebied werden bevraagd door middel van interviews. Enkele belangrijke opties voor een institutioneel beleid zijn:
    - **Beleidsdocument:** Opstellen van een institutioneel beleidsdocument, waarin de verantwoordelijkheden voor opslag en beheer van datasets geregeld wordt en een beleidslijn ten aanzien van beschikbaarstelling wordt vastgesteld.
    - **Data audits:** Uitvoeren van data audits, waarin de huidige datamanagement praktijken onder de loep worden genomen
    - **Datamanagement:** Het verzorgen van trainingen en opleidingen op het gebied van datamanagement en voorlichting en advies hierover.
    - **Dataopslagfaciliteiten:** Zorgdragen voor adequate dataopslagfaciliteiten voor onderzoekdatasets tijdens de datacreatie en het gebruik ervan.
    - **Registratie van onderzoekdatasets:** Opzetten van een registratiesysteem voor onderzoekdatasets gekoppeld aan het onderzoekinformatiesysteem van de instelling
    - **Institutioneel datarepository:** Een institutioneel datarepository is functioneel voor replicatiedatasets, die onderzoekers dienen te deponeren met het oog validatie als gevolg van de Gedragscode Wetenschapsbeoefening en/of als gevolg van verplichtingen van wetenschappelijke tijdschriften en voor die datasets, waarvan het belangrijk is om deze te langdurig

bewaren en ter beschikking te stellen met het oog op hergebruik, maar waarvoor geen disciplinegericht dataarchief bestaat.

- **Stakeholders:** Er zijn zeven belangrijke stakeholders voor wat betreft onderzoeksdata te onderscheiden, ieder met een verschillende rol en met verschillende taken en verantwoordelijkheden:
    - **Dataproducenten en dataconsumenten:** De wetenschappers die datasets produceren en deze ter beschikkingstellen en de wetenschappers die deze datasets hergebruiken. Voor de dataproducerende wetenschappers spelen er belangrijke drempels om door hen geproduceerde datasets ter beschikking te stellen, zowel om inhoudelijke redenen als om praktische redenen. Bij het laatste gaat het vooral om het opstellen van adequate metadata en het leveren van adequate documentatie bij de dataset.
    - **Onderzoeksinstellingen:** De onderzoeksinstellingen zijn verantwoordelijk voor het aanbieden van adequate faciliteiten voor datamanagement van hun onderzoekers, waaronder de mogelijkheid voor de opslag en beschikbaarstelling van onderzoekdatasets.
    - **Het dataarchief of datacentrum:** Dataarchief of datacentra hebben de taak om onderzoekdatasets voor de langere termijn te bewaren en zo wijd mogelijk ter beschikking te stellen.
    - **De onderzoeksfinancier:** Onderzoeksfinanciers hebben een belangrijke stem in het bepalen van het beleid ten aanzien van onderzoeksdata. Een aantal onderzoeksfinanciers vereisen Open Access beschikbaarstelling van datasets van onderzoeksprojecten die door hen gefinancierd worden.
    - **De uitgevers:** Uitgevers van wetenschappelijke tijdschriften hebben als taak om te zorgen dat er een koppeling plaatsvindt tussen de wetenschappelijke artikelen aan de ene kant en de onderliggende datasets aan de andere kant.
    - **Aggregators:** Een rol voor de aggregators is om een search & discovery dienst te onderhouden.
- **Visie op data-infrastructuur:** In meerdere landen worden er acties ondernomen om gebruikmakend van de huidige infrastructuur deze aan te vullen om een dekkend en inter-operabel netwerk van datacentra, data-archieven en institutionele datarepositories in te richten. Daarbij wordt een taakverdeling voorzien tussen onderzoeksinstellingen aan de ene kant, die registratie van datasets en de meer korte termijn beschikbaarstelling van datasets op zich nemen en een netwerk van data-archieven en datacentra, nationaal en internationaal, meestal langs disciplinaire lijnen georganiseerd, waar de lange termijn archivering en beschikbaarstelling van belangrijke onderzoekdatasets wordt georganiseerd.

Op basis van de hiervoor gepresenteerde conclusies uit dit onderzoek zijn de volgende aanbevelingen voor een strategie op nationaal en institutioneel niveau opgesteld.

7.2 Aanbevelingen voor een aanpak op nationaal niveau

Aanpassen Gedragscode Wetenschapsbeoefening

De Gedragscode Wetenschapsbeoefening is in 2004 vastgesteld door de VSNU. Hieronder wordt de meest relevante passage weergegeven. Deze gedragscode wijkt af van de Australian Code for the Responsible Conduct of Research van 2007 [14]. Deze is vastgesteld door de Australische universiteiten en de onderzoeksfinanciers in Australië. In een apart hoofdstuk wordt het management van onderzoeksdata behandeld en worden de verantwoordelijkheden van de instituten en van de onderzoekers afzonderlijk onderscheiden en beschreven. Van elk instituut wordt een instituutbeleid op dit gebied vereist. Gesteld wordt dat instituten veilige opslag moeten garanderen van data, dat er een beleid moet zijn over wie de eigenaar is en wie het bewaart. Als minimum bewaartijd wordt vijf jaar aangehouden, maar voor bepaalde onderzoeken zou dit langer moeten zijn (expliciet genoemd wordt een vijftienjarige bewaartermijn voor klinische trials en permanente bewaartermijnen voor o.a. erfgoed). Onder de verantwoordelijkheden voor de onderzoekers wordt gesteld dat deze onderzoeksdata beschikbaar dienen te stellen voor gebruik door andere onderzoekers, tenzij dit niet mogelijk is vanwege de ethiek, privacy of vertrouwelijkheid.

Op basis van de Australische ervaringen wordt dan ook aanbevolen om de Nederlandse Gedragscode Wetenschapsbeoefening uit te breiden wat betreft de genoemde aspecten van onderzoeksdata en te laten vaststellen door alle betrokkenen bij het onderzoek in Nederland (dus naast de VSNU ook bijvoorbeeld NWO, KNAW en SURFfoundation).

