# Het beheren van onderzoeksdata

MARNIX VAN BERCHUM, MARJAN GROOTVELD

## 1. Introductie

Dit artikel geeft een overzicht van recente ontwikkelingen op het gebied van onderzoeksdatamanagement. Een dergelijk overzicht is nooit volledig: onderzoeksdatamanagement — kort gezegd: het netjes omgaan met onderzoeksgegevens — is zowel nationaal als internationaal erg in beweging en regelmatig worden nieuwe datadiensten op de markt gebracht. Juist deze dynamiek spreekt echter voor een overzichtsartikel.

In het onderstaande is getracht de huidige stand van zaken in Nederland in kaart te brengen. Hiervoor is met name gekeken naar de nationale ontwikkelingen in het hoger onderwijs en wetenschap; commerciële diensten worden genoemd wanneer zij een rol hierin spelen, maar er wordt bijvoorbeeld geen aandacht geschonken aan 'big data' ontwikkelingen buiten de wetenschap of de open data-initiatieven van de overheid. Af en toe worden voorbeelden uit wetenschappelijke vakgebieden gekozen, maar het voert te ver om in te gaan op de (soms grote) verschillen tussen vakgebieden. Wie hierin geïnteresseerd is, verwijzen we graag naar (Borgman, 2015) en naar hoofdstuk 2 in (KNAW, 2012). De auteurs willen de lezer erop wijzen dat zij beiden werkzaam zijn (geweest) bij een instituut met de missie om onderzoeksdata duurzaam beschikbaar te stellen (zie paragraaf 4.4), wat van invloed is op de balans van de besproken onderwerpen.

Hoofdstuk 2 van dit artikel beschrijft het belang van datamanagement, waarna hoofdstuk 3 kernbegrippen introduceert zoals 'data', 'datamanagementplan' en 'onderzoekscyclus'. Het vierde hoofdstuk gaat nader in op de fasen van de onderzoekscyclus: hoewel een eerste *planning* van het benodigde datamanagement vanzelfsprekend in een vroeg stadium moet plaatsvinden, komen verschillende aspecten van de *uitvoering* bij de opeenvolgende fasen aan bod. Hoe data-ondersteuners zoals bibliotheekmedewerkers (kunnen) bijdragen aan de verschillende fasen komt aan bod in hoofdstuk 5, dat een inkijkje biedt in het Nederlandse krachtenveld.

## 2. Het belang van research data management

Of je nu onderzoek verricht in een lab, in het veld of op kantoor, in grote of kleine onderzoeksteams, er komen onderzoeksgegevens oftewel data aan te pas. Die data zijn waardevol en verdienen het om goed beheerd te worden. Het besef dat goed datamanagement een belangrijk onderdeel is van wetenschappelijk gedrag is de laatste jaren sterk gegroeid.

De stroom van door wetenschap gegenereerde data groeit hard, mede door de voor onderzoek relevante data van de social media en open overheidsdata. Het in 2010 door de Europese Unie (High Level Expert Group, 2010) gepubliceerde rapport *Riding the Wave* schetst de gevolgen, kansen en uitdagingen van deze 'vloedgolf van data'.

> The benefits are broad. With a proper scientific e-infrastructure, researchers in different domains can collaborate on the same data set, finding new insights. They can share a data set easily across the globe, but also protect its integrity and ownership. They can use, re-use and combine data, increasing productivity. They can more easily solve today's Grand Challenges, such as climate change and energy supply. Indeed, they can engage in whole new forms of scientific inquiry, made possible by the unimaginable power of the e-infrastructure to find correlations, draw inferences and trade ideas and information at a scale we are only beginning to see. For society as a whole, this is beneficial. It empowers amateurs to contribute more easily to the scientific process, politicians to govern more effectively with solid evidence, and the European and global economy to expand.
>
> But there are many challenges. How can we organise such a fiendishly complicated global effort, without hindering its flexibility and openness? How do we incentivise researchers, companies, and individuals to contribute their own data to the e-infrastructure – while still trusting that they can protect their privacy or ownership? How can we manage to preserve all this data, despite changing technologies and needs? How to convey the context and provenance of the data? How to pay for it all?

Naast deze kansen en uitdagingen kent onderzoek ook risico's. Fraudezaken in de afgelopen jaren en het daaropvolgende onderzoek van de commissie Schuyt (KNAW, 2012) hebben laten zien dat onderzoeksdata niet altijd op de wenselijke wijze worden beheerd en gedeeld, waardoor de wetenschap schade op kan lopen.

Financiers en beleidsmakers in de onderzoekswereld hebben op deze ontwikkelingen gereageerd met de formulering van databeleid, op internationaal, na-

tionaal en institutioneel[1] niveau. Belangrijke thema's hierin zijn transparantie en verifieerbaarheid van studies en hergebruik van onderzoeksgegevens. Concreter streven naar deze doelen moet leiden tot beter en efficiënter onderzoek. Voortbouwen op bestaande gegevens kan onderzoek immers versnellen en het niet opnieuw genereren van data — hergebruik dus — kan daarnaast financieel voordeel betekenen.

Eigenlijk ligt de Nederlandse Gedragscode voor Wetenschapsbeoefening (VSNU, 2014) al ten grondslag aan de noties van transparantie en verifieerbaarheid: "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". Of, zoals het in de strategie van DANS (2015b) staat: "De integriteit van de wetenschapsbeoefening heeft baat bij transparante onderzoeksprocessen en verantwoord datamanagement maakt daar deel van uit."

Naast deze doelen is het vooral voor onderzoekers zelf van belang om tijdens het werk hun werkproces en de gegevens goed te documenteren. Dit lijkt een open deur, omdat onderzoekers in de meeste disciplines ermee vertrouwd zijn dat een publicatie de gehanteerde onderzoeksmethodiek moet beschrijven. Deze beschrijving is echter zeker niet altijd voldoende om zelf na enige tijd nog precies te weten hoe bijvoorbeeld een bepaalde meting is verkregen of welke *query* in het statistiekpakket ten grondslag ligt aan een bepaalde tabel in de publicatie. Voor andere geïnteresseerden is dit dan nog moeilijker, zo niet onmogelijk; in feite kunnen zij dan niet op de data vertrouwen – een gemiste kans voor hergebruik en citatie van de data.

## 3. Kernbegrippen

Datamanagement kent nog geen lange traditie en het begrippenkader is in ontwikkeling. Hierdoor hanteren organisaties vaak hun eigen definitie van belangrijke begrippen zoals 'data' en 'datamanagement'. In zekere zin doet dit artikel dit ook; de auteurs willen geen scherpslijpers zijn en de lezers bewegen zich in verschillende kringen, met hun eigen gewoontes. Daarom volgt hier een beschrijving van de kernbegrippen voor het vervolg van dit artikel: data, me-

1 Een overzicht van databeleid bij Nederlandse universiteiten is te vinden op https://www.edugroepen.nl/sites/RDM_platform/Lists/RDM%20bij%20universiteiten%20in%20Nederland/AllItems.aspx

tadata, FAIR, *research data lifecycle,* datamanagement of RDM, datamanagementplan of DMP en archief[1].

Het begint met de onderzoeksgegevens ofwel **data**. Dit kunnen zijn: feiten, observaties, interviews, opnames, metingen, experimenten, simulaties en software; numeriek, beschrijvend en visueel; ruw, geschoond en bewerkt; al dan niet de onderbouwing van een (beoogde) publicatie; en bewaard en uitgewisseld in uiteenlopende formaten op diverse opslagmedia[2]. Deze niet-uitputtende opsomming wil aanknopingspunten bieden voor alle onderzoeksdisciplines, omdat herbruikbaarheid, controleerbaarheid en transparantie overal relevant zijn. Niet-digitale data, zoals papieren enquêtes en lichaamsweefsel, blijven in dit artikel buiten beschouwing.

Zoals boeken en artikelen voorzien worden van bibliografische informatie, zo worden data dat ook, alleen wordt deze informatie doorgaans **metadata** genoemd. Metadata is informatie over data, waarmee men de data bijvoorbeeld in online-portals kan vinden. Er zijn generieke en domeinspecifieke standaarden voor metadata; de laatste leveren vaak rijkere beschrijvingen op, maar worden zelden ondersteund in brede, meer algemene zoekportals. De grens tussen metadata en documentatie is niet scherp te trekken: het codeboek in de sociale wetenschappen dat de in het onderzoek gebruikte variabelen definieert, valt bijvoorbeeld in dit grensgebied. Belangrijker dan de terminologie is dat dergelijke informatie beschikbaar gesteld moet worden als context om de data goed te kunnen interpreteren. Met rijke metadata kunnen geïnteresseerden immers beter bepalen of een dataverzameling relevant en bruikbaar is voor hun eigen onderzoek. Diverse onderzoeksgemeenschappen hanteren voor hun vakgebied een metadatastandaard[3]; dit bevordert binnen een discipline het kunnen vinden en efficiënt gebruiken van bestaande data. De gedachte dat metadata altijd openbaar zijn, ook wanneer de data niet of slechts beperkt toegankelijk zijn, wordt breed gedragen.