III. Controleerbaarheid

Principe
Gepresenteerde informatie is controleerbaar. Als onderzoeksresultaten openbaar worden gemaakt, blijkt duidelijk waar de gegevens en de conclusies op zijn gebaseerd, waaraan ze zijn ontleend en waar ze te controleren zijn.
Uitwerking
III.1 Onderzoek moet gerepliceerd kunnen worden om de juistheid ervan te testen. De keuze van de onderzoeksvraag, de opzet van het onderzoek, de keuze van de gehanteerde methode en verwijzing naar geraadpleegde bronnen is nauwkeurig gedocumenteerd.
III.2 De kwaliteit van data verzameling, data-invoer, dataopslag en dataverwerking wordt goed bewaakt. Goede verslaglegging van alle stappen en controle op de uitvoering is noodzakelijk (labjournaals, voortgangsrapportages, documentatie van afspraken en beslissingen enz.).
III.3 De bewaartermijn van ruwe onderzoeksgegevens is minimaal 5 jaar. Deze gegevens worden op aanvraag ter beschikking gesteld aan andere wetenschapsbeoefenaren.
III.4 Ruwe onderzoeksgegevens worden zodanig gearchiveerd dat deze te allen tijde met een minimum aan tijd en handelen kunnen worden geraadpleegd.
Uit: Gedragscode voor Wetenschapsbeoefening, VSNU 2004

[14] http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/r39.pdf

Rol NWO en andere onderzoeksfinanciers

Aanbevolen wordt dat NWO, ZonMw en andere onderzoeksfinanciers in Nederland criteria op het gebied van datamanagement opnemen in de beoordeling van onderzoeksvoorstellen. NWO heeft al een dergelijke regel voor alfa en gammawetenschappen wat betreft opname van datasets in DANS maar zou ook een datamanagement paragraaf kunnen vereisen voor onderzoeksvoorstellen op andere vakgebieden. Ook kan NWO mogelijk het door een respondent genoemde nieuwheidsonderzoek (bestaat er niet al een dergelijke dataset?) als eis bij het indienen van een onderzoeksvoorstel invoeren. Tenslotte kan NWO ook een dergelijke aanpak in Europees verband – EUROHORCs - entameren.

Regierol voor een discipline-onafhankelijk DANS

Uit dit onderzoek is naar voren gekomen dat er een lappendeken is wat betreft nationale en internationale data-archieven en datacentra, al dan niet volgens disciplines en thema's georganiseerd. Ook is duidelijk dat deze lappendeken gaten bevat, waardoor belangrijke onderzoekdatasets niet duurzaam worden bewaard en/of ter beschikking worden gesteld. Tevens is duidelijk dat er geen eenduidige toegang is en dat datasets nog niet of moeilijk geciteerd kunnen worden. Daarom is het belangrijk dat er een organisatie binnen Nederland een regierol op zich neemt om deze zaken - in samenwerking met de bestaande datacentra en met het oog op nieuwe ontwikkelingen - verder te ontwikkelen en te organiseren.
DANS is een logische kandidaat om deze regierol op landelijk niveau te vervullen. Blijkens een op moment van schrijven nog niet gepubliceerde visienota van DANS wil men ook graag een dergelijke regierol oppakken. DANS vervult momenteel een data-archiveringsfunctie voor de alfa en gammawetenschappen. Wat betreft de andere disciplines zou DANS dan de regierol op zich kunnen nemen en waar nodig ook een data-archiveringsrol kunnen vervullen.
Andere onderdelen van de regierol voor DANS omvatten:

- Het streven naar een gecentraliseerd zoeksysteem. In de visie van DANS zou dit een geïntegreerd zoeksysteem kunnen zijn met publicaties en onderzoeksgegevens, zoals nu al deels bestaat in NARCIS.
- Zoveel mogelijk uniformering van de opzet van metadata, licenties voor hergebruik en dergelijke voor de verschillende partijen die hierbij betrokken zijn. Een en ander staat hiernaast uitgewerkt in het kader.

Discipline-onafhankelijke activiteiten regierol DANS:

- *Best practices* voor archivering van en toegang tot data (Datakeurmerk, standaarden voor o.a. metadata);
- Gedragscodes voor gebruik van gegevens;
- Beheren van registers van dataverzamelingen;
- Duurzame langetermijnopslag van data (zoals *persistent identifiers*, back-up faciliteit) en het delen van opslagsystemen;
- Koppelen van datasets aan elkaar, aan publicaties en aan onderzoeksinformatie;
- Hulp bij het delen van data, het realiseren van collaboratories en (methoden voor) het verzamelen van gegevens;
- Onderzoek en ontwikkeling op het gebied van oplossingen voor duurzame toegankelijkheid (*data curation*);
- Aanbieden van juridische expertise (auteursrecht, privacy, licenties).

(Naar een discipline-onafhankelijk DANS, visienota

7.3 Aanbevelingen voor een aanpak op institutioneel niveau

Ook voor een instituut is het lastig om in de lappendeken van de bestaande en in ontwikkeling zijnde faciliteiten op het gebied van onderzoekdatasets een eigen rol te af te bakenen. Uit dit onderzoek komt echter duidelijk naar voren dat een dergelijke rol gewenst is en zelfs als noodzakelijk kan worden aangeduid. Een aanpak op institutioneel niveau dient langs de volgende lijnen ontwikkeld te worden:

- **Beleid:** Vaststellen van een institutioneel beleid op het gebied van datamanagement en datasharing. Daarin dienen in elk geval de verantwoordelijkheden voor opslag en beheer van de datasets geregeld te worden en hoelang data bewaard dienen te blijven.
- **Data audit:** De situatie binnen de instelling wat betreft wetenschappelijke datasets en het datamanagement ervan en de behoeften van onderzoekers kunnen in kaart gebracht worden door middel van een data audit, waarbij mogelijk de methode gevolgd kan worden, zoals ontwikkeld voor het Verenigd Koninkrijk en waar nodig aangepast aan de Nederlandse situatie. Het voordeel van een dergelijke data audit is dat de voor de instelling specifieke lacunes in datamanagement worden vastgesteld, waarop de rest van het beleid kan worden gericht. Het nadeel is dat een dergelijke data audit in veel gevallen op weerstand bij de betrokken wetenschappers zal stuiten.
- **Datamanagement:** In dit rapport is vastgesteld dat een onderzoeksinstelling een speciale verantwoordelijkheid heeft voor het datamanagement van de ‘eigen’ onderzoekdatasets. Daarom wordt aanbevolen dat een onderzoeksinstelling trainingen op het gebied van datamanagement aanbiedt aan haar huidige onderzoekers en zorgt dat datamanagement als vak worden opgenomen in het curriculum van de Masterstudenten en/of promovendi. Tevens blijkt er vaak behoefte bij wetenschappers te bestaan aan voorlichting over datamanagement en persoonlijke ondersteuning en advies hierover.
- **Dataopslagfaciliteiten:** De instelling dient adequate dataopslag faciliteiten aan te bieden aan haar onderzoekers. Eventuele lacunes hierin kunnen aan het licht komen bij de eerdergenoemde data audits.
- **Registratie van datasets:** Aanbevolen wordt om de datasets binnen de instelling te laten registreren door de onderzoekers. Een dergelijk data register dient meerdere doelen: datasets kunnen geteld worden als onderzoeksoutput in het jaarverslag van de instelling, er is zicht op wie welke datasets beheert en waar heeft opgeslagen en het kan dienen als input voor een landelijke zoekmachine voor datasets. Op dit moment dienen universitaire wetenschappers hun publicaties te registreren (METIS) en worden gevraagd de full text van hun publicaties te deponeren in de daarvoor bestemde publicatie repositories. Aanbevolen wordt om naar analogie van de universiteiten

van Edinburgh en Melbourne de registratieprocedure van datasets en (eventueel) het deponeren in het institutionele datarepository in dezelfde procedure op te nemen.

- **Datarepository:** Het instituut dient de mogelijkheid aan te bieden voor het opslaan en beschikbaar stellen van datasets, wanneer de discipline zelf geen faciliteiten daarvoor heeft. Ook dient een dergelijke datarepository voor het deponeren van zogenaamde replicatie datasets als gevolg van verplichtingen voortvloeiend uit de Gedragscode Wetenschapsbeoefening en/of uit verplichtingen richting wetenschappelijke tijdschriften. Uit een eerder aangehaald onderzoek blijkt dat een dergelijk datarepository nogal wat menskracht vereist (2,5 FTE): mogelijk dat het voor kleinere instellingen de moeite loont om gezamenlijk een dergelijk datarepository op te zetten. Het is overigens de vraag of een dergelijke institutioneel datarepository ook toegerust moet worden voor langdurige, duurzame opslag gezien de hiervoor noodzakelijke investeringen in personeel en IT-systemen.
- **Samenwerking bibliotheek, IT en onderzoeksinformatie:** Gezien de raakvlakken met zowel de bibliotheek als IT en de afdeling betrokken bij onderzoeksinformatie wordt aanbevolen een aanpak op dit gebied door een samenwerkingsverband met deze drie afdelingen van het instituut op te zetten.

7.4 Naar Open Access toegang voor onderzoekdatasets

Uit dit onderzoek is naar voren gekomen dat er naast lacunes in de huidige infrastructuur voor de opslag en beschikbaarstelling van onderzoekdatasets bij de dataproducerende wetenschappers nog veelal hoge drempels zijn van zowel inhoudelijke als praktische aard om door hen geproduceerde datasets Open Access ter beschikking te stellen. Eventueel in te zetten beleid van SURFfoundation en andere onderzoeksorganisaties dient zich er dan ook op te richten om datasets, die in de academische wereld geproduceerd worden door individuele onderzoekers of (combinaties van) onderzoeksgroepen beter op te slaan en beschikbaar te stellen met het oog op hergebruik. Hiervoor lijken globaal twee - elkaar niet uitsluitende - beleidscenario's mogelijk te zijn: verleiden of verplichten.

Verleiden

Een beleidscenario dat gericht is op het verleiden van onderzoekers om door hen geproduceerde datasets ter beschikking te stellen zal als eerste stap een bewustwording onder onderzoekers van de voordelen hiervan inhouden. Onderdelen van een dergelijk beleid kunnen zijn: (1) opname in een gedragscode (zoals de Gedragscode voor Wetenschapsbeoefening) (2) data audits, waarbij gebreken in het datamanagement worden blootgelegd en (3) trainingen op het gebied van datamanagement, waarbij de onderzoekers worden ondersteund en begeleid om hun datamanagement te verbeteren.
Een tweede stap van een beleid om onderzoekers te verleiden zal beschikbaarstelling zijn met behulp van toegangbeheer dat in handen van de onderzoekers gelegd zal worden. Dit is het beleid van veel data-archieven op dit moment. De onderzoeker bepaalt zelf de toegankelijkheid van de dataset. In veel gevallen zal dit niet een Open Access toegang betreffen, maar een geclausuleerde toegang.
Een derde stap van een dergelijk beleid zal erop gericht zijn om de datasets zoveel mogelijk Open Access toegankelijk te krijgen en de onderzoekers daartoe te stimuleren. Daarbij zal het van belang zijn om de publicatie van datasets voor het ‘academic record’ mee te laten tellen: voorbeelden hiervan zijn initiatieven om datasets citeerbaar te maken (en dus mee te laten tellen in de citatie indexen) en de genoemde peer-reviewed datapublicatie op het gebied van kristallografie[15].