In 2014 werden de **FAIR-principes** voor data geïntroduceerd, die, wanneer ze

1 De definities van data, datamanagement en metadata zijn ontleend aan het (ongepubliceerde) "Beleidskader Datamanagement NWO-instituten". De beschrijving van de research life cycle is, met toestemming, ontleend aan de cursus "Essentials 4 Data Support", zie http://datasupport.researchdata.nl/start-de-cursus/i-begrippen/research-lifecycle. Voor meer datajargon, zie http://datasupport.researchdata.nl/nl/start-de-cursus/i-begrippen/datajargon/.

2 NWO hanteert deze definitie in het Beleidskader Datamanagement NWO-instituten (ongepubliceerd).

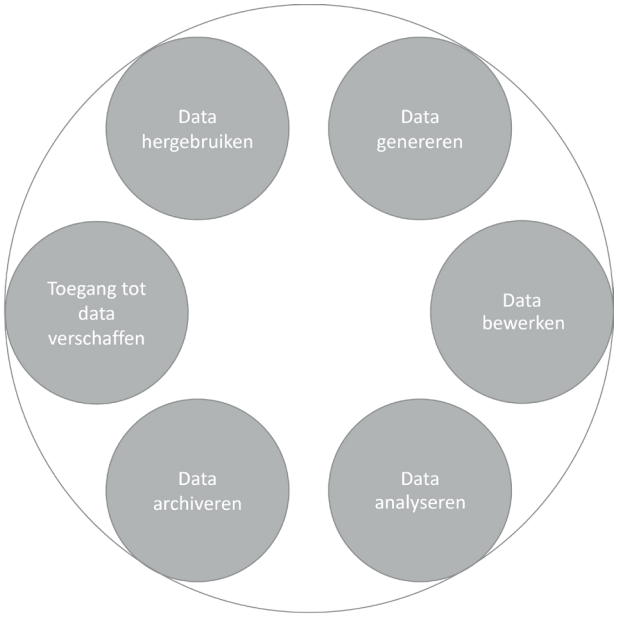
3 Een overzicht van metadatastandaarden voor uiteenlopende domeinen is te vinden op http://rd-alliance.github.io/metadata-directory/standards/.

in praktijk worden gebracht, ervoor zorgen dat data te vinden zijn (*findable*), toegankelijk zijn (*accessible*), te vergelijken en te combineren zijn met andere data (*interoperable*) en – daarmee – herbruikbaar zijn (*reusable*). De ambitie die aan de FAIR-principes ten grondslag ligt, is om dit zowel te bereiken voor mensen als voor machines. Wilkinson et al. (2016) beschrijven het ideaal van data die zo rijk en gestructureerd gedocumenteerd zijn dat ze *machine-actionable* zijn, dat wil zeggen, dat een 'autonomously-acting, computational data explorer' zelfstandig kan bepalen hoe nuttig een digitaal object is voor een gegeven taak en of de eraan verbonden licentie toegang en hergebruik toestaat. Momenteel is dit voor veel vakgebieden toekomstmuziek en is het vooralsnog de uitdaging om de FAIR-principes te concretiseren tot kwaliteitscriteria en richtlijnen voor onderzoekers die data produceren en voor leveranciers van datadiensten. Goed datamanagement ondersteunt de FAIR-principes.

Onderzoeksdata hebben een lange levensduur, vaak langer dan de periode tussen hun ontstaan en het schrijven van de wetenschappelijke publicatie die erop is gebaseerd. In verschillende fasen van een onderzoekscyclus hebben ze een andere functie en een andere waarde. Een ***research data lifecycle*** is een hulpmiddel om in beeld te brengen wat de verschillende fasen zijn, hoe deze in het leven van onderzoeksdata op elkaar aansluiten en hoe de keuzes die een onderzoeker in de ene fase maakt de datakwaliteit in een andere fase beïnvloeden. Een *lifecycle* helpt om het perspectief van de korte termijn naar de lange termijn te verschuiven: wat willen we met deze onderzoeksdata? Hoe zorg je ervoor dat de keuzes die je bij het verzamelen van de data maakt robuust genoeg zijn om archivering en hergebruik mogelijk te maken?

Van oudsher zijn onderzoekers in de meeste disciplines meer gericht op het publiceren van wetenschappelijke artikelen dan op de gegevens die zij genereren en gebruiken. Willen we juist de gegevens benadrukken, dan is het volgende cyclische model goed bruikbaar, dat gebaseerd is op de *research lifecycle* van het UK Data Archive (UKDA)[1]. Terwijl de focus op data ligt, zijn de fasen generiek genoeg om voor veel onderzoekers en andere betrokkenen herkenbaar te zijn.

1 http://www.data-archive.ac.uk/create-manage/life-cycle




*Figuur 1 Research data lifecycle (gebaseerd op UKDA)*

Er zijn ook andere *research lifecycles* in omloop, toegespitst op wat een gebruikersgroep nodig heeft. Voorbeelden zijn te vinden op de website van de cursus 'Essentials 4 Data Support'[1]. In dit artikel volgen we de zes fasen van de datacyclus van Figuur 1. Natuurlijk is het model een abstractie: de fasen zijn niet strikt te scheiden, er kunnen goede redenen zijn om tijdens het onderzoek terug te keren naar een eerdere fase. Bovendien beslaan diverse data-activiteiten, zoals het opslaan van data, meer dan een fase. Maar voordat we de eerste fase ingaan, introduceren we nog enkele andere begrippen die geregeld zullen terugkomen.

Onder **datamanagement** wordt het volledige traject verstaan van het creëren of vergaren van data tot het opslaan, onderhouden, archiveren, ontsluiten en langdurig bewaren (preserveren) van data. Er wordt geen onderscheid gemaakt tussen de doelen van dataopslag zoals controle, verificatie, replicatie, hergebruik of koppeling van de data[2]. *Research data management* en RDM

1 http://datasupport.researchdata.nl/start-de-cursus/i-begrippen/research-lifecycle/
2 NWO hanteert deze definitie in het Beleidskader Datamanagement NWO-instituten (ongepubliceerd).

zijn synoniemen hiervan. Voor dit traject wordt ook wel het begrip 'data stewardship' gebruikt, terwijl anderen die term beperken tot de activiteiten die op duurzaamheid gericht zijn. Goed en gedocumenteerd datamanagement levert de nodige transparantie op waarmee de data controleerbaar en verifieerbaar worden.

Een **datamanagementplan** of DMP is een aanvulling op een onderzoeksplan en beschrijft onder andere welke soort en hoeveel data het project verwacht op te leveren, wat hiervan op welke wijze duurzaam bewaard zal worden en onder welke voorwaarden de data toegankelijk zullen zijn. Indien van toepassing beschrijft het de hardware en software die nodig zijn om de data te gebruiken. Het DMP brengt de hele datalevenscyclus in kaart. Het is een dynamisch document dat in de loop van het onderzoek aanpassing behoeft, bijvoorbeeld omdat bepaalde zaken veranderen (onverwacht wel/geen toegang tot beoogde bronnen, nieuwe partijen in het projectconsortium en dergelijke). Op het moment van schrijven van dit artikel vinden financiers NWO, ZonMw en de Europese Commissie deze dynamiek vanzelfsprekend. Het DMP moet kort na goedkeuring van de projectaanvraag worden geschreven, maar hoeft dus niet 'in één keer goed'. Ook onderzoeksinstellingen en vakgroepen die een DMP eisen, hanteren vergelijkbare uitgangspunten.

**Datamanagementplannen**

Een datamanagementplan of DMP is een handige stimulans voor onderzoekers om in een vroeg stadium advies in te winnen en afspraken te maken over bijvoorbeeld juridische aspecten (*wie mag wat met de data?*) en technische voorzieningen (*welke hard- en software heb ik nodig?*). Omgekeerd is betrokkenheid bij de planning voor de onderzoeksorganisatie, de ondersteunende afdelingen en externe dienstenaanbieders een goede manier om onderzoeksinformatie en (beoogde) werkprocessen te stroomlijnen. Ze weten daardoor namelijk wat er op hen afkomt en kunnen de onderzoekers tijdig adviseren. Data-archieven worden bijvoorbeeld graag al in dit stadium geraadpleegd over mogelijke datadeponeringen in de toekomst, zodat ze de onderzoeker kunnen informeren over de gewenste bestandsformaten, metadata en dergelijke.

Evenals bij de *lifecycles* zijn er voor DMP's wereldwijd veel sjablonen in omloop. Uit de sjablonen van onder andere Nederlandse en Europese onderzoeksfinanciers heeft DANS de elementen geëxtraheerd en toegelicht die een DMP minimaal dient te beschrijven, zie de onderstaande tabel (DANS, 2015a).

| **1** | **Administratieve informatie** |
|---|---|
| 1.a | Projectnaam, hoofdonderzoeker, financier(s), datum van dit plan en van eerdere versies. |

| | |
|---|---|
| 1.b | Wie is de eerstverantwoordelijke voor het datamanagement? |
| **2** | **Beschrijving van de data** |
| 2.a | Worden bestaande data hergebruikt of nieuwe data gegenereerd? |
| 2.b | Om welke soort(en) data gaat het; omvang van de bestanden; groeitempo? |
| **3** | **Standaarden en metadata, ofwel alles wat nodig is om de data te vinden en te benutten** |
| 3.a | Welke metadatastandaarden worden gebruikt (*vindbaarheid*)? |
| 3.b | Welke coderingen e.d. worden gebruikt die toekomstige koppeling met andere data mogelijk maken (*duiding, interoperabiliteit*)? |
| 3.c | Welke software en eventueel hardware wordt er gebruikt (*duiding, bruikbaarheid*)? |
| 3.d | Wat wordt er gedocumenteerd en bewaard om replicatie mogelijk te maken? Wat zijn de afspraken als betrokkenen (voortijdig) vertrekken? |
| **4** | **Ethisch en juridisch** |
| 4.a | Hoe wordt bij het verwerven of genereren van de data de hiervoor eventueel benodigde toestemming verkregen van dataleverancier/ proefpersonen/ …? Welke beperkingen gelden er eventueel tijdens het onderzoek? |
| 4.b | Hoe worden gevoelige gegevens beschermd tijdens en na het project? |
| 4.c | Zijn de data na het project – eventueel na een embargoperiode – als Open Access beschikbaar? Zo nee, welke voorwaarden gelden er? |
| **5** | **Opslag en archivering** |
| 5.a | Hoe wordt voldoende opslag- en back-up-capaciteit tijdens het project geregeld, inclusief versiebeheer? Zijn de kosten hiervoor gedekt; zo nee…? |
| 5.b | Waar en hoe lang worden de data na afloop van het project beschikbaar gesteld voor vervolgonderzoek en verificatie? Is dit een *Trustworthy Digital Repository*, dus met een internationale certificering? Zo niet, hoe worden de data dan vindbaar (*denk aan metadata en aan persistent identifiers zoals DOI, Handle en URN*) en duurzaam toegankelijk en bruikbaar? |
| 5.c | Zijn de kosten voor (het voorbereiden van de data voor) archivering gedekt; zo nee…? |

In aansluiting op het DMP en vooruitlopend op het volgende hoofdstuk noemen we **archief** als laatste kernbegrip. Waar het gaat om het bewaren van data is het namelijk verstandig om verschil te maken tussen het bewaren van data tijdens lopend onderzoek ('opslaan') en het langdurig bewaren van data na afloop van het onderzoek ('archiveren' of 'preserveren'). Het kan in deze twee fasen gaan om dezelfde data, maar er kunnen bijvoorbeeld andere afspraken gelden over wie er toegang heeft tot de data en/of wie verantwoordelijk is voor het behoud van en de toegang tot de data. In dit artikel worden de begrippen 'archief' en 'repository' door elkaar gebruikt. Van een '*Trustworthy Digital Repository*' of TDR is sprake wanneer het archief of de repository als zodanig is gecertificeerd.

## 4. De levenscyclus van data

In dit hoofdstuk gaan we nader in op de fasen die onderzoekers en hun data doorlopen, met de kanttekening dat er grote verschillen kunnen bestaan tussen en zelfs binnen vakgebieden.

### *4.1 Fase 'Data genereren'*

In figuur 1 staan hergebruiken en genereren van data opzettelijk samen bovenaan. Bij de start van een nieuw onderzoek gaat een wetenschapper namelijk idealiter eerst op zoek naar bestaande data, alvorens zelf data (opnieuw) te genereren of te verzamelen. Onderzoeksfinanciers worden ook steeds alerter op hergebruik, omdat dit potentieel tijd en geld bespaart; paragraaf 4.6 gaat verder in op hergebruik.

De wijze van ontstaan van nieuwe data verschilt per discipline en heeft een eigen verloop. Het *verzamelen* van survey-data binnen de sociale wetenschappen is bijvoorbeeld wezenlijk anders dan het verzamelen van de terabytes aan data die de experimenten van de LOFAR-telescoop in Exloo *genereren*.[1] Om te zorgen dat de data in de volgende fasen van de cyclus bruikbaar blijven — en dat de data van de ene naar de andere fase kunnen overgaan — zal in de eerste fase van creatie aan verschillende aspecten aandacht geschonken moeten worden.

- Voor onderzoek met *persoonlijke*, bijvoorbeeld medische gegevens is het verkrijgen van een *informed consent* voor latere verwerking van de data

1 Voor meer voorbeelden van de creatie van data zie http://datasupport.researchdata.nl/start-de-cursus/iii-onderzoeksfase/data-verzamelen/.

noodzakelijk[1]. Wordt dit achterwege gelaten voordat de data verzameld worden, dan zijn deze mogelijk onbruikbaar. Een alternatief kan zijn om de data zo te bewerken dat herleiding tot individuen redelijkerwijs niet mogelijk is. Voor onderzoek met veel en/of geaggregeerde data is dit niet per se een probleem, maar in andere gevallen — na bijvoorbeeld anonimisering ervan — neemt de zeggingskracht van de data wel degelijk af. Kortom, dit vereist al bij het voorbereiden van het onderzoeksvoorstel zorgvuldige overweging. Vergelijkbare afwegingen gelden voor onderzoek met bijvoorbeeld commercieel of militair gevoelige gegevens.

- Ook voor onderzoek dat uitgevoerd wordt in samenwerking met niet-publieke partijen kunnen restricties gelden. Commerciële belangen of mogelijke patenten verhinderen dat data meteen beschikbaar en herbruikbaar zijn. Het vastleggen van dergelijke restricties gebeurt vooraf in bijvoorbeeld een *Consortium Agreement* (in het geval van door de Europese Unie gefinancierde projecten).
- Vanaf het begin zal een onderzoeker correcte en consistente metadata moeten toekennen, om de data in latere fasen begrijpelijk te houden.
- Meer 'technische' zaken zijn de keuze voor bestandsformaten (zie voor de relevantie hiervan fase 'Data hergebruiken'), conventies rondom de naamgeving van bestanden en de organisatie van de mappenstructuur. '151016' als datumaanduiding in de bestandsnaam is bijvoorbeeld niet eenduidig: het kan zowel 15 oktober 2016 als 16 oktober 2015 zijn. Beter is een standaardnotatie als YYYY-MM-DD, dus bijvoorbeeld 20161015 voor 15 oktober 2016. Voor expliciete afspraken binnen een projectteam over mappenstructuur en bestandsnamen wordt ook wel de term 'data-organisatie' gebruikt.

## *4.2 Fase 'Data bewerken'*

In de bewerkingsfase vinden activiteiten plaats zoals vertalen, valideren, anonimiseren en opschonen door bijvoorbeeld statistische uitschieters (gedocumenteerd!) uit de ruwe data te verwijderen, evenals het verder beschrijven en documenteren van de data. De activiteiten worden uitgevoerd op ruwe data, en bewerkte data vormen het resultaat. Ook deze fase vertoont per discipline grote verschillen: soms zijn er opeenvolgende bewerkingsslagen en moeten alle tussenversies bewaard blijven om 'het spoor terug' te volgen, terwijl in zeer data-intensieve vakgebieden zoals deeltjesfysica een geautomatiseerd proces

1 Het UK Data Archive biedt hierover meer informatie en eveneens verschillende sjablonen, zie http://www.data-archive.ac.uk/create-manage/consent-ethics/consent?index=3.

leidt tot een eerste reductie van de data; de data na deze reductieslag worden als ruwe data beschouwd.