Verplichten

Het andere beleidscenario richt zich op verplichtingen ten aanzien van Open Access beschikbaarstelling van datasets. Verplichtingen hiertoe kunnen afkomstig zijn van onderzoeksfinanciers, van wetenschappelijke tijdschriften of vastgelegd zijn in een gedragscode. Er zijn gevallen bekend waarin de verplichting ook wordt afgedwongen: voor de

[15] Er is de opsteller nog één andere datapublicatie bekend: het Earth System Science Data journal. Deze datapublicaties zijn vooralsnog dus een zeldzaamheid.

ontvangers van financiering van een onderzoekvoorstel door de Britse Economic and Social Research Council is Open Access beschikbaarstelling van de resulterende dataset via het data-archief UKDA verplicht en de laatste 10 % van de subsidie wordt pas overgemaakt nadat de dataset in het dataarchief gedeponeerd is.

8. Informatiebronnen

1. Data's shameful neglect (editorial), Nature, vol. 461, 7261, p 145
2. Research data preservation and access: the views of researchers; Beagrie N., Beagrie R. and Rowlands I, Ariadne, July 2009
3. Stewardship of digital research data: a framework of principles and guidelines, Research Information Network, January 2008
4. To Share or Not to Share: publication and quality assurance of research data outputs; Research Information Network, 2008
5. Nederlandse academische repositories, SURFshare Nulmeting, februari 2010
6. The European Repository Landscape 2008, Amsterdam University Press
7. The UK Research Data Service feasibility study, report and recommendations to HEFCE; December 2008
8. Treloar A., D. Groenewegen, C. Harboe-Ree, The Data curation continuum, D-Lib Magazine, 2007, vol 13, 9/10
9. Dealing with Data: Roles, Rights, Responsibilities and Relationships, Liz Lyon, UKOLN, 2007
10. Australian National Data Service (ANDS) Interim Business Plan, 2008/9
11. Advies van de Taskforce Data Archiving & Networked Services (DANS) aan de bestuurder van KNAW en NWO, december 2004 (te downloaden van de website website: www.nwo.nl)
12. Sustainable Digital Data Preservation and Access Network Partners (DataNet), NSF programma, 2008
13. Sustainable Economics for a Digital Planet: ensuring long-term access to digital information, February 2010, Blue Ribbon Task Force on Sustainable Digital Preservation and Access
14. PARSE Insight: Insight into digital preservation of research output in Europe, December 2009 (www.parse-insight.eu)
15. The Research Library's Role in Digital Repository Services, January 2009, Association of Research Libraries
16. DataCite – a global registration agency for research data, Jan Brase, [bijdrage ILDS conference 2009), zie ook www.datacite.org]
17. e-IRG, Report on Datamanagement, Datamanagement Task Force, 2009
18. Transformational Times, ARL 2009
19. OSU Libraris and Research Dataset Curation: a beginning, Research and Innovative Series Report 4, 2009
20. Waardevolle Data en Diensten, Rombouts J. e.a. 2009
21. Piwowar H.A., R.S. Day, D.B. Fridsma, Sharing detailed research data is associated with increased citation rate; PLoS ONE 2(3); e308. doi:10.1371/journal.pone.0000308
22. Van gegevens verzekerd, kennis over de volksgezondheid in Nederland in de toekomst; Raad voor Gezondheidsonderzoek, 2008
23. The STM report: an overview of scientific and scholarly publishing; Mark Ware and Michael Mabe, September 2009

24. Edinburgh Data Audit Implementation Project, Final report, C. Ekmekcioglu, R. Rice, January 2009
25. Perceived documentation quality of social science data; Jinfang Niu, dissertation University of Michigan; 2009
26. Data dimensions: disciplinary differences in research data sharing, reuse and long term viability; a comparative review based on sixteen case studies; DDC Scarp Synthesis report, Key Perspectives Ltd; 2010
27. Selection of Research Data, SURFfoundation, a report by DANS and 3TU Data Centre, in press
28. Waaijers L., Van der Graaf M.; Over de kwaliteit van onderzoeksdata; SURFfoundation rapport in voorbereiding.

Appendix A Case reports datacentra en data-archieven

A.1 De data-archieven DANS, MPI en 3TU Datacentrum

DANS

- Omvang: op dit moment zijn er ongeveer 16.000 datasets gedeponeerd in DANS. Er komen jaarlijks tussen 2000 tot 3000 datasets bij.
- Organisatie: bestaat uit drie secties: archief, beschikbaarstelling en software development (zie ook 11). DANS bestaat uit ruim 20 FTE.
- Deponeren en toegang: het idee vanaf het begin is geweest om de verantwoordelijkheid te leggen bij de onderzoeker. De deponerende onderzoeker bepaalt ook hoe de toegang is geregeld. DANS bevordert Open Access zo veel mogelijk, maar houdt rekening met de realiteit, dat niet iedere onderzoeker dat wil. Er is een optie voor beschermde toegang, waarbij de onderzoeker zelf dan ook de Acces-authority is. De derde optie is 'other access': de toegang verloopt dan niet via DANS maar via een ander dataarchief/ datarepository .
- Hergebruik: Het hergebruik is aanzienlijk gestegen in de laatste jaren. In 2006 werden ruim 1000 datasets gedownload, in 2009 is dit opgelopen naar meer dan 8000 downloads. Gebruikers moeten zich registreren en wanneer zij een dataset gedownload hebben wordt dit in het systeem weergegeven zodat transparant is wie een dataset hergebruikt.
- Diensten:
    - Data projecten: dit zijn projecten waarin medewerkers van DANS samen met onderzoekers het datamanagement van grote projecten uitvoeren.
    - Visualisatie van data: de huidige opzet van het archief is degelijk, maar nogal saai. De nieuwe generatie gebruikers wil data kunnen visualiseren. DANS probeert hier op in te spelen met de ontwikkeling van visualisatie software.

3TUDatacentrum

- Omvang: op dit moment zijn er van ongeveer 15 projecten bijna 3500 datasets in het dataarchief opgenomen. Men verwacht aan het eind van het jaar ongeveer 15 terabyte aan datasets
- Organisatie: Het 3TU Datacentrum heeft officieel nog steeds de status van een project. Er zijn ongeveer 10 FTE bij het project betrokken. Een verlenging van het project is onlangs aangevraagd, het is de bedoeling om e.e.a. na enige tijd over te laten gaan in de staande organisaties van de drie technische universiteiten. Een belangrijk deel van het backoffice wordt uitgevoerd door de Bibliotheek van de TU Delft, de frontoffice taken (met name acquisitie) worden ook door de

bibliotheken van de TU Eindhoven en de Universiteit Twente uitgevoerd.