Een belangrijke generieke activiteit is de opslag van de data tijdens het onderzoek, ook vaak '*storage*' genoemd (paragraaf 4.4 beschrijft opslag na afloop van het onderzoek, oftewel de archivering). Hierbij moet wederom aandacht geschonken worden aan aspecten die het vervolg van de data in de cyclus mogelijk maken. Regelgeving en gedragscodes (nationaal, institutioneel, wettelijk) en bestaande afspraken spelen hierbij vaak een grote rol. Een voorbeeld is de geografische locatie van de fysieke opslag: universiteiten staan soms niet toe om de data buiten de campus op te slaan; in andere gevallen geldt de voorwaarde dat de data binnen Europa blijven. Technisch moet het mogelijk zijn om de opgeslagen data in een latere fase te kunnen migreren naar een archief; een offline storage-locatie met een exotisch besturingssysteem maakt migratie van de data moeizaam. Voor storage in het algemeen zijn verschillende *best practices* te noemen[1]:

- Sla data op in een open standaardformaat dat niet gebonden is aan een bepaalde softwareleverancier — omwille van hergebruik en toekomstige conversie door een archief.
- Volg zelfs voor een kortetermijnproject een data-opslagstrategie waarbij twee verschillende typen opslagmedia gebruikt worden, bijvoorbeeld CD en harddisk. Afhankelijkheid van slechts één type levert risico's op wanneer dit medium (of de benodigde software) corrupt raakt of veroudert.
- Kopieer of migreer data elke twee tot vijf jaar naar nieuwe opslagmedia. Opslagmedia gaan in kwaliteit achteruit en zijn daardoor op termijn niet meer te openen met de dan gangbare hardware en software.
- Overschrijf een oude back-up niet met een nieuwe back-up. Het is beter om een geheel nieuwe back-up maken van files die zijn veranderd, zodat je altijd 'terug in de tijd' kunt naar een specifieke versie.
- Bereken een *checksum* (controlegetal) van de data. Controleer vervolgens regelmatig de integriteit van de data met een *checksum checker.* Hierdoor weet je zeker dat je met eenzelfde versie werkt. Als er 'een bitje omvalt', leidt dit tot een afwijkende *checksum*; een vorige of eerdere versie moet dan teruggehaald worden om uit te zoeken wat er is misgegaan.

De markt van online-toepassingen die bovenstaande punten kunnen ondersteunen is groot. Er is aanbod van private partijen zoals Mendeley, Figshare en Dropbox, evenals van non-profit organisaties: in Nederland biedt bijvoorbeeld SURF Beehub en SURFdrive aan, en DANS DataverseNL. Het zal duidelijk zijn dat zulke opslagdiensten in functionaliteit verschillen. Bij de

1 http://datasupport.researchdata.nl/start-de-cursus/iii-onderzoeksfase/data-opslaan/

keuze van een systeem spelen eveneens de (beleids)kaders van een instelling mee: soms zijn er reeds bestaande contracten met aanbieders waar de onderzoekers gebruik van dienen te maken of is het gebruik van een dienst juist niet toegestaan. Een instelling kan ook de voorkeur geven aan een systeem dat in een internationale samenwerking wordt gebruikt of minimale veiligheidseisen stellen.

Tijdens het bewerken van de data en ook in de hierna volgende analysefase worden protocollen, software en tools gebruikt. Ook in dit opzicht moet transparant zijn wat deze zijn en hoe deze zijn gebruikt. Hoe zijn bijvoorbeeld statistische *outliers* verwijderd? Hoe zijn de uitspraken in interviews gecodeerd (dat wil zeggen, in categorieën ingedeeld): machinaal, door een of meer assistenten, door de onderzoeker zelf? In het ideale geval kan de onderzoeker bij het plannen en later bij het documenteren van zijn of haar werk verwijzen naar standaardmethoden en -tools binnen het vakgebied.

### *4.3 Fase 'Data analyseren'*

Na het bewerken van de data kan interpretatie en analyse plaatsvinden. In deze fase worden visualisaties vanuit de data gegenereerd en ontstaan artikelen en andere wetenschappelijke output. Voor veel onderzoekers is dit de kern van het onderzoeksproces; het moment waarop ze de onderzoekshypothese toetsen aan de data of, in onderzoek dat niet vanuit een hypothese wordt opgezet, patronen in de data blootleggen. Goede verslaglegging van alle stappen en controle op de uitvoering, zoals de Nederlandse Gedragscode voor Wetenschapsbeoefening (VSNU, 2014) bij de uitwerking van het controleerbaarheidsprincipe beschrijft, bepalen in deze fase het datamanagement. De stap van (ruwe) onderzoeksdata naar een artikel dient transparant te zijn: het moet voor een andere onderzoeker duidelijk zijn hoe de data de conclusie ondersteunen en wat er met de data gebeurd is om tot de conclusie te komen. Ook in deze fase is de beschrijving van de gebruikte protocollen, software en tools nodig om het hergebruik van de data te garanderen. Zo'n beschrijving maakt doorgaans deel uit van een wetenschappelijke publicatie, maar niet altijd in de mate van detail die voor replicatie van het onderzoek nodig is. De Gedragscode noemt bijvoorbeeld ook documentatie van afspraken en beslissingen (VSNU, 2014, p. 8). Afhankelijk van de standaarden en gewoontes in een vakgebied, is een uitgebreide beschrijving nodig of kan met een verwijzing naar de relevante standaard worden volstaan. De vier dataniveaus bijvoorbeeld die in de deeltjesfysica[1] worden gebruikt, zijn

1 Zie bijvoorbeeld de vier niveaus van datapreservering in https://www.dphep.org/sites/site_dphep/content/e36435/e67191/e221646/chep13validationPoster.pdf.

binnen deze community bekend en hoeven dus niet uitgebreid toegelicht te worden.

### *4.4 Fase 'Data archiveren'*

Zoals vermeld bij de kernbegrippen is het verstandig om verschil te maken tussen het beheren van data *tijdens* onderzoek ('opslaan') en het langdurig beheren van data *na afloop* van het onderzoek ('archiveren' of 'preserveren'). Bij de overgang van de analysefase naar de archiveringsfase is de onderzoeker in het tweede stadium beland. Wellicht moeten er nu andere afspraken worden gemaakt over wie er toegang heeft tot de data en/of wie verantwoordelijk is voor het behoud van en de toegang tot de data. Het begrip 'preserveren' omvat bovendien de activiteiten die nodig zijn om veroudering en daarmee onbruikbaarheid van data te voorkomen. Anders gezegd: het gaat nu om de activiteiten die digitale duurzaamheid bewerkstelligen.

**Digitale duurzaamheid**
"Duurzame toegankelijkheid gaat over het op lange termijn toegankelijk houden van digitaal en gedigitaliseerd erfgoed. Snelle technologische veranderingen zijn kenmerkend voor de digitale wereld. Hardware en software verouderen snel. Wie kan er tegenwoordig nog grote of kleine floppy's lezen? Toch is er in het verleden veel opgeslagen op dit medium. Het is dus belangrijk om van begin af aan goed na te denken over de vorm waarin je je digitale data voor de toekomst wilt bewaren."[1] Dit appel van DEN/Kenniscentrum Digitaal Erfgoed geldt net zo goed voor onderzoeksdata. DEN is, evenals de Koninklijke Bibliotheek (KB), het Nationaal Archief, DANS en het Nederlands Instituut voor Beeld en Geluid, lid van de Nationale Coalitie Digitale Duurzaamheid[2]. Deze koepelorganisatie bundelt informatie voor de leden en voert projecten uit in het nationale Netwerk Digitaal Erfgoed.
Een van de NCDD-projecten richt zich op certificering van repositories, en juist in de onderzoekssector is hier al veel ervaring mee. De meeste dataarchieven die het Data Seal of Approval[3] mogen voeren of lid zijn van ICSU World Data System[4] preserveren namelijk onderzoeksgegevens. Om bijvoorbeeld het Data Seal of Approval (DSA) te verwerven, moet het archief er aantoonbaar voor zorgen dat:

1 http://www.den.nl/thema/16/
2 http://www.ncdd.nl/
3 http://www.datasealofapproval.org/
4 https://www.icsu-wds.org/community/membership

- de data op internet te vinden zijn;
- de data goed toegankelijk zijn (duidelijke rechten en licenties);
- de data een bruikbaar formaat hebben;
- de data betrouwbaar zijn;
- de data een unieke en persistente identifier hebben zodat ernaar verwezen kan worden.

Niet geheel toevallig vertonen de DSA-criteria voor repositories veel overeenkomsten met de FAIR-data-principes. Voor meer informatie over zogenoemde *Trustworthy Digital Repositories* en DSA verwijzen we graag naar Dillo en de Leeuw (2014).
Hoewel hiervoor nog weinig duurzame repositories beschikbaar zijn, is er een groot besef dat ook *software sustainability* een belangrijk onderdeel van datamanagement en digitale duurzaamheid is. Een groeiende hoeveelheid data is voor interpretatie en hergebruik namelijk afhankelijk van software. Software in GitHub kan eenvoudig gearchiveerd worden in het Zenodo-archief[1] van CERN; daarmee blijft de code lang beschikbaar, zij het dat beleid voor verdere preservering nog ontbreekt. In Nederland werken de NCDD, de KB, het Netherlands eScience Center en DANS samen om kennis, diensten en ondersteuning te ontwikkelen; DANS is hiervoor ook een samenwerking aangegaan met het Franse Inria voor de verdere ontwikkeling van het Software Heritage Initiative[2].

Voor FAIR data adviseren we dat de data met alle documentatie die voor hergebruik nodig is, worden ondergebracht oftewel 'gedeponeerd' in een archief voor onderzoeksdata dat zowel de expliciete doelstelling als de expertise heeft om onderzoeksgegevens duurzaam te bewaren en bruikbaar te houden. Instellingen waar het onderzoek is uitgevoerd, bieden ook opslag op hun eigen server aan, maar doorgaans ontbreekt daar die specifieke expertise, omdat langetermijnpreservering en -toegang niet tot de kerntaken van die instelling behoren.