- Deponeren en toegang: op dit moment worden de data in overleg met de onderzoekers gedeponeerd in het archief. Binnenkort heeft men een self-deposit mogelijkheid. In principe is de toegang Open Access, de onderzoeker kan wel een embargoperiode instellen. Er is nog geen mogelijkheid om de dataset slechts voor een beperkte groep gebruikers open te stellen. Binnenkort zullen gebruikers van het dataarchief zich eerst moeten registreren. De licentieovereenkomsten zijn identiek aan die van DANS.
- Diensten:
    - Zoeken: via de website kan men zoeken en browsen. Ook worden er resource maps gepubliceerd van de metadata ten behoeve van andere instellingen.
    - Training: het doel is om een cursus datamanagement op te nemen in de programma's voor promovendi.
    - Datacite: dit is een registratiesysteem voor datasets waardoor deze citeerbaar worden (16). Aan een dataset wordt er een DOI (digital object identifier) toegekend. Dit gebeurt voor alle datasets die worden opgenomen.

Max Planck Instituut

- Omvang: op dit moment bevat het archief 50 terabyte, dit zijn ongeveer 400.000 files.
- Organisatie:
    - Er zijn twee ‘archive managers’. De taken van deze ‘archive managers’ zijn:
        - het begeleiden van de gebruikers van het data archief bij het zoeken en hergebruiken van de datasets
        - het begeleiden van de deponerende wetenschappers bij het deponeren van door hen geproduceerde datasets
        - het inventariseren van de gebruikerswensen en deze doorgeven aan de software ontwikkelaars.
    - Daarnaast is er een groep van circa 10 software ontwikkelaars. Een belangrijk deel van deze software ontwikkelaars worden gefinancierd uit gerelateerde projecten.
- Deponeren en toegang:
    - Het softwarepakket is het zelf ontwikkelde Language Archive Management Upload System. Dit systeem maakt het mogelijk dat een wetenschapper zelf zijn datasets in het archief kan deponeren. Er worden hieraan vooraf enkele eisen gesteld: (1) het data archief vraagt inzage in de dataset vooraf om te checken of de dataset inhoudelijk in het archief past (2) de deponerende wetenschapper dient ook metadata aan te leveren en (3) een beperkt aantal data formats zijn toegestaan.

    - Elke deponerende wetenschapper zelf kan bepalen hoe zijn datasets toegankelijk worden gemaakt. De metadata zijn wel voor iedereen toegankelijk. Er zijn vier mogelijkheden om de toegang tot de datasets regelen:
        - de dataset is vrij toegankelijk voor iedereen
        - de dataset is helemaal niet toegankelijk
        - de dataset is toegankelijk voor een bepaalde groep
        - de dataset is toegankelijk voor iedereen die een bepaalde licentie tekent.
    - Omdat de deponerende wetenschapper zelf de toegang kunnen regelen, is het gebruik van het systeem lastig in te schatten. De respondent schat dat er tussen de 500 en 2000 wetenschappers zijn die datasets in het archief deponeren.
- Diensten:
    - Enrichment software: dit zijn software pakketten, waarmee men de data kan verrijken. Het gaat bijvoorbeeld om een annotatie tool (Elan) en een syntactische analyse tool (Sympathy) en Lexus voor het opzetten van digitale woordenboeken.
    - Satellietarchieven: een dataarchief wordt in de omgeving van de vreemde taal zelf opgezet zodat de lokale taalgemeenschap er optimaal gebruik van kan maken. Het dataarchief in Nijmegen bewaart dan wel de metadata van deze datasets, maar niet de datasets zelf.
    - Visualisatie: er worden pogingen ondernomen om de datasets op een andere manier te presenteren: zo is er al een Google Earth overlay en worden er ook andere manier overwogen om datasets te presenteren.
    - Cursus: een twee jaarlijkse training over hoe de software werkt en hoe men datasets kan archiveren.
- Citeren van datasets in het dataarchief: men heeft een manier hiervoor ontwikkeld, maar deze is nog niet geheel uitontwikkeld. Inmiddels citeert men zelf in de eigen publicaties al wel datasets op deze manier, maar men heeft nog geen zicht in hoeverre wetenschappers dit systeem reeds gebruiken.

A.2 De datacentra van de NOCD, LOFAR en het KNMI

NODC

- Organisatie: Het NODC (Nationale Oceanografische Data Commissie) is een netwerk van op dit moment zeven deelnemende organisaties, waaronder het Koninklijk Nederlands Instituut voor Zeeonderzoek (NIOZ). Het grote voordeel van deze opzet is dat er geen apart instituut nodig is: het netwerk maakt gebruik van de faciliteiten en expertise van de deelnemende instituten. Het NODC heeft een beleid ontwikkeld om gegevens onderling en naar buiten toe te kunnen uitwisselen. Het NODC is gebaseerd op een convenant dat de deelnemende instituten hebben ondertekend. Dit is in 2009 aangepast aan de ontwikkelingen en opnieuw vastgesteld. De kosten van het NODC zijn minimaal: de deelnemende instellingen betalen jaarlijks enkele duizenden euro's, waarvan de website wordt onderhouden en wat andere zaken worden geregeld. In Europa is er Seadatanet opgericht om te zorgen dat de verzamelde meetgegevens van de verschillende Europese zeeën onderling uitwisselbaar zijn. Het NODC verzorgt het Nederlandse onderdeel van dit Seadatanet en sluit aan op de afspraken binnen dit netwerk.
- Datacentrum NIOZ: Het NIOZ beheert een schip dat een nationaal vaarprogramma uitvoert voor wetenschappers. Wetenschappers kunnen voorstellen indienen, zij krijgen dan vaartijd toegewezen waarin bepaalde metingen worden uitgevoerd. Deze data worden door de datamanagementgroep van het NIOZ beheerd en ter beschikking gesteld aan de aanvragende wetenschappers.
- Toegang tot datasets: Enkele jaren geleden is een project gestart dat ertoe heeft geleid dat de datasets per datapunt raadpleegbaar zijn. Een datapunt is een reeks meetgegevens van een bepaalde geografisch plek - de plaats waarop de meetgegevens zijn verzameld. De verschillende datasets met de verschillende gegevens zijn nu zodanig opgezet, dat een gebruiker alle meetgegevens van de verschillende datasets van een bepaalde geografische plek in één keer naar zich toe kan halen. In veel gevallen zijn de datasets direct na een eerste kwaliteitscontrole beschikbaar, in andere gevallen krijgen de aanvragende onderzoekers een periode van exclusieve toegang van meestal een jaar voordat deze worden vrijgegeven. Men dient zich overigens eerst te registreren voordat men een dataset kan downloaden.

- Hergebruik: verschillende databanken van het NIOZ met gegevens worden dagelijks tientallen tot honderden keren geraadpleegd.
- Duurzame toegankelijkheid: In principe bewaart men de datasets voor de eeuwigheid. Binnen het NODC heeft men afgesproken het te melden wanneer men data zou gaan verwijderen. Maar in de praktijk verwijderen datacentra niet gauw datasets. Wat betreft de data formats ziet de respondent voorlopig weinig problemen wat betreft

duurzame toegankelijkheid. De problemen doen zich vooral voor bij de dragers: de tapes, cd's en dvd's zijn na tientallen jaren niet meer leesbaar. Zij hanteren een systeem waarbij om de 5-7 jaar de data worden overgezet op een andere drager.

LOFAR

- Omvang: LOFAR is een radiotelescoopsysteem met meer dan 40 stations en is zeer recentelijk van start gegaan. Elk station bestaat uit een veld met honderden antennes, per veld worden de signalen gecombineerd. De gecombineerde signalen gaan naar het rekencentrum in Groningen. Daar worden de data realtime verwerkt. Er vindt een correctieslag plaats, data van de verschillende velden worden bij elkaar opgeteld en dergelijke. De supercomputer slaat vervolgens de verwerkte gegevens op: er is ruimte voor tijdelijke opslag van twee petabyte. Wanneer LOFAR observeert, komen er data vrij in de orde van grootte van een tot twee petabytes per week. Dit zijn hoeveelheden data die slechts enkele rekencentra aankunnen. Naast de tijdelijke opslag is er een rekencluster, waar de eerste bewerkingsslag en analyse plaatsvindt. Een belangrijk onderdeel van de analyse is of de kwaliteit van de data goed genoeg is. Het gaat erom te kijken of er geen storingen zijn (bijvoorbeeld dat iemand een mobiele telefoon in de buurt van een antenneveld heeft gebruikt) of dat alle apparaten voldoende hebben gefunctioneerd. Ook vindt er datareductie plaats. Er blijven dan gegevens over van enkele terabytes per uur: deze data worden dan beschikbaar gesteld via het LOFAR archief aan de astronoom die een en ander heeft aangevraagd. Het is een gedistribueerd archief, met locaties in Groningen, Amsterdam en Duitsland. Voor deze langere termijn opslag zijn er petabytes aan opslagruimte beschikbaar.
- Organisatie: Er is een afdeling support, die astronomen kan ondersteunen bij het verwerken van de data. Deze afdeling beslaat zes medewerkers. Daarnaast is een softwareontwikkelingsgroep van circa acht personen. Tenslotte is er nog een onderhoudsgroep van circa zes man.
- Hergebruik: de respondent stelt dat de mate van hergebruik mede afhangt van hoe laagdrempelig de toegang tot de datasets is georganiseerd. Bij de radiotelescoop Westerbork worden de datasets wel hergebruikt, maar niet erg veel vanwege de wijze van beschikbaarstelling. Bij andere astronomische waarnemingsinstrumenten is wel bekend dat er vrij veel hergebruik plaatsvindt bij een laagdrempelige toegang. Hergebruikers dienen zich overigens te registreren, dit is vooral bedoeld om te zorgen dat er niet onnodig grote datasets gedownload worden. Daarnaast zijn er regels voor een hergebruikende astronoom: er dient in elk geval een referentie naar LOFAR in de publicatie te zijn opgenomen. Het is niet verplicht om ook de oorspronkelijke onderzoekers, die de waarneming

hebben aangevraagd, te citeren, maar in de praktijk gebeurt dat wel veel, temeer dat deze onderzoekers het meest van die dataset weten.

- Digitale duurzaamheid: het uitgangspunt is om de datasets zo lang mogelijk te bewaren, dit vindt plaats op basis van 'best effort'. Voor Westerbork zijn alle datasets bewaard vanaf 1970, het startjaar van Westerbork. Voor LOFAR zal men waarschijnlijk onderscheid moeten maken: er zullen datasets zijn die te grootschalig zijn om op langere termijn te blijven bewaren.
- Diensten: er wordt periodiek een cursus over LOFAR data gegeven. Daarnaast is er de genoemde supportgroep, die ondersteuning biedt aan wetenschappers.