Onderzoeksfinanciers zoals de Europese Commissie spreken geregeld een voorkeur uit voor gecertificeerde data-archieven, hoewel die nog niet in alle disciplines bestaan. NWO stelt: "Na het onderzoek worden de data bij voorkeur gearchiveerd bij een (inter)nationale data repository. Wanneer dit niet mogelijk is, dienen de data te worden gearchiveerd door de institutionele repository." De toelichting op het DMP (NWO, 2016) verwijst onder meer naar de duurzame richtlijnen van DSA.

1 https://zenodo.org/
2 https://www.softwareheritage.org/mission/

Wanneer er geen specifieke archieven beschikbaar zijn, kan ook gebruik gemaakt worden van een generieke oplossing als Zenodo. Het Registry of Research data Repositories (re3data.org) is een register van datarepositories wereldwijd. Het register geeft van de repositories informatie over eventuele certificeringen, *persistent identifiers*, licenties en toegangscategorieën en is eenvoudig doorzoekbaar op land, discipline of type data.[1] Voor de voorbereiding op het deponeren van data – dus in een veel eerdere fase in de onderzoekscyclus – bieden archieven zoals Dryad[2] en DANS EASY[3] adviezen en instructies. Het is dan ook verstandig om tijdig contact te zoeken met een archief dat expertise heeft op het betreffende vakgebied en/of het soort data, zoals enquêtedata of sensordata.

Om toekomstig hergebruik van data te garanderen, adviseren data-archieven om zogenoemde duurzame formaten te kiezen: bestandsformaten die veel worden gebruikt, open specificaties hebben en onafhankelijk zijn van specifieke software en leveranciers. Zulke bestanden kan het data-archief op termijn converteren naar een volgende generatie duurzaam formaat, waardoor de data bruikbaar blijven. Afhankelijk van de doelgroepen – wetenschappelijke disciplines kunnen hun eigen specifieke formaten hanteren – en de expertise van het archief waar de data uiteindelijk terecht zullen komen, zal het archief bepaalde bestandsformaten wel of niet tot duurzame formaten rekenen; zie bijvoorbeeld de informatie van 4TU.ResearchData[4] en DANS[5] over de door hen gehanteerde voorkeursformaten.

De deponeerder moet bij de overgang van de analysefase naar de archiveringsfase een aantal zaken weten en zo nodig afronden of regelen:

- Accepteert het archief de bestandsformaten, of moeten de data eerst nog naar een ander formaat geconverteerd worden om ze duurzaam bruikbaar te houden?
- Welke metadata horen bij de data? Bieden ze genoeg informatie voor derden om de data te vinden en een indruk te krijgen van de relevantie en bruikbaarheid? Als er een disciplinespecifieke metadatastandaard bestaat of vereist wordt, is die dan ook gebruikt?
- Is alle documentatie volledig en in passende bestandsformaten, zodat een archiefbezoeker en potentiële datagebruiker de data in hun context kan interpreteren?

1 http://re3data.org
2 https://datadryad.org/pages/faq
3 https://dans.knaw.nl/nl/deponeren/toelichting-data-deponeren
4 http://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/
5 http://dans.knaw.nl/nl/deponeren/toelichting-data-deponeren/DANSpreferredformatsNL.pdf

- Ondersteunt het archief de gewenste toegangscategorie: Open dan wel Restricted Access, met eventuele gradaties?
- Wie krijgt toegang tot de data? Zijn eventuele voorwaarden en beperkingen, zoals 'wij berekenen kosten voor de selectie van de data' of 'alleen voor onderzoekers', duidelijk geformuleerd? Is het duidelijk hoe geïnteresseerden om toegang tot de data kunnen verzoeken als Open Access niet van toepassing is?
- Kent het archief aan de data een *persistent identifier* toe zoals de Digital Object Identifier (DOI), waarmee gebruikers de data correct kunnen citeren, zoals dat ook regel is voor het verwijzen naar publicaties en boeken?[1]

Diverse archieven formaliseren de datadeponering op het moment van overdracht met een deponeerlicentie. Met de licentie verklaart de deponeerder dat hij of zij rechthebbende is, dan wel met toestemming van de rechthebbende handelt. De licentie legt bijvoorbeeld vast dat de deponeerder de data ook elders beschikbaar kan stellen, maar dat het archief het recht en de verantwoordelijkheid heeft om de data tijdig naar een nieuw bestandsformaat te converteren. Ook kan erin staan wat het archief na afloop van een eventuele embargoperiode doet en na het overlijden van de deponeerder of de opheffing van het deponerende instituut.[2]

### *4.5 Fase 'Toegang verschaffen tot data'*

Toegang verschaffen tot de data betekent in dit stadium: ervoor zorgen dat buitenstaanders, dat wil zeggen, personen die niet bij het onderzoek betrokken waren, bij de data kunnen. *Findable* en *accessible* zijn de voornaamste FAIR-karakteristieken van deze fase, terwijl archieven zoals de Australian National Data Service (ANDS)[3] stellen dat de data nu 'gepubliceerd' zijn: "This domain involves the public sphere (publication in the sense of making public)". De data zijn, voorzien van metadata en overige benodigde documentatie, gedeponeerd in een extern archief of opgeslagen in het repository van de onderzoeksinstelling. Het woord 'public' impliceert overigens niet dat alle data open

1 Een *persistent identifier* (PID) is een unieke identificatiecode van een digitaal object die op een afgesproken plaats wordt geregistreerd. Hij blijft gegarandeerd werken, ook al verandert het webadres van een organisatie. Voor informatie over DOI's en andere PID's, zie het dossier op http://www.ncdd.nl/pid/.

2 Zie bijvoorbeeld DANS http://www.dans.knaw.nl/nl/over/organisatie-beleid/juridische-informatie.

3 http://www.ands.org.au/guides/curation-continuum

en voor iedereen toegankelijk zijn, net zomin als dit voor alle tijdschriftpublicaties geldt.

In het ideale geval zijn data voor iedereen toegankelijk en goed gedocumenteerd; dan vergt deze fase weinig aandacht. In praktijk echter zijn er wel wat aandachtspunten, en deze grijpen terug op eerdere fasen. Zo moet bijvoorbeeld duidelijk zijn wie wat mag met de data, en dit hangt mede af van eventueel *informed consent* en contracten met projectpartners en/of derden, zie paragraaf 4.1. Hiervoor dient een gebruikerslicentie zoals de Creative Commons licenties[1], die meer en minder restrictieve vormen kennen. Het is van belang om ook aan open data een licentie toe te kennen[2]: ten eerste ligt de openheid dan expliciet vast en ten tweede kan de oorspronkelijke rechthebbende hiermee desgewenst vastleggen dat data die ervan worden afgeleid eveneens open beschikbaar gesteld moeten worden. Dit principe van *share alike* kan men bijvoorbeeld ook hanteren bij data die niet commercieel gebruikt mogen worden: voor zogenoemd *downstream use*, dus in een volgende ronde van hergebruik, geldt dan dezelfde beperking. In een gebruikslicentie kan ook staan dat de gebruiker gehouden is aan auteurs- en databankrecht, aan academische mores zoals wetenschappelijk correct citeren en aan de Wet Bescherming Persoonsgegevens (WBP) – dit laatste is relevant wanneer data tot individuen te herleiden zijn[3]. Gebruikslicenties worden vaak verstrekt door het archief waar de data gedeponeerd zijn en worden samen met de deponeerlicentie opgesteld (zie paragraaf 4.4).

Niet alleen is het aan te bevelen om de keuze tussen Open Access en Restricted Access in een vroeg stadium te maken; als men voor Restricted Access kiest, dienen namelijk de voorwaarden voor toegang duidelijk gemaakt te worden. Wil de rechthebbende onderzoeker of onderzoeksgroep bijvoorbeeld opgevoerd worden als co-auteur bij publicaties op basis van de data of worden er kosten in rekening gebracht voor het selecteren van de gevraagde data uit een grotere verzameling? Welke informatie, zoals onderzoeksvraag en -methodiek, moet een geïnteresseerde voorleggen om toegang te krijgen, en wie beslist er over een dergelijk verzoek: de oorspronkelijke onderzoeker (niet bepaald een duurzame aanpak), diens instelling, het data-archief (dat niet per se inhoudelijke kennis over de data heeft)?

1 https://creativecommons.org/

2 Voor open software en de bijbehorende documentatie komen de GNU-licenties in aanmerking, https://www.gnu.org/licenses/licenses.html.

3 VSNU Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek, http://vsnu.nl/code-pers-gegevens.html; Wet Bescherming Persoonsgegevens, http://wetten.overheid.nl/BWBR0011468/2016-01-01.

Een laatste aspect van deze fase dat het UK Data Archive noemt, is 'promoting data'. De gearchiveerde data zullen via de catalogus van het data-archief te vinden zijn, en het archief zal de metadata ook – geautomatiseerd – aan zoekportals of *discovery services* beschikbaar stellen, zoals NARCIS[1], DataCite[2], OpenAIRE[3] en B2FIND[4]; bij instellingsrepositories spreekt dit niet altijd vanzelf. Daarnaast kunnen de onderzoeker zelf, diens instelling, het eventuele projectteam en de financier attenderen op de beschikbaarheid van de data. Hierbij is het zaak om de *persistent identifier* te gebruiken die het archief aan de data heeft toegekend, want dit is de betrouwbare, duurzame verwijzing die ook in citaties gebruikt dient te worden.

### *4.6 Fase 'Data hergebruiken'*

Hergebruik van data is een belangrijk doel van de hele datacyclus. Door goed management van de opslag en de archivering, door de documentatie en de juiste formaten van de data, zijn deze geschikt voor controle, follow-up en/of geheel nieuw, origineel onderzoek. De mate van hergebruik, en de gewoonten en afspraken hierover, kunnen verschillen per discipline, afhankelijk van de mate van 'data-intensiteit' die het vakgebied kenmerkt. Binnen de ene onderzoeksgroep zal het gebruikelijk zijn dat een promovendus voortbouwt op de dataverzameling van zijn voorgangers of teamgenoten; binnen een vakgebied waar onderzoekers meestal alleen of in kleine teams werken, worden data vaker zelf verzameld of gegenereerd. Ook de mate waarin tools, software, modellen of methodieken hergebruikt worden, varieert sterk. Wanneer een discipline of een vakgroep afspraken maakt over de omgang met data, (mogelijk) hergebruik inbegrepen, is het aan te raden om die afspraken goed te laten aansluiten op de eigen workflow en mores; zie Aerts en Doorn (2016) evenals het tekstkader over databeleid in paragraaf 5.1.

Naast hergebruik in nieuw onderzoek kunnen wetenschappers een dataset ook beschrijven en beoordelen in een zogenaamde *data paper*. Beschikbare onderzoeksdata kunnen eveneens gebruikt worden voor educatiedoeleinden[5]. Afhankelijk van de toegangsrechten kunnen ook derden gebruik maken van

1 NARCIS biedt de toegang tot wetenschappelijke informatie in Nederland, http://www.narcis.nl/.
2 http://search.datacite.org/
3 https://www.openaire.eu/search/find?keyword=
4 http://b2find.eudat.eu/
5 Een bijzonder voorbeeld van data geschikt voor hergebruik in educatieve context, zijn de verschillende datasets die het Centraal Bureau voor de Statistiek beschikbaar stelt op https://www.cbs.nl/nl-nl/onze-diensten/in-de-klas.

de data: burgers (de zogenaamde *citizen science*), overheidsinstellingen en bedrijven.

In deze fase worden de mogelijkheden van het hergebruik van data grotendeels bepaald door activiteiten uitgevoerd in de eerdere, hierboven beschreven, fasen:

- Keuze van archief: aangezien het archief dat de data beschikbaar stelt voor een groot deel bepalend is voor de mate van herbruikbaarheid van de data (denk aan de toegangsrechten die toegekend kunnen worden of de ondersteuning die het archief biedt voor bepaalde, disciplinaire metadata), is de initiële keuze voor een archief zeer belangrijk. In paragraaf 4.4 zijn hiervoor criteria en hulpmiddelen genoemd. Dit veronderstelt wel dat de onderzoeker/deponeerder kennis heeft van dergelijke hulpmiddelen (zie paragraaf 5.2).
- Documentatie en metadata: de kwaliteit van de beschrijving van de data heeft grote invloed op de begrijpelijkheid van de data. Wanneer uit de metadata en documentatie niet af te leiden is hoe de data tot stand zijn gekomen, wat de betekenis is van gebruikte variabelen etc., zal het voor een gebruiker lastiger zijn om de toepasbaarheid van de data te beoordelen. Ook het beschikbaar stellen van bijbehorende software, scripts en methodiek maakt het inzichtelijker wat er met de data gedaan kan worden. Kortom, hoe meer documentatie over de context van de data, des te beter.
- Bestandsformaten: de gekozen bestandsformaten bepalen door welke software de data hergebruikt kunnen worden. Een keuze voor open standaarden, bijvoorbeeld OpenDocument Tekst (*.odt) in plaats van Word© (*.docx), maakt hergebruik minder afhankelijk van een bepaald softwarepakket. Bovendien is het door de openheid van deze formaten (in tegenstelling tot de zogenaamde *proprietary* formaten van bijvoorbeeld Microsoft) mogelijk om nieuwe software te ontwikkelen voor het gebruik van deze formaten. *Trustworthy Digital Repositories* bieden daarom vaak hun data aan in deze open, duurzame formaten (zie paragraaf 4.4).
- Toegangsrechten: wellicht het belangrijkste aspect voor hergebruik vormen de rechten die een gebruiker heeft op een dataset. Is in een eerdere fase besloten dat de data niet open beschikbaar zijn voor hergebruik, dan kan en zal dit laatste dus ook niet plaatsvinden! Het is niet voor niets dat subsidiegevers en beleidsmakers het volledig open beschikbaar stellen van data stimuleren, al erkennen ze dat dit niet altijd mogelijk is. ‘Open als het kan, beschermd als het moet’ is het motto.

## 5. De rollen en verantwoordelijkheden, in Nederland

### *5.1 De betrokken stakeholders*

In het voorafgaande zijn de verschillende fasen van de levenscyclus van onderzoeksdata besproken, waarbij gerefereerd is aan betrokken stakeholders die een rol spelen: onderzoekers, financiers, archieven etc. Dit hoofdstuk bekijkt die rollen en de bijbehorende verantwoordelijkheden in meer detail.

**Stakeholders in datamanagement en hun verantwoordelijkheden**
De vele stakeholders hebben deels eigen en deels gezamenlijke verantwoordelijkheden voor aspecten van datamanagement[1]:

- de hoofdonderzoeker – verantwoordelijk voor de data en het datamanagement; eerstverantwoordelijke voor een DMP, tenzij anders afgesproken met de projectleider;
- overige onderzoekers, onderzoeksassistenten en/of datamanagers – betrokken bij het dagelijkse datamanagement in de praktijk;
- senioronderzoekers en onderzoekscoördinatoren – voorbeeldgedrag, mentorschap en supervisie van (aankomende en jonge) onderzoekers;
- directie en/of bestuur van de onderzoeksinstelling – databeleid opstellen en monitoren; databewustzijn bevorderen bij alle wetenschapsbeoefenaren;
- front-office van de onderzoeksinstelling, met onder meer bibliotheek-, ICT- en juridische staf – toegang verschaffen tot externe data en tools, veilige opslag en toegang tot data; expertise op juridisch en ethisch gebied, datacitatie, metadata, toegang en licenties, voorwaarden van financiers; databewustzijn bevorderen (zie verder paragraaf 5.2);
- onderzoeksfinanciers – goed datamanagement stimuleren; databewustzijn bevorderen; investeren in data-infrastructuur;
- projectpartners bij onderzoeksinstellingen en bedrijven – zie ‘overige onderzoekers’;
- wetenschappelijke uitgevers – bij ingediende en/of gepubliceerde artikelen voorwaarden stellen aan de beschikbaarheid van de data die aan de publicatie ten grondslag liggen; *persistent identifiers* toekennen aan publicaties en (laten) verwijzen naar de bijbehorende data;
- archieven en repositories voor onderzoeksdata – langetermijntoegang bieden tot de data; *persistent identifiers* toekennen aan de data; data vindbaar maken (*data discovery service*);
- datagebruikers, zoals onderzoekers, overheid, journalisten, bedrijven, *citizen scientists*, docenten, studenten – data conform toegangsrechten en licenties gebruiken en verspreiden; naar wetenschappelijk gebruik correct refereren aan gebruikte data.

Hoewel het in theorie mogelijk is dat het beheer in de gehele data *lifecycle* door één organisatie uitgevoerd wordt, leert de praktijk dat het efficiënter is om de verschillende fasen door verschillende organisaties/instituten uit te laten voeren – een instelling kan immers niet specialist zijn in alles of alle fasen in haar missie hebben staan. Er is in Nederland dan ook een landschap ontstaan waarin verschillende organisaties samenwerken om goed datamanagement tot stand te brengen. Generieke diensten voor langetermijnarchivering, het toekennen van *persistent identifiers* aan datasets en training voor ondersteuners worden geleverd door centrale organisaties als Research Data Netherlands (RDNL)[2], het verband van universiteitsbibliotheken en de Koninklijke Bibliotheek (UKB) en SURF.