KNMI

- Organisatie: Op dit moment heeft het KNMI een aantal datacentra draaien: Climate Explorer met klimaat gegevens, TEMIS met gegevens van satellieten etc. Men is nu van start gegaan met het opzetten van een overkoepelend datacentrum, waarbij de wetenschappers van het KNMI door hen geproduceerde data kunnen publiceren en het datacentrum de technische afhandeling (archivering, beschikbaarstelling etc.) verzorgt. Het datacentrum is in principe uitsluitend bedoeld voor datasets geproduceerd door wetenschappers van het KNMI en data geproduceerd in de operationele product ketens van het KNMI. Er is onderscheid tussen de voorkant: de portal/website en de webservices, waar ook andere partijen diensten kunnen aanbieden, en de achterkant. Die laatste bestaat uit data archivering en processing. Men verwacht dat er aan de achterkant circa één FTE menskracht op continue basis nodig zal zijn, en aan de voorkant circa 0,5 FTE. Op dit moment is men het datacentrum aan het oprichten. Dit is een vrij lichte organisatie, express gekozen omdat men zo efficiënt mogelijk wil werken. In het licht daarvan is het ook de bedoeling dat de wetenschappers zelf door hen geproduceerde data in het systeem kunnen plaatsen. Men probeert zoveel mogelijk te automatiseren: na het plaatsen van de dataset zouden de metadata gedeeltelijk automatisch gegenereerd moeten worden. Een ander deel van de metadata zal door de wetenschappers moeten worden toegevoegd. Men hanteert bij de metadata verschillende standaarden, onder andere de ISO standaarden, maar ook de verschillende standaarden van de verschillende wetenschapsgebieden. Het wordt dus een soort self-service archief, maar wel met de mogelijkheid van ondersteuning. Dit jaar is men bezig met het opstellen van de functionele specificaties, volgend jaar begint men met bouwen en middels halfjaarlijkse releases verwacht men in 2013 het datacentrum geheel gereed te hebben.
- Deponeren en toegang: Wat betreft specifieke meetcampagnes willen de wetenschappers van het KNMI vaak eerst de gegevens valideren en erover publiceren. Meestal worden de datasets na een half of een heel jaar ter beschikking gesteld. De respondent stelt geen datasets binnen het KNMI te kennen, die niet worden gepubliceerd. Met andere

woorden, er is een cultuur van delen van datasets. Dit is ook gewoon binnen de wereld van de meteorologie. Overigens, de door het KNMI verrichtte dagelijkse waarnemingen zijn vrijwel direct op Internet beschikbaar. In dit verband wordt ook de INSPIRE richtlijn genoemd: dit is een Europese richtlijn, die stelt dat geodatasets geharmoniseerd moeten worden en ter beschikking gesteld moeten worden. Hiervoor wordt een Europese infrastructuur aangelegd (http://www.inspire-geoportal.eu/), een portal, diensten en catalogi etc. Het Nationale Georegister (http://www.nationaalgeoregister.nl) is er een onderdeel van.

- Hergebruik: de huidige datacentra binnen het KNMI worden goed gebruikt. Dit zijn specifieke communities, erg gespecialiseerd, met gebruikers, die precies weten wat ze willen. Enkele klimaatdatasets zijn gepubliceerd in speciale datapublicaties en zijn citeerbaar door middel van een DOI. Wanneer wetenschappers andere datasets gebruiken, wordt er wel gevraagd om een acknowledgment.
- Duurzame toegankelijkheid: de meetgegevens binnen Nederland, zoals verzameld door de KNMI, moeten voor altijd bewaard worden. Dit is wettelijk vastgelegd. Satellietmetingen worden echter vaak vanuit projecten gefinancierd. Men wil nu een slag maken om ook deze metingen voor de langere termijn te bewaren en ter beschikking te stellen. Het meeste werk is dat veel gegevens worden geproduceerd met een bepaald algoritme. Na een tijdje wordt er een beter algoritme ontwikkeld en dient de dataset gereprocessed te worden. Voor een langere periode zal het mogelijk ook wel voorkomen dat het dataformaat gemigreerd moet worden naar een ander dataformaat. Overigens, in de klimatologie en in de meteorologie veranderen de dataformaten niet snel volgens de respondent.
- Diensten: Men ziet dat er een toenemende behoefte is aan web services: services om opgehaalde datasets te presenteren. Als voorbeeld geeft de respondent een samenwerking met het RIVM: zij halen bepaalde gegevens automatisch op en gebruiken die direct in hun eigen workflow. Zo wordt als het ware direct ‘ingeprikt’ met de juiste data in het eigen werkproces.

Appendix B: Case reports instituten

Universiteit van Melbourne

De Universiteit van Melbourne heeft in 1996 richtlijnen opgesteld voor datamanagement. Geleidelijk aan werd duidelijk dat door de lage status van die richtlijnen e.e.a. onvoldoende serieus werden genomen. Daarom werden de richtlijnen omgezet tot officieel beleid van de universiteit (2005). In een survey en audit, die beide daarna plaatsvonden, werd vastgesteld dat er behoefte was binnen de universiteit aan o.a. dataopslag, datacuratie en duurzame opslag na afloop van het onderzoeksproject. Bovendien werden gevallen aangetoond van tekortschietende datamanagement praktijken. Men concludeerde daaruit dat er duidelijk moet worden vastgelegd wie waarvoor verantwoordelijk is en dat een en ander dient te worden ondersteund met faciliteiten. Binnen de Universiteit van Melbourne heeft men dit opgezet als een samenwerkingsproject van het research office, de bibliotheek en het rekencentrum. Inmiddels wordt het beleid van de universiteit wat betreft datamanagement herzien en zal binnenkort een gewijzigde versie worden goedgekeurd. Een van de redenen van deze wijziging is dat in Australië de 'Australian Code for the Responsible Conduct of Research' is opgesteld door de gezamenlijke universiteiten en onderzoeksfinanciers. Er zijn een aantal duidelijke eisen gesteld aan datamanagement in deze Code of Conduct, waaraan de Universiteit van Melbourne moet voldoen om zijn financiering te blijven behouden. Tevens wordt steeds duidelijker dat wetenschappelijke tijdschriften in toenemende mate vereisen dat de onderliggende datasets bij de artikelen beschikbaar worden gesteld.