Ook buiten dit verband spelen de Nederlandse universiteitsbibliotheken (UB's) en in toenemende mate ook de hogeschoolbibliotheken hierin een rol (zie paragraaf 5.2 voor de activiteiten die zij uitvoeren). In het zogenoemde Frontoffice/Backoffice-model van het RDNL-consortium worden de UB's beschouwd als frontoffice: zij staan in direct contact met de onderzoeker en kunnen er door advisering voor zorgen dat de onderzoeksdata in één van de RDNL-archieven (de backoffices) worden ondergebracht[3]. De betreffende medewerkers van de bibliotheek begeleiden de onderzoekers van de instellingen, zodat zij kunnen voldoen aan wetgeving en het beleid van financiers, van hun instelling of van een tijdschrift waarin zij (willen) publiceren.

Samenwerking, kennisuitwisseling en het delen van *best practices* vinden onder andere plaats in de Werkgroep Research Data van het UKB.[4] De Werkgroep heeft het UKDA *lifecyclemodel* naar drie fases vertaald: start, tijdens en na het onderzoek. Voor elke fase is geïdentificeerd wat de activiteiten zijn waarvoor de UB ondersteuning kan bieden.[5]

Het in 2015 opgerichte Landelijk Coördinatiepunt Research Data Management (LCRDM) faciliteert een landelijke aanpak van research datamanagement in Nederland. Op verzoek van de universiteitenkoepel VSNU vervult

1 Dit overzicht is deels gebaseerd op (OpenAIRE, 2015).
2 http://researchdata.nl/
3 Een overzicht van de frontoffices aan de Nederlandse universiteiten is te vinden op: https://www.edugroepen.nl/sites/RDM_platform/Lists/RDM%20bij%20universiteiten%20in%20Nederland/AllItems.aspx; informatie over een eerste bijeenkomst van de hogeschoolbibliotheken is te vinden op http://www.shb-online.nl/kennisdeling/publicaties/presentaties-themadag-data-management-16-juni-2016/.
4 https://www.ukb.nl/research-data
5 Zie https://www.ukb.nl/sites/ukb/files/docs/Opdracht-wg-ukd-data-2015.pdf

SURFsara – als onderdeel van het innovatieprogramma Duurzame data[1] – deze coördinerende rol, in samenwerking en afstemming met de Nederlandse onderwijs- en onderzoeksinstellingen, de RDNL-partners, de *special interest group* Research Data[2] en de UKB Werkgroep Research Data. Vijf essentiële thema's zijn geïdentificeerd, die in landelijke werkgroepen worden aangepakt: faciliteiten en data-infrastructuur; juridische aspecten en zeggenschap; financiën; onderzoeksondersteuning en advies; en bewustwording van de voordelen van RDM. Het doel is dat datamanagement een vanzelfsprekend onderdeel vormt van de manier van denken en doen aan de Nederlandse universiteiten en onderzoeksinstellingen. Belangrijke te realiseren doelen zijn: synergie tussen beleid, ICT en onderzoeksondersteuning; verbinding tussen experts van onderzoeksuitvoerende organisaties, facilitaire organisaties en onderzoeksfinanciers; en bestuurlijke verankering van het data(management)beleid.

**Databeleid**

De onderwerpen in de onderstaande lijst zijn ontleend aan data(management) beleid van een kleine twintig instellingen in binnen- en buitenland. Ook databeleid is een thema met vele variaties; sommige instellingen hebben aan drie pagina's voldoende, terwijl andere databeleidsdocumenten tientallen pagina's beslaan. Breed gedragen is wel de aanpak dat instellingsbreed, dat wil zeggen op het niveau van het College van Bestuur of de Raad van Bestuur, het beleidskader wordt vastgesteld en dat hierin de verantwoordelijkheid voor concretisering en uitvoering wordt gedelegeerd aan bijvoorbeeld faculteiten en onderzoeksscholen (zie (Aerts en Doorn, 2016) voor een pleidooi voor het opstellen van disciplinespecifieke protocollen).
De vele begrippen uit die beleidsdocumenten zijn enigszins gegroepeerd, maar andere indelingen zijn zeker mogelijk, zoals men ook nog onderwerpen zou kunnen toevoegen. Deze *long list* vormt een goede basis om te bespreken wat er in het data-instellingsbeleid aan de orde zou kunnen of moeten komen.

**Doel**

1. Doel van het data(management)beleid, zoals herhaalbaarheid van onderzoek, controleerbaarheid of betere zichtbaarheid van de resultaten van de instelling.
2. Visie van de bestuurder en/of de organisatie op data en/of databeleid en/of datamanagement.

1 https://www.surf.nl/innovatieprojecten/duurzame-data.html
2 https://www.surf.nl/themas/onderzoek/management-van-onderzoeksdata/special-interest-group-sig-research-data/index.html

3. Open Access tot publicaties en data, 'Open als het kan, Restricted als het moet', de omgang met gevoelige data, eventuele versleuteling van data, interoperabiliteit/uitwisselbaarheid van bestandsformaten, een eventuele embargoperiode.
4. Het bevorderen en onderhouden van vertrouwen in de wetenschap, in de eigen instelling en bij derden, bijvoorbeeld door data waar mogelijk te archiveren in gecertificeerde *Trustworthy Digital Repositories*.

**Kwaliteit van de data**

1. De herkomst ofwel *provenance* van de data (zijn de data bijvoorbeeld aangeschaft bij een externe dataleverancier, hoe zijn ze verzameld, hoe zijn ze tijdens het onderzoek bewerkt), betrouwbaarheid van de data, FAIR-principes, metadata.
2. Minimale gedragscodes, de wetenschappelijke context, relevante wet- en regelgeving binnen en buiten de instelling, academische normen, de eventuele rol van de (medisch-)ethische toetsingscommissie.

**Middelen**

1. Zeggenschap over de data: eventuele voorwaarden voor hergebruik van data die bij de instelling worden gegenereerd, inzage in de data door een niet-reguliere partij zoals een tijdschriftreviewer of een integriteitscommissie, hoe moet men eigen en andermans data citeren, afspraken bij contractonderzoek.
2. Opslag van de data tijdens het onderzoek en archivering na afloop: de data veiligstellen voordat medewerkers vertrekken, veilige opslag, back-up en versiebeheer; *replication package* als de eenheid van archivering (alles wat nodig is om een studie te repliceren), minimale en/of maximale bewaartermijnen, zijn er selectiecriteria, moeten/mogen er data vernietigd worden; zie ook de punten onder 'Kwaliteit van data'.
3. Ondersteuning van de onderzoekers door bijvoorbeeld training, datamanagers, hulp bij het schrijven van DMP's, het 'opvoeden' van jonge onderzoekers.
4. De financiering van de technische en menselijke infrastructuur voor datamanagement, waar veel punten van deze lijst binnen vallen.

**Verantwoording**

1. De erkenning en zichtbaarheid vergroten, zowel van onderzoekers als van bijvoorbeeld software-ontwikkelaars en *data supporters*; datasets correct (laten) citeren, een datacitatie-index stimuleren als pendant voor de H-index voor publicaties.

2. Naleving van dit databeleid in ontwikkel- of beoordelingsgesprekken ("Waar hebt u de data die bij deze publicatie horen gearchiveerd?"); andere vormen van toezicht, zoals interne audits.
3. Kosten van datamanagement begroten in project- of afdelingsplannen; kosten bijhouden op één of meer niveaus in de organisatie; zie ook het laatste punt onder 'Middelen'.
4. De verantwoordelijkheden inzake data van de onderzoeker, de supervisor, de instelling en van andere stakeholders.
5. Periodiek evalueren van dit databeleid, op verschillende niveaus in de organisatie.

### *5.2 Datasupport aan de instellingen*

Hoe de instellingen voor hoger onderwijs en onderzoek de ondersteuning van datamanagement invullen, verschilt per instelling en is nog zeer in beweging.[1] De diversiteit komt bijvoorbeeld tot uiting in de naamgeving: de verantwoordelijke medewerkers worden aangeduid als 'data librarian', 'data steward' of 'RDM-support'. De activiteiten die deze medewerkers (voornamelijk gepositioneerd binnen de universiteitsbibliotheken) ontplooien, zijn echter zeker onderling vergelijkbaar en sluiten aan op bestaande werkzaamheden en expertise in de UB's: beschrijving/metadatering van data, *access control* en het vinden en hergebruiken van data. De UB's werken in veel gevallen samen met de universitaire ICT-dienst. Aanvullende expertise wordt opgedaan in beschikbare cursussen als 'Essentials 4 Data Support'[2] en in de bovengenoemde samenwerkingsverbanden van UKB, LCRDM en SURF.

Een rondje langs de RDM-webpagina's van de Nederlandse universiteiten laat zien dat de UB's ondersteuning bieden voor de gehele *lifecycle*: van support bij het schrijven van een RDM-plan tot het bemiddelen tussen onderzoekers en langetermijnarchieven. Het volgende tekstkader somt per fase diverse activiteiten op. Activiteiten die niet gebonden zijn aan een specifieke fase zijn: het trainen van onderzoekers en ondersteuners, het aanbieden van disciplinespecifieke contactpersonen en ondersteuning ('maatwerk') en het geven van algemeen bruikbare informatie over RDM (checklists, begrippenlijsten, FAQs, links), geïllustreerd met *good practices* aan de instelling.

1 In het onderstaande wordt verder gesproken over 'universiteitsbibliotheken', aangezien aan alle Nederlandse universiteiten ondersteuning van RDM plaatsvindt. Binnen de HBO-instellingen groeit de aandacht voor (onderzoeks)data ook, maar is de aanwezige ondersteuning nog minder wijdverbreid.

2 http://datasupport.researchdata.nl/nl/

Een goed voorbeeld van het ondersteunen van onderzoekers – en eveneens van het eerdergenoemde Frontoffice/Backoffice-model – is de dienst die de Radboud Universiteit Nijmegen aanbiedt: het lokale RDM-systeem (genaamd 'Research Information Services') biedt onderzoekers de mogelijkheid publicaties en datasets gelijktijdig te registreren en te bewaren (Simons et al., 2016). Het systeem is gekoppeld aan het langetermijnarchief EASY van DANS. De medewerkers van de Radboud Universiteit bieden ondersteuning bij het gebruik van RIS en zorgen dat de datasets in het lokale, universitaire systeem voldoen aan de preserveringseisen van EASY. Onderzoekers hoeven maar één systeem te gebruiken om zowel aan institutionele eisen als aan eisen van de financier te voldoen.

**Activiteiten per fase van de onderzoekscyclus**
Onderstaande lijst is samengesteld uit activiteiten die de UB's op hun websites beschrijven als dienstverlening inzake datamanagement. We hebben ze per fase gegroepeerd.

**Fase 'Data genereren':**
- Ondersteunen bij het schrijven van RDM-plannen en projectaanvragen
- Aanbieden van templates van RDM-plannen
- Informeren en adviseren over subsidievoorwaarden en datagerelateerde eisen van financiers (NWO, ZonMw, Horizon2020)
- Informeren en adviseren over institutioneel databeleid en gedragscode
- Ondersteunen bij het zoeken naar bestaande data/verwijzingen naar systemen en zoekmachines (bijvoorbeeld NARCIS, DataCite, B2Find)
- Ondersteunen bij juridische (bijvoorbeeld WBP, risicoclassificatie van data) en ethische vraagstukken en het anonimiseren van data
- Adviseren over kosten van datamanagement en hoe die in een projectaanvraag te begroten

**Fase 'Data bewerken':**
- Aanbieden van (data)lab-omgevingen en Virtual Research Environments (VREs)
- Aanbieden van lokale opslag of repository voor opslaan en delen tijdens onderzoek (bijvoorbeeld DataverseNL, Beehub, SURFdrive, Figshare, SharePoint)
- Aanbieden van software-management systemen
- Adviseren over de (technische) beveiliging van data
- Verwijzen naar disciplinespecifieke opslag
- Aanbieden *persistent identifier* services (bijvoorbeeld handle, DOI)
- Adviseren over tijdelijke opslag

- Adviseren over metadatering en standaarden
- Ondersteunen van het documenteren van data

**Fase 'Data analyseren':**
- Aanbieden visualisatietools en infrastructuur
- Aanbieden rekenkracht (op campus of bij bijvoorbeeld SURFsara)
- Aanbieden of overzicht geven van relevante software (bijvoorbeeld lab journals)
- Ondersteunen van het documenteren van data

**Fase 'Data archiveren':**
- Adviseren over duurzame archivering
- Aanbieden van eigen TDR (op dit moment alleen in het geval van 4TU)
- Verwijzen of koppelen naar nationaal of internationaal TDR
- Aanbieden *persistent identifier* services (handle, DOI)
- Archiveren van niet-digitale data

**Fase 'Toegang verschaffen tot data':**
- Adviseren over toegang, Open Access, Restricted Access, eventueel embargo
- Registreren van datasets in lokaal CRIS-systeem
- Informeren en adviseren over eisen van financiers en uitgevers
- Informeren en adviseren over licenties
- Verwijzen naar relevante *data journals*

**Fase 'Data hergebruiken':**
- Aanbieden zoekmachines (lokaal, NARCIS et cetera)
- Informeren en adviseren over datacitatie
- Aanbieden cursusonderdelen voor research masterstudenten en PhD's

Binnen de RDM-activiteiten in de instellingen wordt ook veel aandacht geschonken aan het *waarom* van goed datamanagement, gekoppeld aan een uitleg van het datamanagementbeleid van de universiteit, wanneer aanwezig. Het vergroten van het besef onder onderzoekers van wat goed RDM is, is een belangrijk gebied. Voor onderzoekers en informatiemedewerkers worden dan ook datamanagementcursussen, -trainingen en -workshops aangeboden.

Door het bewustzijn over research data management te vergroten – in het gehele veld: van financiers, naar onderzoeker, naar ondersteuners – wordt nu en in de toekomst goed omgegaan met onderzoeksdata. Meer data zal beter herbruikbaar beschikbaar komen, meer data zal hergebruikt worden. Uiteindelijk leiden alle vormen van hergebruik tot de in de introductie genoemde oplossingen van "*today's Grand Challenges, such as climate change and energy supply*".

## 6. Literatuur


Aerts, P. en Doorn, P. (2016) *Data doordacht - Strategie voor de omgang met data en software in de wetenschap*. DANS. https://dans.knaw.nl/nl/over/organisatie-beleid/informatiemateriaal

Borgman, B. (2015) *Big data, little data, no data – Scholarship in a networked world*. MIT Press.

DANS (2015a) *Datamanagementplan voor wetenschappelijk onderzoek*. https://dans.knaw.nl/nl/over/organisatie-beleid/informatiemateriaal

DANS (2015b) *Samen data delen – Strategienota DANS 2015-2020*. https://dans.knaw.nl/nl/over/organisatie-beleid/informatiemateriaal

Dillo, I. en De Leeuw, L. (2014) Het Data Seal of Approval: keurmerk voor duurzame en betrouwbare databewaarplaatsen. *Handboek Informatiewetenschap*. de Jong, A. S. M., van Trier, G. M., Sieverts, E. & Koren, M. (eds.). Alphen aan de Rijn: Vakmedianet, Vol. aanvulling 69, p. IV B 630 1-29 29 p. IV B 630

European Commission (2016) *H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020*. Version 3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

High Level Expert Group on Scientific Data (2010) *Riding the wave: How Europe can gain from the rising tide of scientific data*. European Union. http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

KNAW (2012) *Zorgvuldig en integer omgaan met wetenschappelijke onderzoeksgegevens.* Koninklijke Nederlandse Akademie van Wetenschappen. ISBN 978-90-6984-655-2 | https://www.knaw.nl/nl/actueel/publicaties/zorgvuldig-en-integer-omgaan-met-wetenschappelijke-onderzoeksgegevens

NWO (2016) *Formulier Datamanagementplan*. Versie september 2016. http://www.nwo.nl/documents/nwo/datamanagement%5B2%5D/formulier-nwo-datamanagementplan

OpenAIRE (2015) *Research data management - Briefing paper for National Open Access Desks*. https://www.openaire.eu/briefpaper-rdm-infonoads

Simons, E., Jetten, M., van Berchum, M., Messelink, M., Schoonbrood, H., Wittenberg, M. (2016) The important role of CRIS's for registration and archiving of research data *Proceedings of the 13th International Conference on Current Research Information Systems*. EuroCRIS. http://hdl.handle.net/11366/524.

VSNU (2014) *De Nederlandse Gedragscode Wetenschappelijk Onderzoek - Principes van goed wetenschappelijk onderwijs en onderzoek*. Vereniging van Universiteiten (VSNU). http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/Code_wetenschapsbeoefening_2004_(2014).pdf

Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, article number 160018 http://dx.doi.org/10.1038/sdata.2016.18

## 7. Weblinks

De hier genoemde organisaties en initiatieven bieden discipline- en instellingsoverstijgende informatie over goed beheer van onderzoeksgegevens.

*Data Seal of Approval*
http://www.datasealofapproval.org/

*Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), Wetenschappelijke integriteit*
https://www.knaw.nl/nl/thematisch/ethiek/wetenschappelijke-integriteit

*Landelijk Coördinatiepunt Research Data Management (LCRDM)*
https://www.lcrdm.nl

*Nationale Coalitie Digitale Duurzaamheid (NCDD), persistent identifiers*
http://www.ncdd.nl/pid/

*NWO, Datamanagementprotocol*
http://www.nwo.nl/beleid/open+science/datamanagement

*Research Data Netherlands (RDNL), Cursus 'Essentials 4 Data Support'*
http://datasupport.researchdata.nl/nl/