De respondent stelt dat een universiteitsbeleid in de praktijk weinig effectief zal blijken te zijn wanneer er niet tegelijkertijd faciliteiten aan de onderzoekers worden aangeboden. Zij bieden de volgende diensten aan:

- Dataopslag: dataopslag wordt nu aangeboden tot vijf jaar na publicatie van het artikel. Men realiseert zich dat dit een vrij korte tijd is, waarschijnlijk moet het aantal datasets dat in aanmerking komt voor langere termijn opslag worden beperkt.
- Datamanagement adviesdiensten: de respondent stelt dat datamanagement inmiddels een aparte discipline aan het worden is. Op dit moment heeft men drie 'data librarians' in dienst, men streeft er naar om dit verder uit te breiden met 2-3 medewerkers (de Universiteit van Melbourne heeft 12.000 onderzoekers).
- Data registratie: men heeft een data registratie opgezet, die geïncorporeerd is in het onderzoeksinformatiesysteem van de universiteit (vergelijkbaar met het METIS systeem bij de Nederlandse universiteiten).
- Meer voorlichting over de voordelen van een goed datamanagement: in dit verband noemt de respondent een onderzoeksgroep die hun datamanagement sterk verbeterd hebben. Het blijkt nu dat zij steeds

meer publiceren, omdat het voor hen duidelijker is welke gegevens zij hebben en hoe zij deze kunnen (her)-gebruiken. De respondent stelt dat een ‘compliance only’ benadering minder effectief zal zijn bij de onderzoekers: het benadrukken van de voordelen voor het eigen onderzoek daarentegen zal meer aandacht trekken.

De respondent geeft aan dat de Universiteit van Melbourne op dit moment vooral is gefocust op het verbeteren van het datamanagement en minder op het delen van datasets. Men ziet dit als een tweede stap in het proces.

Universiteit van Toronto

Bij de Universiteit van Toronto is reeds geruime tijd geleden een ‘data library’ opgezet. Men acquireert meer dan 900 datasets van partijen buiten de universiteit en stelt deze beschikbaar aan gebruikers binnen de Universiteit van Toronto. Het gaat om onderzoekers, maar ook om studenten die in het kader van cursussen met dergelijke grote datasets leren omgaan. De data library biedt daarbij ondersteuning, naast persoonlijke ondersteuning verzorgt men onderdelen van cursussen. De data library bestaat uit een fulltime medewerker en enkele tijdelijke krachten, veelal studenten.
Men biedt ook een datarepository aan voor onderzoekers die de door hen geproduceerde datasets willen deponeren. Binnenkort verwacht men ook een self-deposit module te kunnen aanbieden, zodat dit zonder tussenkomst van de data library kan plaatsvinden. Op dit moment zijn er slechts enkele tientallen datasets gedeponeerd. De respondent wijt dit met name aan de tot voor kort overheersende cultuur binnen Canada om datasets uitsluitend onder zeer strikte voorwaarden ter beschikking te stellen vanwege privacyredenen. Inmiddels zijn er ontwikkelingen gaande om dit te versoepelen.

Universiteit van Edinburgh

De Universiteit van Edinburgh is een van de weinige universiteiten in het Verenigd Koninkrijk met een data library: volgens de respondent zijn er nog drie andere: de Universiteit van Southampton, de Universiteit van Oxford en de Londen School of Economic. Een belangrijke taak van de data library is om datasets te acquireren en deze beschikbaar te stellen aan gebruikers binnen de universiteit. Men ondersteunt deze beschikbaarstelling met persoonlijke ondersteuning.
Recentelijk heeft men ook vastgesteld dat er behoefte is aan ondersteuning bij het datamanagement van onderzoeksprojecten. Men heeft een aantal R&D projecten hiertoe uitgevoerd:

- De inrichting van een datarepository: dit datarepository is bedoeld om datasets die niet in het UKDA (het Nationale dataarchief voor de geesteswetenschappen en sociale wetenschappen) worden gedeponeerd (ofwel omdat zij tot een andere discipline behoren ofwel om een andere reden) op te slaan. Dit is mogelijk door een self-deposit optie. In de praktijk blijkt dat onderzoekers vaak persoonlijke

ondersteuning bij het deponeren willen hebben. Het datarepository is eind 2009 opgeleverd: op dit moment zijn er nog slechts een tiental datasets gedeponeerd. Overigens heeft de Universiteit van Edinburgh verplicht gesteld dat publicaties van de onderzoekers worden gedeponeerd in het publicatie repository. Men streeft er naar om in dezelfde procedure het eventuele deponeren van de bijbehorende dataset in het datarepository op te nemen.

- Data audit: men heeft volgens een bepaalde methode (data asset framework [16]) een audit uitgevoerd binnen de instelling naar datamanagement.
- Universitair beleid: op dit moment zijn er twee commissies door het Universiteitsbestuur ingesteld, die beleidsvoorstellen zullen ontwikkelen voor respectievelijk dataopslag en datamanagement. De respondent verwacht dat dit leidt tot o.a. een beleid over eigenaarschap van de data en hoelang data bewaard moet blijven.
- Online guidance: enige tijd geleden heeft men een aantal webpagina's gemaakt over datamanagement[17]. Deze webpagina's zijn tot nog toe redelijk uniek. Men streeft er naar dit op te volgen door het opzetten van een online cursus over datamanagement voor PhD studenten. Een subsidie hiertoe is aangevraagd.

De Universiteit van Amsterdam

De UB Amsterdam heeft een aantal R&D projecten uitgevoerd op het gebied van dataopslag en beschikbaarstelling:

- 'Testweeklab' zijn gegevens waarin eerstejaarsstudenten psychologische testen uitvoeren. Dit is een langlopend onderzoek dat elk jaar wordt herhaald met een nieuwe lichting eerstejaarsstudenten. In samenwerking met de betreffende onderzoekers heeft de UB een elektronische samenwerkingsomgeving ingericht, waarbij de datasets worden opgeslagen in Fedora.
- Verrijkte publicaties: men heeft een project uitgevoerd in samenwerking met het Tijdschrift voor Archeologie (uitgegeven door de Amsterdam University press) en met DANS. De gedachte was dat de onderzoeksdata bij DANS zouden worden gedeponeerd, maar dat de visualisaties van die onderzoeksdata in het tijdschriftartikel. De visualisaties zijn goed gelukt, nu is het zaak om deze dienst voor de langere termijn in een veranderende IT omgeving te handhaven.

[16] http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework
[17] http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt