



Nationale Verkenning Digitale Duurzaamheid

Inputnotitie sector wetenschap

René van Horik/DANS

**Nationale Coalitie Digitale Duurzaamheid
1 juli 2009**



Dit document vormt een inputnotitie bij het rapport 'Toekomst voor ons digitaal geheugen: duurzame toegang tot digitale informatie in Nederland', NCDD, 1 juli 2009, beschikbaar op <http://www.ncdd.nl/activiteiten-natverkenning.php>.

De rapportages zullen besproken worden tijdens de nationale werkconferentie *Toekomst voor ons digitaal geheugen* op 18 september 2009 in Den Haag; meer informatie en aanmelden op <http://www.ncdd.nl/werkconferentie2009.php>.

1 juli 2009

NCDD – Nationale Coalitie Digitale Duurzaamheid
Postbus 90407
2509 LK Den Haag
<http://www.ncdd.nl>
info@ncdd.nl

Het rapport werd gefinancierd door het ministerie van Onderwijs, Cultuur en Wetenschap, directie Onderzoek en Wetenschapsbeleid.

De Sector Wetenschap

1	De sector wetenschap en digitale duurzaamheid	7
1.1	Wetenschap in Nederland.....	7
1.2	Digitale resultaten van wetenschappelijk onderzoek.....	8
1.2.1	Wetenschappelijke publicaties	8
1.2.2	Octrooien en patenten.....	9
1.2.3	Onderzoeksdata	10
1.2.4	Verrijkte publicaties	11
1.3	Digitale duurzaamheid in de wetenschap.....	12
1.3.1	Het belang van digitale duurzaamheid	12
1.3.2	Archivering van digitale onderzoeksdata	12
1.3.3	De periode 1990 - 2000.....	14
1.3.4	Drie belangrijke pijlers	15
1.3.5	Scenario's voor duurzame toegang tot onderzoeksdata	16
1.3.6	De onderzoekscyclus.....	17
1.3.7	Verantwoordelijkheden	19
2	Kaders en codes in relatie tot duurzame toegang tot onderzoeksdata	21
2.1	Kwaliteitszorgsysteem voor wetenschappelijk onderzoek	21
2.2	Gedragscode wetenschapsbeoefening	22
2.3	ICT-strategie van de koepelinstellingen.....	22
2.4	Toetsingskaders	23
2.4.1	Data Seal of Approval.....	23
2.4.2	DRAMBORA.....	24
2.4.3	TRAC	25
2.4.4	nestor	25
2.5	Internationale beleidskaders.....	25
2.5.1	Open Access	25
2.5.2	OECD principles for access to research data.....	26
2.5.3	ESFRI-roadmap	26

3	Infrastructuur voor duurzame opslag van en toegang tot digitale onderzoeksobjecten	28
3.1	Instellingen en organisaties.....	28
3.1.1	Landelijke voorzieningen voor digitale archivering	28
3.1.2	Institutionele digitale bewaarplaatsen.....	30
3.1.3	Universitaire repositories	31
3.1.4	Registerdata	31
3.1.5	Enkele onderzoeksinfrastructuren	32
3.2	Wetenschap en ICT	34
3.3	Financiële aspecten.....	35
4	Conclusies en aanbevelingen.....	37

Samenvatting

Deze inputnotitie bevat de resultaten van de nationale verkenning digitale duurzaamheid, sector "wetenschap", die uitgevoerd is in de eerste helft van 2009 door de Nationale Coalitie Digitale Duurzaamheid (NCDD). Doel van de verkenning is een beeld te schetsen van de langetermijntoegang tot digitale dataobjecten van belang voor de wetenschap, zoals computerbestanden en computerprogramma's. Daarnaast wordt geïnventariseerd welke richtingen ontstaan om de digitale duurzaamheid van wetenschappelijke dataobjecten te verbeteren. De langetermijntoegang tot deze objecten wordt bedreigd door het ontstaan van nieuwe hardware, door het verouderen van bestandsformaten en software, en door het ontbreken van goede documentatie.

De verkenning bestond enerzijds uit het afnemen en analyseren van een serie semigestructureerde interviews van mensen die een rol spelen in de wetenschappelijke informatievoorziening in Nederland en anderzijds uit literatuuronderzoek: het verwerken van een aantal nationale en internationale publicaties die betrekking hebben op de rol en functie van informatietechnologie en digitale onderzoeksdata in de wetenschap.

Het gebruik van databestanden kent binnen de wetenschap verschillende tradities. Binnen de sociale wetenschappen kent men al vanaf de jaren zestig van de vorige eeuw het gebruik van computers bij het verrichten van survey-onderzoek. Binnen de geesteswetenschappen zijn de economisch en sociale historici en onderzoekers op het gebied van taal en tekst de eerste gebruikers van databestanden. Dit geldt ook voor de archeologen. Binnen de exacte wetenschappen bestaat een lange traditie op datagebied, waarbij veel zaken pragmatisch worden aangepakt en er verschillende producten ontwikkeld worden die hun weg vinden naar de commerciële ICT wereld. Voorbeelden zijn het Linux besturingssysteem en het HTML-protocol.

Bij digitale duurzaamheid in de wetenschap ligt de nadruk eerder op het beheer van en toegang tot relevante digitale onderzoeksobjecten, dan op de archivering in de traditionele zin van het woord.

Dit rapport bestaat uit de volgende hoofdstukken:

Hoofdstuk 1. "De sector wetenschap"

In dit hoofdstuk wordt een overzicht gegeven van de organisaties die een rol spelen in het wetenschappelijk onderzoek in Nederland. Verder is er in dit hoofdstuk aandacht voor de digitale objecten die van belang zijn voor de wetenschap en waarvoor maatregelen genomen moeten worden ten behoeve van de bruikbaarheid op lange termijn. Het derde onderdeel van dit hoofdstuk behandelt de totstandkoming van databewaarplaatsen en de belangrijkste standaarden en richtlijnen.

Hoofdstuk 2. "Kaders en codes"

Dit hoofdstuk bevat een overzicht van de strategieën van wetenschappelijke koepelorganisaties voor zover ze een relatie hebben met digitale onderzoeksobjecten, en kaders om de duurzaamheid ervan te toetsen.

Hoofdstuk 3. "Infrastructuur voor duurzame toegang tot onderzoeksdata"

In dit hoofdstuk wordt de bestaande infrastructuur beschreven voor de archivering en beschikbaarstelling van digitale wetenschappelijke objecten in Nederland. Het tweede deel van dit hoofdstuk gaat in op het onder invloed van de ICT veranderende landschap voor wetenschappelijk onderzoek. Ten slotte is er in dit hoofdstuk aandacht voor de financiële aspecten van digitale archivering.

Hoofdstuk 4. "Conclusies en aanbevelingen".

Een belangrijke constatering is dat er in Nederland weinig organisaties zijn die expliciet en structureel activiteiten ontplooiën om de langetermijnarchivering van wetenschappelijke digitale objecten te verzorgen. Er zijn daarentegen wel een vrij groot aantal repositories die de toegang tot een groot aantal digitale wetenschappelijke objecten mogelijk maken, maar die niet ingericht zijn op langetermijnbewaring. Er dient meer aandacht te komen voor de langetermijnarchivering van digitale wetenschappelijke objecten. Hierbij is het van belang in samenwerking te komen tot een gemeenschappelijke basisinfrastructuur voor digitale archivering en voor de manier waarop digitale wetenschappelijke objecten kunnen worden hergebruikt.

1 De sector wetenschap en digitale duurzaamheid

Dit hoofdstuk bestaat uit drie delen. Allereerst wordt beknopt de organisatiestructuur van de sector wetenschap in Nederland beschreven. Daarna worden de belangrijkste digitale objecten behandeld die een rol spelen in het wetenschapsbedrijf. Dit zijn digitale publicaties, octrooien en patenten, onderzoeksdata en het recentelijk geïntroduceerde concept "verrijkte publicatie". Het derde deel van dit hoofdstuk beschrijft kort de ontwikkelingen betreffende de digitale duurzaamheid in de wetenschap en gaat in op de belangrijkste pijlers onder de initiatieven die een oplossing trachten te bieden voor de duurzame opslag van en toegang tot digitale onderzoeksobjecten.

1.1 Wetenschap in Nederland

Op het terrein van het wetenschappelijk onderzoek in Nederland zijn vele personen en organisaties actief. Volgens de Nederlandse Onderzoek Databank (NOD), een openbare online databank met informatie over wetenschappelijk onderzoek, onderzoekers en onderzoeksinstituten in Nederland, zijn er in Nederland ca. 7.600 hoogleraren en universitair hoofddocenten, bijna 40.000 onderzoekers, 750 universitaire en niet-universitaire onderzoeksinstellingen en 120 onderzoeksscholen. De NOD bevat 20.000 beschrijvingen van lopende projecten en 18.000 beschrijvingen van afgesloten onderzoek¹. Wetenschap is bij uitstek een internationale aangelegenheid, waardoor internationale ontwikkelingen veel invloed hebben op de situatie in Nederland.

Nederland heeft vijftien algemene universiteiten en vier bijzondere universiteiten op levensbeschouwelijke grondslag. Aan acht universiteiten in Nederland is een academisch ziekenhuis verbonden. Een derde van het wetenschappelijk onderzoek in Nederland wordt door universiteiten uitgevoerd. Circa 80% van de financiering is afkomstig van de overheid. Het merendeel van het universitair wetenschappelijk onderzoek is ondergebracht in een van de ruim 120 onderzoeksscholen. Zes van deze onderzoeksscholen, alle op het gebied van de exacte wetenschappen, zijn erkend als toponderzoeksscholen. Het systeem van onderzoeksscholen is uniek in de wereld. Het buitenland kent geen vergelijkbaar systeem.

De NOD telt ca. 4.500 onderzoeksinstituten die meestal verbonden zijn aan een of meerdere universiteiten of aan een van de drie grote koepelorganisaties: de Koninklijke Nederlandse Academie van Wetenschappen (KNAW), de Organisatie voor Wetenschappelijk Onderzoek (NWO) en de Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO).

De KNAW beheert 19 instituten voor fundamenteel en strategisch onderzoek op de gebieden levenswetenschappen en geestes- en sociale wetenschappen. Er zijn circa 1.300 mensen werkzaam bij deze instituten. NWO richt zich van oudsher op zuiver wetenschappelijk onderzoek. Onder NWO vallen tien onderzoeksinstituten, twee stichtingen (waaronder de Stichting Nationale Computerfaciliteiten, NCF) en drie tijdelijke aansturingorganen (waaronder het regieorgaan voor ICT-onderzoek en -innovatie, ICTRegie). Bij NWO als koepelorganisatie zijn ca. 2.400 mensen werkzaam. Daarnaast zijn ca. 3.600 fte's met subsidie van NWO in dienst bij een universiteit of onderzoeksinstelling. TNO is een kennisorganisatie voor bedrijven, overheden en maatschappelijke organisaties. Met meer dan 5.000 medewerkers in dienst werkt TNO aan de ontwikkeling en toepassing van kennis in de vorm van contractresearch en advisering. Daarnaast houdt TNO zich bezig

¹ Zie <<http://www.onderzoekinformatie.nl/nl/oi/landschap>> [bezocht 18 mei 2009]

met testen en certificeren van producten en diensten, en het geven van onafhankelijk kwaliteitsoordeel. De TNO-organisatie is opgebouwd rond vijf kerngebieden: kwaliteit van leven, defensie en veiligheid, industrie en techniek, bouw en ondergrond en informatie- en communicatietechnologie. Daarnaast omvat TNO 30 kenniscentra waarin TNO op structurele basis samenwerkt met de universiteiten.

Toepassingsgericht wetenschappelijk onderzoek vindt plaats bij het bedrijfsleven, de overheid en hogescholen. Vooral grote bedrijven in de industriesector verrichten veel onderzoek. Daarnaast hebben ministeries instellingen die onderzoek uitvoeren, zoals het Koninklijk Nederlands Meteorologisch Instituut (KNMI) dat verbonden is aan het ministerie van Verkeer en Waterstaat. Het Centraal Bureau voor de Statistiek (CBS) valt onder het ministerie van Economische Zaken.

De instrumenten die de overheid heeft om het wetenschappelijk onderzoekstelsel te reguleren zijn wet- en regelgeving, financiële middelen en bestuurlijk overleg. Elk ministerie is verantwoordelijk voor het onderzoeksbeleid op zijn of haar specifieke terrein. De minister van Onderwijs, Cultuur en Wetenschap is binnen de rijksoverheid verantwoordelijk voor de coördinatie van het wetenschapsbeleid en voor het goed en doelmatig functioneren van het onderzoeksbestel in Nederland. De minister van Economische Zaken coördineert het technologiebeleid.

1.2 Digitale resultaten van wetenschappelijk onderzoek

Van oudsher vormen publicaties, octrooien en patenten de belangrijkste resultaten van wetenschappelijke onderzoek. Met de opkomst van de informatietechnologie zijn deze resultaten beschikbaar gekomen in digitale vorm. Daarnaast is er een nieuw soort digitaal dataobject ontstaan dat van belang is voor de wetenschap, namelijk digitale onderzoeksbestanden. In deze paragraaf worden deze digitale resultaten van wetenschappelijk onderzoek nader toegelicht.

1.2.1 Wetenschappelijke publicaties

De resultaten van wetenschappelijk onderzoek worden van oudsher opgenomen in publicaties, zoals proefschriften en artikelen in wetenschappelijke tijdschriften. Het *peer review* proces, de collegiale toetsing, bepaalt in hoge mate de kwaliteit van wetenschappelijke publicaties. Daarnaast hebben bepaalde tijdschriften in de loop der tijd een reputatie gekregen die van invloed is op de waardering van het wetenschappelijk onderzoek. Een publicatie in *Nature* of *The Lancet*, bijvoorbeeld, heeft een positieve invloed op de *citation index* van de wetenschapper.

Alle belangrijke wetenschappelijke tijdschriften zijn digitaal beschikbaar, hetzij via een abonnement bij een uitgever, hetzij als Open Access tijdschrift. Voor de wetenschappers is de toegang tot de wetenschappelijke literatuur van groot belang in verband met de controleerbaarheid van de uitkomsten van het onderzoek en de mogelijkheid te kunnen refereren aan het onderzoek, nu en in de toekomst. Het e-Depot van de Koninklijke Bibliotheek is een belangrijke faciliteit voor de duurzame opslag van deze digitale wetenschappelijke tijdschriften². Dit e-Depot bevat een groot aantal digitale publicaties waarvan de langetermijntoegang vastgelegd is in afspraken en contracten. De belangrijkste pijlers onder de digitale duurzaamheid van wetenschappelijke publicaties zijn het toepassen van het als duurzaam beoordeelde PDF/A³ bestandsformaat en het

² Zie: <<http://www.kb.nl/dnp/e-depot/dm/dm.html>> [bezocht 2 juni 2009]

³ Meer informatie over het PDF/A formaat kan bijvoorbeeld gevonden worden op de site van het "PDF/A competence center, zie: <<http://www.pdfa.org/>> [bezocht 2 juni 2009]

inrichten van een betrouwbare opslagfaciliteit met de expliciete missie om digitale publicaties voor de lange termijn te archiveren.

Ook het Noord-Amerikaanse Portico initiatief⁴ heeft als doel elektronische wetenschappelijke literatuur duurzaam op te slaan. Zowel de KB als Portico maken de literatuur toegankelijk voor aangesloten organisaties in het geval er een "trigger event" optreedt. Dit houdt in dat als de wetenschappelijke literatuur niet langer op een andere wijze toegankelijk is (bijvoorbeeld via de uitgever) dat het archief dan zorgt voor de toegang. Zowel het e-Depot en Portico publiceren via hun website lijsten van tijdschrifttitels die opgenomen zijn in de respectievelijke digitale depots. Portico en het e-Depot van de KB archiveren voor een gedeelte dezelfde titels als extra beveiliging in het geval één van beide conserveringsstrategieën problemen geeft.

Het e-Depot bevat ook tijdschriften van Open Access uitgevers. In 2009 is een contract gesloten met de "Directory of Open Access Journals" om ook die tijdschriften duurzaam te archiveren⁵.

1.2.2 Octrooien en patenten

Uitkomsten van toegepast wetenschappelijk onderzoek worden vastgelegd in patenten en octrooien. Dit zijn tijdelijk door de overheid verstrekte monopolies op een uitvinding. Bij het *European Patent Office* (EPO)⁶ worden patenten en octrooien geregistreerd. Sinds 1988 werkt EPO aan een digitale infrastructuur om patenten te ontsluiten. In 2008 bestond deze database uit 16 miljoen documenten⁷. De digitale duurzaamheid van het informatiesysteem dat het intellectueel eigendom van wetenschappelijke resultaten garandeert, is tijdens het in 2007 uitgevoerde *Scenario's for the future* project aan bod gekomen. Het *blue skies* scenario gaat uit van een situatie waarin de technologie de belangrijkste factor is bij het beheer en gebruik van patenten. Bijgaand citaat bevat een beschrijving van dit scenario:

In the end, the patent system responds to the speed, interdisciplinarity and complex nature of the new technologies by abandoning the one-size-fits-all model: the former patent regime still applies to classic technologies while the new one uses other forms of intellectual property protection, such as the license of rights. The patent system increasingly relies on technology, and new forms of knowledge search and classification emerge⁸.

Ten behoeve van de onlinetoegang tot octrooien en patenten heeft het EPO de "Esp@cenet portal" opgezet⁹. Dit informatiesysteem maakt 60 miljoen octrooipublicaties uit 80 landen toegankelijk. Het oudste octrooi dat online toegankelijk is dateert uit 1827 en betreft het destilleren van alcohol.

⁴ Zie: <<http://www.portico.org>> [bezoekt 2 juni 2009]

⁵ Zie: <<http://www.doaj.org>> [bezoekt 2 juni 2009]

⁶ Zie: <<http://www.epo.org>> [bezoekt 20 mei 2009]

⁷ Zie: <<http://www.epo.org/topics/patent-system/patent-information.html>> [bezoekt 20 mei 2009]

⁸ Zie: <<http://www.epo.org/topics/patent-system/scenarios-for-the-future.html>> (Scenarios for the future. P. 10) [bezoekt 20 mei 2009]

⁹ Zie <<http://www.espacenet.com>> [bezoekt 20 mei 2009]

In 2007 heeft het EPO ruim één miljard euro aan inkomsten afkomstig uit vergoedingen voor de procedures om patenten te registreren en beheren¹⁰. Er zijn circa 6.500 mensen werkzaam bij het EPO, waarbij de online dienstverlening het hart van de organisatie vormt¹¹. Het EPO spreekt nergens expliciet over de digitale duurzaamheid van de elektronische patent- en octrooi-informatie. Het grote belang van betrouwbare informatie op dit gebied, zowel van patenten die al lange tijd geleden geregistreerd zijn, als van patentinformatie die momenteel geregistreerd wordt, rechtvaardigt de veronderstelling dat de registratie, de opslag, het beheer en de toegang tot patentgegevens duurzaam geregeld zijn; in dit rapport zal niet nader worden ingaan op de digitale duurzaamheid van patenten en octrooien. De nadruk zal liggen op publicaties en onderzoeksdata.

1.2.3 Onderzoeksdata

Naast digitale publicaties en octrooien speelt een groot aantal andere digitale objecten een rol in het wetenschapsbedrijf. In algemene zin wordt vaak over "onderzoeksdata" gesproken. Onderzoeksdata kunnen het resultaat zijn van verschillende soorten processen en verschillende doelen dienen. Er is op dit moment geen consensus over de classificatie, benaming en functie van de soorten onderzoeksdata. Het ontbreken van deze consensus belemmert het realiseren van de digitale duurzaamheid van deze objecten. Zonder een duidelijke identiteit van de soorten onderzoeksdata is het moeilijk de essentiële kenmerken¹² te beschrijven en daarmee het langetermijnbeheer te organiseren.

Er bestaat een aantal voorstellen om tot een classificatie voor onderzoeksdata te komen, waaruit een eerste aanzet tot indeling kan worden gedestilleerd¹³. Deze is hieronder weergegeven:

- gegevens die het resultaat zijn van wetenschappelijke experimenten die in principe herhaald kunnen worden, al dan niet tegen hoge kosten;
- modellen en simulaties, waarbij het model en de metadata van hogere waarde zijn dan de gegevens die uit de modellen voortkomen;
- observatiegegevens die betrekking hebben op specifieke fenomenen op een bepaald tijdstip of plaats. De data is uniek en kan niet opnieuw gegenereerd worden;
- afgeleide data: van "ruwe" data afgeleide gegevens die ontstaat door verschillende bronnen te verwerken of te combineren;
- referentie data, bijvoorbeeld teksten uit de literatuur, genendatabank, chemische structuren.

Als aparte categorie kan ook nog het web als onderzoekscorpus genoemd worden. Het web staat in toenemende mate in de belangstelling van

¹⁰ Informatie afkomstig uit "Financial Statements Accounting period 2007" te vinden op <<http://www.epo.org/about-us/publications/general-information/financial-report.html>> [bezocht 18 juni 2009]

¹¹ Zie: "Facts and figures 2009" te vinden op <<http://www.epo.org/about-us/publications/general-information/facts-figures.html>> [bezocht 18 juni 2009]

¹² In veel publicaties over oplossingen voor digitale duurzaamheid wordt het begrip "significant properties" gebruikt, de kenmerken van digitale objecten die essentieel voor de duurzame toegang.

¹³ Zie bijvoorbeeld de paragraaf "The diversity of data" in: Liz Lyon, *Dealing with data: Roles, Rights, Responsibilities and Relationship*. P15 <http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf> [bezocht 9 mei 2009].

wetenschappers. De KB archiveert een selectie van websites op het gebied van Nederlandse geschiedenis, cultuur en samenleving, met als doel te voorkomen dat ze niet meer toegankelijk zullen zijn¹⁴. Ook is het mogelijk het door de tijd ontstaan, veranderen en verdwijnen van websites te betrekken bij onderzoek. Zo kan het web bijvoorbeeld een onderzoeksbron worden voor taaltechnologen. Ook nieuwsgroepen op internet kunnen een belangrijke bron voor onderzoek zijn. Het IISG heeft bijvoorbeeld een verzameling nieuwsgroepen uit de periode 1986-2002, betreffende de oorlog in voormalig Joegoslavië, gearchiveerd onder de naam "Occasio"¹⁵.

Onderzoek heeft uitgewezen dat het delen van onderzoeksdata kan leiden tot een toename van het aantal citaties. Hiermee wordt duidelijk dat van het toegankelijk maken van onderzoeksdata niet alleen de gebruikers profiteren, maar ook de maker van de onderzoeksdata. Als dit inzicht in brede kring steun krijgt in de wetenschappelijke wereld zal hiermee het belang van goed ingerichte databewaarplaatsen bredere erkenning krijgen¹⁶.

De wetenschap creëert en gebruikt een complexe en heterogene data-infrastructuur. Hieronder vallen niet alleen databases, maar ook de organisaties die betrokken zijn bij het beheer en de beschikbaarstelling. De rol, functie en waarde van de digitale objecten verschilt per wetenschappelijke discipline.

1.2.4 Verrijkte publicaties

De informatietechnologie heeft de wetenschappelijke publicatie verlost van de beperkingen van het papieren tijdperk. Alle wetenschappelijke publicaties verschijnen nu als elektronisch artikel, maar als ze geopend of geprint worden lijken ze nog steeds op een gedrukt artikel. In toenemende mate worden publicaties verbonden met gerelateerde digitale dataobjecten, zoals databases en multimediatekstbestanden.

Inmiddels bestaat het algemene inzicht dat in toenemende mate publicaties een relatie hebben met andere soorten digitale objecten. Men spreekt dan van een "verrijkte publicatie" of "enhanced publication". Een verrijkte publicatie kan omschreven worden als een publicatie die verrijkt is met drie soorten informatie: (1) onderzoeksdata, waardoor toegang kan worden gekregen tot de onderzoeksgegevens waarop de uitkomsten zijn gebaseerd, (2) extra materiaal om het onderzoek te illustreren of nader toe te lichten, en (3) "post-publicatie" gegevens, zoals commentaren en waarderingen¹⁷.

SURFfoundation werkt samen met de universiteiten, DANS, de KB, 3TU.Datacentrum en andere organisaties aan de ontwikkeling van een infrastructuur voor verrijkte publicaties¹⁸. De langetermijntoegang en

¹⁴ Zie: <http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index.html> [bezoekt 20 mei 2009]

¹⁵ Zie: <<http://www.iisg.nl/occasio>> [bezoekt 11 juni 2009]. Overigens toont het Occasio project de noodzaak van lange termijn beheer duidelijk aan. Gedurende de eerste weken van juni 2009 was het niet mogelijk toegang te krijgen tot het online nieuwsarchief van Occasio.

¹⁶ Zie het artikel: Heather A. Piwowar e.a., *Sharing detailed research data is associated with increased citation rate*. In: PLOS ONE, March 2007, issue 3, p1-5.
<<http://www.plosone.org/article/fetchArticle.action?articleURI=info:doi/10.1371/journal.pone.0000308>>

¹⁷ Bron: "Report on enhanced publications state-of-the-art" Deliverable D4.1 van het DRIVER project <<http://www.driver-repository.eu>> [bezoekt 4 april 2009]

¹⁸ Zie: <<http://www.surfoundation.nl/nl/themas/openonderzoek/verrijktepublicaties>> [bezoekt 23 mei 2009]

bruikbaarheid van verrijkte publicaties maakt vanzelfsprekend deel uit van deze infrastructuur.

In dit rapport wordt het begrip “wetenschappelijk digitaal dataobject” gebruikt als algemene term voor alle digitale resultaten die van belang zijn voor wetenschappelijk onderzoek.

1.3 Digitale duurzaamheid in de wetenschap

Beknopt komen de belangrijkste argumenten met betrekking tot de duurzame bewaring van digitale onderzoeksobjecten aan de orde. Vervolgens is er in deze paragraaf aandacht voor het ontstaan van wetenschappelijke data-archieven en worden de standaarden en begrippen beschreven die centraal staan in het huidige denken over digitale duurzaamheid van onderzoeksdata. Gebaseerd op de huidige stand van zaken worden ten slotte een aantal scenario's gegeven voor de manier waarop de toekomstige toegang tot onderzoeksdata kan worden georganiseerd.

1.3.1 Het belang van digitale duurzaamheid

Er zijn drie argumenten te noemen die het belang van langetermijntoegang tot digitale wetenschappelijke objecten onderbouwen. Op de eerste plaats is het vanzelfsprekend dat onderzoeksdata die slechts eenmalig te creëren zijn, langetermijnaandacht vereisen om het risico van verlies van de data te minimaliseren. Op de tweede plaats is hergebruik van de onderzoeksdata in de toekomst een belangrijke reden om duurzame toegang te faciliteren. Ten derde kan de controleerbaarheid van wetenschappelijk onderzoek door toegang te krijgen tot de onderliggende onderzoeksdata genoemd worden.

In algemene zin kan gesteld worden dat de aanwas van onderzoeksdata te groot is om in zijn geheel voor duurzame bewaring in aanmerking te laten komen. Daarom vormt de selectie van het te bewaren materiaal een belangrijk aspect van het verwerkingsproces. Selectie is ook nodig om de daadwerkelijk belangrijke objecten te bewaren en bijvoorbeeld testbestanden en conceptversies te verwijderen. Om te kunnen bepalen welke objecten bewaard dienen te worden en in de toekomst te kunnen gebruiken is adequate documentatie of metadata noodzakelijk.

1.3.2 Archivering van digitale onderzoeksdata

Aan de hand van drie voorbeelden afkomstig uit verschillende wetenschappelijke disciplines wordt het ontstaan beschreven van organisaties die wetenschappelijke data archiveren.

In 1964 wordt de Steinmetz Stichting opgericht, genoemd naar Prof. Mr. Dr. S. R. Steinmetz, een van de grondleggers van empirisch georiënteerde sociologie in Nederland. Doelstelling van de stichting was te voorkomen dat waardevolle sociaal-wetenschappelijke databestanden verloren zouden gaan. Het Steinmetzarchief werd daarmee het eerste wetenschappelijke data-archief in Nederland¹⁹. De Steinmetz Stichting werd in de jaren zeventig een afdeling van het Sociaal- Wetenschappelijk Documentatiecentrum (SWIDOC) en kwam zo onder de vleugels van de KNAW. De datasets, verworven door het Steinmetz archief, zijn vandaag de dag toegankelijk via het EASY archiefsysteem van DANS²⁰. Doordat de datasets gedocumenteerd zijn, ze in de loop der tijd op nieuwe informatiedragers zijn overgezet, het bestandsformaat leesbaar bleef en er een organisatie is die toegang verschaft tot de bestanden, wordt bijvoorbeeld

¹⁹ Bron: <<http://www.moaweb.nl/bibliotheek/jaarboeken/1976/jaarboek-1976-16.pdf/>> [bezocht 8 mei 2009]

²⁰ Zie: <<http://easy.dans.knaw.nl>> [bezocht 6 juni 2009]

de “Jongerenenquête 1961” nog steeds gebruikt door wetenschappers²¹. Dit voorbeeld geeft aan dat institutionele inbedding van wetenschappelijke onderzoeksobjecten een belangrijke voorwaarde is voor de instandhouding er van.

Een ander vroeg voorbeeld van een relevant initiatief op het gebied van de data-infrastructuur voor de Nederlandse wetenschap zijn de activiteiten van het Instituut voor Nederlandse Lexicologie (INL)²². Het INL is opgericht in 1967 om met de modernste technieken de Nederlandse woordenschat in kaart te brengen. In 2001 publiceerde het INL “Blauwdruk voor onderhoud, beheer en distributie van door de overheid gefinancierde digitale materialen”²³ met daarin aandacht voor de duurzaamheid van materialen voor taal- en spraaktechnologie. Deze publicatie heeft mede geleid tot de oprichting van de “Centrale voor Taal- en Spraaktechnologie” in 2004²⁴. Deze TST-Centrale wordt gefinancierd door de Nederlandse Taalunie en is als project ondergebracht bij het INL. Dit projectmatig karakter van de TST-Centrale geeft per definitie duurzaamheidsproblemen in de toekomst, omdat het risico bestaat dat er na afloop van het project geen middelen meer zijn om de organisatie voort te zetten.

Dan een voorbeeld uit de exacte wetenschappen, bij uitstek internationaal georiënteerde. In de periode van 1989 tot 2000 liep bij CERN²⁵, de Europese organisatie op het gebied van kernonderzoek, het grote Aleph experiment, de voorloper van de Large Hydron Collider (LHC). De resultaten van het Aleph experiment zijn nog steeds beschikbaar via: <<http://aleph.web.cern.ch/aleph/>>²⁶. Over deze site met wetenschappelijke data is geen expliciete informatie beschikbaar met betrekking tot de betrouwbaarheid en duurzaamheid. De meest betrouwbare informatiebron over het Aleph-experiment lijkt de publicatie *The Aleph experience* te zijn, die in digitale vorm beschikbaar is via de eerder genoemde website²⁷. De elektronische publicatie bevat beschrijvingen van de opzet van de experimenten en de data-infrastructuur die gebruikt is. Tijdens het Aleph-project is het protocol geïntroduceerd waarop het World Wide Web is gebaseerd en is ook een van de eerste web servers ingericht. De natuurwetenschappers behoren tot de grootverbruikers van dataopslag met een hoge verwerkingsnelheid. In dit opzicht zijn nadere gegevens over het Aleph experiment bij CERN illustratief. In de periode 1989 tot 2000 werd regelmatig de hardware en het besturingssysteem van de onderzoeksinfrastructuur aangepast, maar er werd wel steeds gebruik gemaakt van de ANSI standaard Fortran77 (“with no machine specific extensions”), waardoor de software gemakkelijk gemigreerd kon worden en tot op de dag van vandaag bruikbaar blijft. De toename van de processorsnelheid zorgde ervoor dat een complete “reprocessing” van alle data uit een bepaald jaar in twee tot vier weken uitgevoerd kon worden op een nieuwe generatie apparatuur. In 1989 werd

²¹ De dataset “Jongeren enquête 1961” heeft de *persistent identifier* <urn:nbn:nl:ui:13-mne-2pt> waardoor deze een eenduidige en daarmee duurzame identiteit krijgt.

²² Zie: <<http://www.inl.nl>> [bezocht 6 juni 2009]

²³ Zie: <http://www.inl.nl/images/stories/taalbank/publicaties/blauwdruk_volledig.pdf> [bezocht 6 juni 2009]

²⁴ Zie: <<http://www.tst.inl.nl/>> [bezocht 6 juni 2009]

²⁵ Zie: <<http://www.cern.ch>> [bezocht 6 juni 2009]

²⁶ [bezocht 13 juni 2009]

²⁷ *The Aleph Experience* (second edition (2006)) zie: <<http://aleph.web.cern.ch/aleph/alpub/draft/AlephHistory.pdf>> [bezocht 13 juni 2009]

begonnen met een opslagcapaciteit van 20 gigabyte²⁸. Ruim 10 jaar later, in het jaar 2000, wordt er ruim 668 gigabyte aan data gegenereerd en is de totale vereiste opslagcapaciteit bijna 2.700 gigabyte²⁹. De data worden opgeslagen op tapes en deze tapes kunnen nog steeds opgevraagd worden. Voor zover bekend zijn er geen plannen om de gegevens op de tapes te migreren naar nieuwe dragers om te voorkomen dat de gegevens ontoegankelijk worden. Ook is niet bekend of er procedures bestaan om de leesbaarheid van de tapes te controleren. Het is overigens de vraag of de wetenschappers het een probleem vinden als de Aleph data verloren zou gaan. Het experiment is afgesloten, de uitkomsten zijn gepubliceerd en er ontstaat een nieuwe generatie instrumenten die kwalitatief betere onderzoeksdata oplevert.

1.3.3 De periode 1990 - 2000

Bij de formele overdracht van het archief van de Deltawerken aan het Rijksarchief in Zeeland in 1990 bleek dat de 55 strekkende meter machineleesbare gegevensdragers met primaire gegevens van metingen, waterdiepte en golf- en getijdenwaarnemingen voor een groot deel als verloren moesten worden beschouwd, omdat de documentatie ontbrak, de informatiedragers beschadigd waren of apparatuur ontbrak om de dragers te lezen. Dit verlies van databestanden was mede de aanleiding voor een onderzoek door de Algemene Rekenkamer. In 1991 publiceerde de Algemene Rekenkamer het rapport "Machineleesbare gegevensbestanden: archivering en beheer bij het Rijk". In dit rapport schenkt de overheid aandacht aan het probleem dat zonder beleidsmaatregelen computerbestanden verloren gaan. In het rapport wordt gesteld dat ook wetenschappelijk onderzoek gebaat is bij een effectief beleid op het gebied van het beheer van computerbestanden. Veel van de aanbevelingen uit 1991 zijn nog steeds actueel. Zo wordt gesteld dat secundair gebruik van bestaande bestanden meer aandacht dient te krijgen en dat de eigenaar en beheerder van de bestanden schriftelijke afspraken dienen te maken over het beheer en de archivering van de bestanden.

Tussen 1990 en 2000 werden er diverse projecten uitgevoerd waarbij origineel bronnenmateriaal gedigitaliseerd werd en digitale collecties en databases werden opgezet ten behoeve van wetenschappelijk onderzoek. Het ontstaan van het World Wide Web midden jaren negentig was een grote stimulans voor het aanbieden en gebruiken van digitale informatiebronnen. Hoewel er werd nagedacht over de organisatie van het beheer van en de langetermijntoegang tot deze digitale bronnen, zijn er in deze periode weinig concrete implementaties gerealiseerd. Overigens zijn momenteel nog steeds veel digitale collecties toegankelijk die in deze periode ontstaan zijn, zoals de "American Memory" collectie van de Library of Congress, waarmee men in 1990 begonnen is³⁰ en het hierop gebaseerde "Geheugen van Nederland"³¹ initiatief waarmee in 2000 een begin werd gemaakt .

Hoewel er in de jaren negentig van de vorige eeuw meer aandacht komt voor de duurzaamheid van digitale onderzoeksgegevens, is deze nog steeds incidenteel en versnipperd. Vanaf 1990 is er in Nederland een projectorganisatie die computerbestanden op het gebied van de Geschiedwetenschap archiveert en toegankelijk maakt: het Nederlands Historisch Data Archief (NHDA). Net als het

²⁸ Marco Gattaneo, "More on Aleph computing" in: *the Aleph Experience* pp 121-123), zie noot 27

²⁹ Zie: <http://aleph.web.cern.ch/aleph/aleph/Cartridges/Cartridges_real_data.html> [bezoekt 10 juni 2009]

³⁰ Zie: <<http://memory.loc.gov/ammem/index.html>> [bezoekt 10 juni 2009]

³¹ Zie: <<http://www.geheugenvannederland.nl>> [bezoekt 10 juni 2009]

Steinmetzarchief is het NHDA ontstaan als een stichting. Daarna werd het NHDA een KNAW-instituut om vervolgens samen met het eerder genoemde SWIDOC op te gaan in het Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI). Vanaf 2005 maakt de datacollectie van het NHDA deel uit van DANS.

In het midden van de jaren negentig wordt in Nederland de aanzet gegeven tot een aantal initiatieven die vanaf het jaar 2000 meer aandacht voor beheer en (her)gebruik van digitale onderzoeksdata opleveren. Dit zijn de activiteiten van het Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI), het "testbed digitale bewaring" van ICTU, de ICT uitvoeringsorganisatie van de overheid en de projecten op het gebied van de digitale duurzaamheid van elektronische publicaties (bijvoorbeeld het Europese NEDLIB project, gecoördineerd door de Koninklijke Bibliotheek³²), waaruit later het e-Depot voor publicaties zal voortkomen dat beheerd wordt door de KB.

Een laatste belangrijke ontwikkeling die in de periode 1990 – 2000 heeft plaatsgevonden is het ontstaan van een kwalitatief hoogwaardige netwerkinfrastructuur. Nederland hoort wereldwijd tot de koplopers waar het gaat om netwerkbandbreedte en dataverwerkingscapaciteit. SURFnet initieert en beheert deze nationale infrastructuur en verzorgt de netwerkverbindingen met het buitenland³³. Deze recentelijk tot stand gekomen netwerkinfrastructuur wordt momenteel als vanzelfsprekend ervaren.

1.3.4 Drie belangrijke pijlers

Met betrekking tot de gedachtevorming op het gebied van de digitale duurzaamheid van onderzoeksdata spelen drie concepten, ontwikkeld in de eerste jaren na 2000, een belangrijke rol. Dat is op de eerste plaats het OAIS referentiemodel, vervolgens het nauw daar aanverbonden concept *Trusted Digital Repository* en ten slotte het begrip *data curation*.

Het *Reference Model for an Open Archival Information System (OAIS)*³⁴, sinds 2003 ISO standaard ISO 14721, beoogt een raamwerk te bieden voor termen en concepten die relevant zijn voor de langetermijnarchivering van digitale data. De standaard heeft haar wortels in de wetenschap en is opgesteld door de organisatie die verantwoordelijk is voor het beheer van data afkomstig van de Amerikaanse ruimtevaartorganisatie NASA. Een OAIS wordt gedefinieerd als "*an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a designated community*". Belangrijk in het referentiemodel zijn de zes functionele entiteiten "ingest", "archival storage", "data management", "administration", "preservation planning" en "access". Er is nauwelijks een project, initiatief of organisatie te noemen die niet op een of andere wijze refereert aan deze standaard. Voor het wetenschapsdomein kan bijvoorbeeld het EU-project CASPAR³⁵ genoemd worden. Dit project ontwikkelt oplossingen voor de langetermijnbruikbaarheid van digitale onderzoeksdata op basis van de OAIS standaard. Het is belangrijk om te realiseren er slechts een begin is gemaakt met concrete implementatierichtlijnen voor OAIS-standaard en dat er geen consensus is over welke diensten en organisaties de standaard goed hebben toegepast.

³² Zie: <<http://nedlib.kb.nl/>> [bezoekt 13 juni 2009]

³³ Zie: <<http://www.surfnet.nl/nl/netwerk/Pages/Default.aspx>> [bezoekt 18 juni 2009]

³⁴ Zie: <public.ccsds.org/publications/archive/650x0b1.pdf> [bezoekt 13 juni 2009]

³⁵ Zie: <www.casparpreserves.eu> [bezoekt 13 juni 2009]

In 2002 verscheen het rapport *Trusted Digital Repositories. Attributes and Responsibilities*³⁶ dat de kenmerken en verantwoordelijkheden beschrijft van een zogenaamde TDR (trusted digital repository) die als doel heeft om betrouwbare langetermijntoegang te geven tot beheerde digitale bronnen voor haar *designated community* (letterlijk: *A TDR is a "mission to provide reliable long-term access to managed digital resources to its designated community, now and into the future"*). Het rapport vormt de basis van een aantal belangrijke vervolgstappen in relatie tot de ontwikkeling van digitale archieven voor wetenschappelijke data, zoals het ontwikkelen van certificeringsprocedures, het ontwikkelen van "tools" om digitale archivering te faciliteren, het opstellen van kwaliteitsraamwerken en het ontwikkelen van strategieën om digitale duurzaamheid te realiseren. De term "repository" is inmiddels goed ingeburgerd waar het gaat om de opslag van en toegang tot digitale data. De universiteit van Leiden gebruikt de Nederlandse term "repositorium" voor haar systeem dat via internet toegang geeft tot Open Access elektronische publicaties³⁷.

Een "repository" is niet automatisch een "trusted" repository. Er bestaat consensus dat een TDR naast het toegang verschaffen tot digitale objecten ook additionele functies op het gebied van digitale archivering kent. Deze functies zijn opgenomen in het OAIS-model en aan de implementatie ervan wordt in internationaal verband gewerkt. Zo ontstaan er toepassingen van metadatastandaarden, zoals de PREMIS metadatastandaard³⁸, en services om data te migreren en software te emuleren. Het Internationaal Instituut voor Sociale Geschiedenis heeft het voornemen om een TDR te bouwen opgenomen in haar *"Strategy memorandum 2007-2010"*.

Met het groeien van het aantal digitale objecten van belang voor de wetenschap alsmede de toename van de complexiteit er van, ontstond het inzicht dat voortdurende aandacht en bewerkingen noodzakelijk zijn om gebruik en onderhoud van deze objecten gedurende de totale levenscyclus mogelijk te maken. De term "data curation" (of "datacuratie" in het Nederlands) werd voor het eerst gebruikt in 2001 en heeft betrekking op de activiteiten die naast digitale conservering uitgevoerd worden om onderzoeksdata optimaal te kunnen hergebruiken en te kunnen verrijken³⁹. De tools die ontwikkeld en gebruikt worden om een TDR te realiseren, kunnen bijvoorbeeld gezien worden als datacuratie activiteiten. Een indicator dat datacuratie een ingeburgerd begrip is geworden is het feit dat sinds 2006 het *peer-review* elektronische tijdschrift *International journal of digital curation* bestaat⁴⁰.

1.3.5 Scenario's voor duurzame toegang tot onderzoeksdata

Vanaf het jaar 2000 is het gebruik van informatietechnologie in het wetenschapsbedrijf enorm gegroeid. Er is er een aantal standaarden, strategieën

³⁶ Zie: <www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf> [bezocht 13 juni 2009]

³⁷ Zie: <<http://www.bibliotheek.leidenuniv.nl/catalogi-databases/catalogi/leids-repositorium.html>> [bezocht 12 mei 2009]

³⁸ Zie: <<http://www.loc.gov/standards/premis/>> [bezocht 15 mei 2009]

³⁹ Digital Curation: digital archives, libraries and e-science seminar" sponsored by the Digital Preservation Coalition and the British National Space Centre held in London, October 19th 2001. See: <<http://www.dpconline.org/graphics/events/digitalarchives.html>> [bezocht 11 juni 2009]. Voor een recente beschrijving van de "data-curatie" activiteiten van bibliotheken, zie: Inge Angevaere *Taking care of digital collections and data: 'curation' and organisational choices for research libraries*. In: *Liber Quarterly* 19 (1) april 2009 P1-12 <<http://liber.library.uu.nl/publish/articles/000278/article.pdf>> [bezocht 29 juni 2009]

⁴⁰ Zie: <<http://www.ijdc.net>> [bezocht 10 juni 2009]

en organisaties ontstaan die de toekomstige toegang tot wetenschappelijke onderzoeksobjecten faciliteren. Het lijkt erop dat momenteel de eerste stappen worden gezet om te komen tot een brede consensus binnen het wetenschapsbedrijf waar het gaat om de noodzakelijke instrumenten om de duurzame opslag en toegang tot wetenschappelijke onderzoeksobjecten te realiseren. Het gaat dan om zaken als selectie van de objecten die voor langetermijntoegang in aanmerking komen, opslag van data in duurzame bestandsformaten, adequate documentatie van de objecten en het inrichten van een beheersorganisatie.

Op het vlak van de organisatorische inbedding van taken en functies op het gebied van digitale duurzaamheid zijn momenteel verschillende scenario's denkbaar, waarbij verschillende niveaus van ondersteuning, samenwerking en dienstverlening te onderscheiden zijn. Hieronder worden er acht genoemd:

1. Het meest basale alternatief stelt dat onderzoekers zelf voor opslag en toegang tot hun data zorgen.
2. Internationale vakspecifieke repositories.
3. Universitaire repositories, gelieerd aan de universiteitsbibliotheek.
4. Nationale repositories voor bepaalde wetenschappelijke disciplines.
5. "Research infrastructure network" voor alle wetenschapsgebieden, waarbij data, tools en expertise worden samengebracht.
6. Een groot nationaal domeinonafhankelijk digitaal archief, waarin data-archieven, de nationale bibliotheek en het nationaal archief participeren.
7. "Cloud computing" door gebruik te maken van generieke dataopslag- en analysefaciliteiten, bijvoorbeeld aangeboden door Google.
8. Het BIG GRID model, een "technology driven" oplossing die veel lijkt op "cloud computing" maar waarbij de infrastructuur onder beheer van de sector blijft.

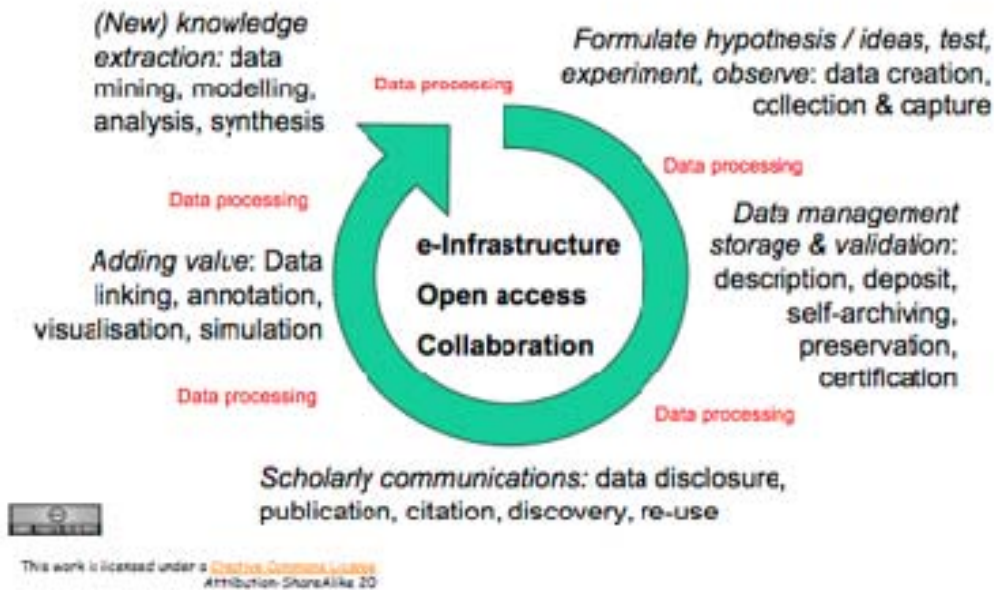
1.3.6 De onderzoekscyclus

Wetenschappelijk onderzoek bestaat uit een aantal stadia. Lange tijd kon wetenschappelijk onderzoek als een lineair proces beschouwd worden, beginnend bij het formuleren van een onderzoeksvraag en eindigend bij het rapporteren van de uitkomsten van het onderzoek in een publicatie. Met de opkomst van de informatietechnologie is dit proces veranderd.

Er bestaan verschillende modellen die de rol van digitale objecten in het wetenschappelijk onderzoeksproces visualiseren, waarbij er altijd sprake is van een cyclisch karakter. Een goed voorbeeld van een model dat de "levenscyclus"-visie op wetenschappelijk onderzoek visualiseert staat hieronder afgebeeld⁴¹.

⁴¹ Het model is afkomstig uit Liz Lyon, Dealing with data: Roles, Rights, Responsibilities and Relationship" p48 <http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf> [bezoekt 9 mei 2009]

(e)-Research Life Cycle view of Data Curation?



Het model heeft als titel "(e)-Research Life Cycle view of Data Curation. Het "(e)-Research" aspect van het model vertaalt zich in de vele vormen van dataprocessing die onderdeel zijn van het model. Centraal in het model staat de infrastructuur die de levenscyclus mogelijk maakt. De scenario's genoemd in de vorige paragraaf kunnen dit onderdeel van het model concretiseren. Open Access, dat verderop nog nader wordt toegelicht, en samenwerking zijn ook deel van het vliegwiel dat creatie, beheer, disseminatie, verrijking en analyse van onderzoekdata mogelijk maakt. Lyon stelt dat er weinig kennis is over de manier waarop individuele onderzoekers aankijken tegen datacuratie. Ook is er over het algemeen weinig kennis bij onderzoekers op het gebied van beheer en (her)gebruik van onderzoeksdata. Ze stelt dat met name longitudinale cohort studies op (bio-)medisch terrein interessant zijn om de werking van het model in de praktijk te toetsen.

De aanpak van het 3TU.Datacentrum is een goed voorbeeld van het in de praktijk brengen van het samenwerkingsprincipe. Op de website roept het 3TU.Datacentrum onderzoekers op om onderzoeksgegevens aan te leveren en onder te brengen in een samenwerkingsomgeving, een "collaboratory"⁴². Ook het Internationaal Instituut voor Sociale Geschiedenis werkt aan het inrichten van collaboratories.

Hoewel het model bestaat uit vijf gelijkwaardige fasen, worden deze op dit moment zelden met gelijk gewicht toegepast. Dienstverleners op het gebied van datacuratie richten zich met name op "data management, storage and validation" en op het "scholarly communication" aspect. Respectievelijk het DANS data-archief en het KB e-Depot voor wetenschappelijke publicaties zijn hier een goed voorbeeld van.

⁴² Zie: <<http://datacentrum.3tu.nl>> [bezoekt 18 juni 2009]

1.3.7 Verantwoordelijkheden

Verschillende partijen spelen een rol bij de creatie en het gebruik van bewaarplaatsen van digitale wetenschappelijke objecten. Onderstaand overzicht, afkomstig uit het rapport *Dealing with data*⁴³ kan gebruikt worden als referentiekader bij het bepalen van de verantwoordelijkheden die er zijn bij creatie, (her)gebruik, beheer en archivering van onderzoeksdata.

Rol	Rechten	Verantwoordelijkheden	Relaties
<i>Wetenschapper:</i> Creatie en gebruik van data	Primair gebruik van de data Wetenschappelijke erkenning Eigendomsrecht Ontvangen van training en advies	Beheer van de data gedurende het onderzoeksproject In acht nemen van relevante richtlijnen en regelgeving Data bewerken zodat anderen deze ook kunnen gebruiken	Werknemer van instelling Met experts Met data centrum Met financier
<i>Instelling:</i> Bewaren van en toegang verschaffen tot data	Verkrijgen van kopie van de data	Instellen interne richtlijnen voor databeheer Databeheer op korte termijn Toepassen van standaarden Geven van training en advies aan wetenschappers Promoten van "repository" dienst	Werkgever van wetenschapper Via experts met data centrum
<i>Data centrum:</i> Curatie van en toegang verschaffen tot data	Verkrijgen van kopie van de data Selectie van data van waarde voor lange termijn bewaring	Beheer van data voor de lange termijn Toepassen van standaarden Aanbieden van training mbt datadeponering Promoten van dienstverlening Beschermen van relevante rechten Aanbieden van <i>tools</i> voor hergebruik van de data	Wetenschapper is "klant" Met gebruikersgroepen Via experts met instelling Met financier van de dienstverlening
<i>Gebruiker:</i> (her)gebruiken van bestaande data	Hergebruiken van data (<i>non-exclusive</i>) Toegang tot metadata om bruikbaarheid te kunnen bepalen	Erkennen van licentievoorwaarden Erkennen van data producenten Effectief beheer van (afgeleide) data	Met het data centrum als leverancier van data Met de instelling als de leverancier van de data
<i>Financier:</i> Opstellen van, aansluiten bij, reageren op regelgeving	Implementeren van regelgeving Eisen dat uitvoerders zich aan regelgeving houden	Breed perspectief mbt belanghebbenden in acht nemen Participeren in coördinatie van de datastrategie Ontwikkelen van beleid met belanghebbenden Coördinatie van beleidsontwikkeling Monitoren en opleggen van regelgeving op het gebied van data Zorg dragen voor lange termijn databeheer na afloop van het project Stimuleren van datacuratie en financieren van expert dienstverlening Steunen van ontwikkeling van expertise van data curatoren	Als financier van de wetenschapper Met instelling Als financier van data centrum Met andere financiers Met andere belanghebbenden
<i>Uitgever:</i> integriteit van de wetenschappelijke output garanderen	Verwachten dat data beschikbaar is om de publicatie te ondersteunen Verzoeken om data op te slaan in lange termijn opslagplaats	Betrekken van stakeholders van ontwikkeling van publicatiestandaarden Verwijzen naar data Ontwikkelen en monitoren van standaarden	Met de wetenschapper als producent, auteur en lezer Met data centra en instellingen als aanbieders van data

⁴³ Zie noot 37, p22-23 [bezocht 9 mei 2009]

Er bestaan verschillende initiatieven om de verantwoordelijkheden met betrekking tot de duurzaamheid van wetenschappelijke data nader in te vullen. Zo voert het 3TU.Datacentrum het project "Waardevolle Data en Diensten" uit om de functionele eisen te formuleren voor de inrichting van het 3TU.Datacentrum. Daarnaast organiseert SURFfoundation bijeenkomsten met experts om de ontwikkelingen en issues in kaart te brengen op het gebied van onderzoeksdata. Een belangrijke actielijn hierbij is het stimuleren en belonen van onderzoekers om onderzoeksdata te deponeren en te hergebruiken.

Een laatste voorbeeld betreft de symposia die DANS regelmatig organiseert waarbij de kennis over (her)gebruik, toegankelijkheid en houdbaarheid van onderzoeksdata centraal staat⁴⁴.

⁴⁴ Zie: <http://www.dans.knaw.nl/nl/dans_symposia/> [bezocht 28 juni 2009]

2 Kaders en codes in relatie tot duurzame toegang tot onderzoeksdata

In dit hoofdstuk wordt een aantal strategieën en gedragscodes genoemd die een rol spelen bij wetenschappelijk onderzoek in Nederland en een relatie hebben met het duurzaam beheer van en toegang tot digitale objecten van belang bij wetenschappelijk onderzoek.

2.1 Kwaliteitszorgsysteem voor wetenschappelijk onderzoek

In 2003 is een kwaliteitszorgsysteem voor wetenschappelijk onderzoek opgesteld dat bestaat uit interne evaluaties en externe visitaties. KNAW, NWO en VSNU (als vertegenwoordiger van alle Nederlandse universiteiten) hebben een *Standard Evaluation Protocol 2003-2009 for Public Research* opgesteld op basis waarvan de evaluaties en visitaties worden uitgevoerd⁴⁵. De belangrijkste evaluatiecriteria voor publiek gefinancierd onderzoek zijn kwaliteit, productiviteit, relevantie en levensvatbaarheid. Hierbij speelt het aantal en de kwaliteit van *academic publications* (zonder onderscheid te maken tussen papieren en digitale versies) een belangrijke rol. De *academic reputation* wordt als volgt nader ingevuld:

The academic reputation of the institute may be indicated in several ways. Institutes and disciplines may refer to the practice of presenting a bibliometric analysis of the citations of the scientific results. Previous peer reviews, rewards and prizes may be cited (p30 Standard Evaluation Protocol)

Bovenstaand citaat uit het *Standard Evaluation Protocol* illustreert het belang van de impact (in de vorm van citaties) van wetenschappelijk onderzoek waar het gaat om het vaststellen van de kwaliteit van wetenschappelijk onderzoek. De bibliometrische analyse die genoemd wordt geeft aan dat de wetenschappelijke publicatie hierbij een belangrijke rol speelt. Met betrekking tot het kwaliteitscriterium *productivity*, onderdeel van het genoemde *standard evaluation protocol*, is er ook aandacht voor de rol van digitale objecten in het onderzoek. Dit wordt geïllustreerd door onderstaand citaat:

Furthermore, new tools for mapping and analysing productivity are emerging to take account of changes in publication behaviour. As more and more results of research become available through the Internet, these tools become increasingly appropriate and valuable. The research organisations will follow these developments closely and consider the introduction of such new tools into the evaluation process once they have proven their credibility and can provide significant added value to the evaluation process. (p10 Standard Evaluation Protocol)

Nieuwe 'tools' zullen dus in de toekomst bij het evaluatieproces betrokken worden als ze betrouwbaarder worden en toegevoegde waarde bieden bij het evaluatieproces. Het evaluatieprotocol is gepubliceerd in januari 2003 en sindsdien is de rol en waarde van digitale onderzoeksobjecten, zoals databanken, toegenomen, waardoor de langetermijntoegang tot deze objecten en daarmee de digitale duurzaamheid van wetenschappelijke objecten belangrijker wordt.

⁴⁵ Zie: <<http://www.qanu.nl/comasy/uploadedfiles/sep2003-2009.pdf>> [bezoekt 9 mei 2009]

2.2 Gedragscode wetenschapsbeoefening

Op verzoek van de VSNU is in oktober 2004 de "Gedragscode wetenschapsbeoefening" opgesteld met daarin de principes van goed wetenschappelijk onderzoek en onderwijs⁴⁶. De gedragscode bevat vijf kernbegrippen: "zorgvuldigheid", "betrouwbaarheid", "controleerbaarheid", "onpartijdigheid" en "onafhankelijkheid". Op het gebied van de digitale duurzaamheid van onderzoeksresultaten bevat de gedragscode bij de uitwerking van het principe van de "controleerbaarheid" een aantal relevante details voor deze verkenning. De gedragscode stelt dat onderzoek gerepliceerd moet kunnen worden om de juistheid ervan te kunnen testen. Nauwkeurige documentatie van de geraadpleegde bronnen is hierbij belangrijk. Ook wordt gesteld dat de kwaliteit van de dataverzameling, data-invoer, dataopslag en dataverwerking goed bewaakt dient te worden. De bewaartermijn van ruwe onderzoeksgegevens wordt op minimaal 5 jaar gesteld. Deze gegevens dienen op aanvraag ter beschikking gesteld te worden aan andere wetenschapsbeoefenaren. Deze ruwe onderzoeksgegevens dienen zodanig gearchiveerd te worden dat deze te allen tijde met een minimum aan tijd en handelen kunnen worden geraadpleegd.

De gedragscode wetenschapsbeoefening beschrijft welk gedrag gewenst is en bevat geen bepalingen over klachtenregelingen of sancties. Daarom is de gedragscode weliswaar een goede bijdrage aan het verbeteren van de langetermijntoegang en kwaliteit van digitale onderzoeksobjecten, maar zijn minder vrijblijvende vervolgstappen nodig.

2.3 ICT-strategie van de koepelinstanties

In deze paragraaf wordt aandacht geschonken aan de rol van de informatietechnologie binnen de strategie van KNAW en NWO, de twee belangrijkste koepelinstanties op het gebied van wetenschappelijk onderzoek.

Het strategisch plan van de KNAW 2007-2010, getiteld "Duurzame wetenschap" pleit voor een vrije toegang tot wetenschappelijke informatie⁴⁷. Onderstaand fragment beschrijft de strategische overwegingen van de KNAW op dit terrein:

De moderne informatietechnologie biedt grote mogelijkheden voor snelle en effectieve toegang tot wetenschappelijke informatie en kan de reikwijdte van wetenschappelijk onderzoek sterk verbreden. Grote veranderingen treden al op met betrekking tot publiceren, het opslaan, verwerken en analyseren van data, het benutten van collecties, en niet in de laatste plaats in het onderzoek zelf. Bij haar advisering over zulke onderwerpen is open access voor de KNAW het uitgangspunt. Dat kleurt ook de concrete activiteiten van bijvoorbeeld de KNAW-instituten. Met het ministerie van OCW overlegt de KNAW over advies- en andere taken op het terrein van de wetenschappelijke informatieverzorging.

Met name het belang van Open Access toegang tot onderzoeksresultaten heeft invloed op de kwaliteit van de data-infrastructuur voor wetenschappelijk onderzoek. Zo zullen er betrouwbare bewaarplaatsen of repositories nodig zijn om deze Open Access toegang mogelijk te maken

De NWO-strategie voor de periode 2007-2010 heeft als titel "Wetenschap gewaardeerd". De strategie kent de actielijn "bundeling van krachten"

⁴⁶ De gedragscode is te downloaden via: <<http://www.vsnu.nl/Universiteiten/Publicaties.htm>> [bezocht 12 mei 2009]

⁴⁷ Het strategisch plan van de KNAW is te vinden op: <http://www.knaw.nl/cfdata/publicaties/detail.cfm?boeken__ordernr=20061041>). [bezocht 12 mei 2009]

waarbinnen een aantal activiteiten wordt genoemd waarbij digitale duurzaamheid een rol speelt: bij de ontwikkeling van “nationale research initiatieven” (programma’s van 30-50 miljoen euro op gebieden waar Nederland een toppositie inneemt), bij het realiseren van grootschalige onderzoeksfaciliteiten en het stimuleren van Europese onderzoekssamenwerking in de vorm van een *roadmap*⁴⁸. In juli 2007 is de “Commissie Nationale Roadmap Grootschalige Onderzoeksfaciliteiten” ingesteld. Deze heeft in oktober 2008 een rapport uitgebracht waarin vijftientig grootschalige onderzoeksfaciliteiten opgenomen zijn, verspreid over verschillende wetenschapsgebieden⁴⁹. Het rapport stelt dat ICT voor alle grootschalige onderzoeksfaciliteiten een noodzakelijke randvoorwaarde is en wijst in dit verband op het belang van het initiatief van de Europese *Alliance for Permanent Access* om binnen Europa te komen tot een organisatorische infrastructuur voor de permanente toegang tot digitale onderzoeksdata en wetenschappelijke publicaties⁵⁰.

Bij subsidies die NWO toekent waarbij een dataverzameling wordt aangelegd, zoals bijvoorbeeld binnen het programma investeringssubsidies middelgroot of bij subsidie-instrumenten voor dataverzamelingen, dient de uitvoerder een *datacontract* met DANS af te sluiten. Zodoende blijven onderzoeksdata ook na afloop van het project toegankelijk en bruikbaar. De onderzoeksdata dienen te voldoen aan het *Data Seal of Approval* (DSA), waarover meer informatie beschikbaar is elders in dit rapport.

2.4 Toetsingskaders

Er bestaat een aantal initiatieven om kaders te scheppen waarbinnen duurzame toegang tot onderzoeksdata gerealiseerd kan worden. Deze recentelijk tot stand gekomen kaders kunnen beschouwd worden als een theoretische onderbouwing bij de praktische werkzaamheden die verricht worden, of die idealiter verricht zouden moeten worden. Momenteel worden de eerste stappen gezet om tot komen tot een structurele organisatorische inbedding. Hieronder wordt een aantal toetsingskaders beschreven.

Het opstellen van een toetsingskader is de eerste stap naar een situatie waarbij gesteld kan worden dat de duurzame toegang tot digitale objecten goed geregeld is. Het toepassen van de criteria en bepalen of aan de criteria voldaan is, is een volgende stap. Hierbij is inzet van domeindeskundigen van groot belang. Zij zijn op de hoogte van de stand van de wetenschap en kunnen bepalen welke objecten bewaard dienen te worden. In zekere zin sluit dit proces aan op het *peer review* principe dat in de loop der tijd is ontstaan waar het gaat om de creatie en bewaring van wetenschappelijke artikelen.

2.4.1 Data Seal of Approval

Het Data Seal of Approval⁵¹ (DSA) heeft als doel om ervoor te zorgen dat onderzoeksdata op lange termijn vindbaar, toegankelijk, bruikbaar, betrouwbaar en refereerbaar blijven. DANS heeft het initiatief genomen voor het opstellen van dit datakeurmerk en dit inmiddels overgedragen aan een internationaal “editorial board” bestaande uit vertegenwoordigers van verschillende Europese data-archieven. Het bestuur zal het DSA periodiek beoordelen en indien nodig bijstellen. Daarnaast zal het bestuur het beoordelingstraject bewaken dat

⁴⁸ Zie: <http://www.nwo.nl/nwohome.nsf/pages/NWOP_5SME25> [bezocht 12 mei 2009]

⁴⁹ Zie: <<http://www.minocw.nl/documenten/Roadmapdefinitief.pdf>> [bezocht 12 mei 2009]

⁵⁰ Zie: <<http://www.alliancepermanentaccess.eu/>> [bezocht 12 mei 2009]

⁵¹ Zie: <<http://www.datasealofapproval.org/>> [bezocht 18 mei 2009]

ingesteld gaat worden om organisaties te laten voldoen aan de richtlijnen die onderdeel zijn van het DSA. Het DSA beoogt op basis van onderling vertrouwen de digitale duurzaamheid van onderzoeksdata te bewaken.

Hoewel de meeste van de 16 richtlijnen van het DSA betrekking hebben op de activiteiten van het data-archief, worden ook de dataproducent en de datagebruiker aangesproken. In de tabel hieronder worden de DSA richtlijnen beknopt weergegeven.

Data Seal of Approval Guidelines (< www.datasealofapproval.org >)	
01	The data producer deposits the research data in a data repository qualified according to the DSA guidelines
02	The data producer provides the research data in formats recommended by the data repository
03	The data producer provides the research data together with the metadata requested by the data repository
04	The data repository has an explicit mission in the area of digital archiving and promulgates it
05	The data repository uses due diligence to ensure compliance with legal regulations and contracts
06	The data repository applies documented processes and procedures for managing data storage
07	The data repository has a plan for long-term preservation of its digital assets
08	Archiving takes place according to explicit workflows across the data life cycle
09	The data repository assumes responsibility from the data producers for access and availability of the digital objects
10	The data repository enables the users to utilize the research data and refer to them
11	The data repository ensures the integrity of the digital objects and the metadata
12	The data repository ensures the authenticity of the digital objects and the metadata
13	The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS
14	The data consumer complies with access regulations set by the data repository
15	The data consumer conforms to and agrees with any codes of conduct that are generally accepted in higher education and scientific research for the exchange and proper use of knowledge and information
16	The data consumer respects the applicable licences of the data repository regarding the use of the research data

2.4.2 DRAMBORA

DRAMBORA staat voor *Digital Repository Audit Method Based on Risk Assessment* en bestaat uit een methode om de kwaliteit van een repository vast te stellen⁵². Zoals de titel al aangeeft speelt het beheersen van risico's de centrale rol binnen de methode. DRAMBORA heeft een uitgebreide toolkit opgesteld waarmee onderstaande doelen bereikt kunnen worden:

- Het definiëren van de doelgroep en functies van een repository

⁵² Zie: <<http://www.repositoryaudit.eu>> [bezoekt 20 mei 2009]

- Het identificeren van de activiteiten en inhoud van de repository
- Het identificeren van de risico's en kwetsbaarheden in relatie tot de doelstellingen, activiteiten en inhoud van de repository
- Het benoemen en calculeren van deze risico's
- Het definiëren van maatregelen om de risico's te beheren
- Het rapporteren van deze zelfevaluatie.

Terwijl de DSA-richtlijnen veel meer voorschrijvend zijn, helpt de DRAMBORA toolkit een organisatie die al enigszins bekend is met digitale archivering om alle aspecten ervan in detail te beoordelen en een pro-actief beheersplan op te stellen.

2.4.3 TRAC

TRAC staat voor *Trusted Repositories Audit and Certification* en heeft zijn oorsprong in de Verenigde Staten⁵³. TRAC levert tools voor auditing, beoordeling en potentiële certificering van digitale repositories van allerlei aard. TRAC levert een controlelijst op drie gebieden, namelijk op het gebied van de organisatorische infrastructuur, op het gebied van het beheer van digitale objecten en op het gebied van de techniek, technische infrastructuur en beveiliging. TRAC sluit aan bij een aantal ISO-standaarden, waaronder het OAIS reference model.

2.4.4 nestor

nestor is een Duits initiatief met als doel een netwerk op te zetten van expertise op het gebied van langetermijnarchivering van digitale bronnen⁵⁴. nestor heeft een catalogus opgesteld van criteria waaraan "Trusted Digital Repositories" dienen te voldoen.

Het Data Seal of Approval (DSA) richt zich voornamelijk op toegang en gebruik van onderzoeksdata, terwijl DRAMBORA en TRAC zich meer richten op de bewaring en archivering van onderzoeksdata, gezien vanuit de organisaties die zich ten doel stellen digitale objecten te archiveren.

2.5 Internationale beleidskaders

Naast nationale strategieën en richtlijnen voor beheer en toegang tot onderzoeksdata is er een aantal internationale ontwikkelingen dat veel invloed heeft op de Nederlandse situatie. Dit zijn de Open Access beweging, de OECD principes en de recentelijk tot stand gekomen ESFRI onderzoeksinfrastructuren.

2.5.1 Open Access

Onder invloed van de nieuwe mogelijkheden van het world wide web ontstond de Open Access beweging. De Open Access benadering houdt in dat onderzoeksresultaten (inclusief ruwe data, metadata, bronnenmateriaal en gerelateerde multimedia bestanden) zonder belemmeringen vrij toegankelijk worden voor iedereen. Voorwaarde hierbij is dat de rechthebbenden expliciet toestemming geven voor dit gebruik en dat de complete onderzoeksresultaten tenminste in één online repository worden opgenomen, waarbij betrouwbare langetermijntoegang geregeld moet zijn. De Open Access principes vormen dus een belangrijk aspect van de duurzame infrastructuur voor wetenschappelijk onderzoek.

De Open Access-beweging heeft veel invloed op de juridische en financiële aspecten van de disseminatie van wetenschappelijke kennis. Onder regie van

⁵³ Zie: <<http://www.crl.edu/PDF/trac.pdf>> [bezoekt 20 mei 2009].

⁵⁴ Zie: <<http://www.langzeitarchivierung.de>> [bezoekt 6 juni 2009]

SURF is 2009 uitgeroepen tot het jaar van Open Access⁵⁵. Gedurende het hele jaar wordt samengewerkt aan het formuleren van beleid, het ontwikkelen en verbeteren van de kennisinfrastructuur, het zorgdagen voor een duidelijk juridisch kader en het geven van voorlichting aan alle betrokkenen.

Ook "Nederland Open in Verbinding", dat zicht richt op het gebruik van open standaarden en open source software bij de (semi) publieke sector, is van belang voor de wetenschap⁵⁶. Hoewel het plan zich op de overheid in brede zin richt, stimuleert het plan ook het breder toegankelijk worden voor wetenschappelijk gebruik van belangrijke databestanden die aangelegd worden door de overheid.

2.5.2 OECD principles for access to research data

De snelle ontwikkelingen in de informatietechnologie hebben ervoor gezorgd dat er nieuwe manieren van wetenschappelijke communicatie en nieuwe soorten van wetenschappelijke data zijn ontstaan. Omdat doorgaans hoge investeringen nodig zijn om onderzoeksdata te creëren, ontstaat in toenemende mate het besef dat toekomstige toegang tot deze data van groot belang is om verder wetenschappelijk onderzoek en innovaties mogelijk te maken. De OECD (*Organisation for Economic Co-operation and Development*) werkt vanaf 2000 aan het opstellen van algemene principes ten behoeve van toegang tot onderzoeksresultaten die met publieke middelen gefinancierd zijn. Dit leidde in 2007 tot de publicatie van de *OECD Principles and Guidelines for Access to Research Data From Public Funding*⁵⁷. Hierin worden dertien principes genoemd in relatie tot het ontwikkelen van toegang tot met behulp van publieke middelen gefinancierde onderzoeksdata. Eén van deze principes, relevant voor deze verkenning, is *sustainability*, nader beschreven als "*taking measures to guarantee long term access to data*". Ondanks het feit dat de er geen concrete implementatierichtlijnen bij de OECD-aanbevelingen zijn opgenomen, hebben ze veel invloed, omdat beleidsorganisaties en financiers op het gebied van de wetenschappelijke data-infrastructuur er vaak aan refereren en met behulp van de aanbevelingen verantwoordelijkheden benoemen.

2.5.3 ESFRI-roadmap

In het jaar 2000 ontstond het idee om op Europees niveau de ontwikkeling van onderzoeksinfrastructuren te coördineren. Het "European Strategy Forum on Research Infrastructures" (ESFRI) vormt het strategisch instrument om de positie van de wetenschap op Europees niveau te verbeteren⁵⁸. Een expert groep werd ingesteld die richtlijnen opstelde voor het inrichten van een Europese onderzoeksinfrastructuur. In oktober 2006 verscheen de eerste *European Roadmap for Research Infrastructures* met daarin opgenomen 35 internationale onderzoeksfaciliteiten. Bij een groot aantal daarvan zijn Nederlandse wetenschappers betrokken. Een "*research infrastructure*" of grootschalige onderzoeksfaciliteit bestaat uit essentiële gereedschappen voor wetenschappelijke vooruitgang en voor het verrichten van toponderzoek, heeft een groot maatschappelijk en economisch belang, leidt tot een concentratie van menselijk kapitaal en fungeert als knooppunt. Hieronder vallen zowel apparaten (zoals microscopen), als databanken en onderzoekscollecties⁵⁹.

⁵⁵ Zie: <<http://www.surffoundation.nl/nl/themas/openonderzoek/OpenAccess/Pages/default.aspx>> [bezoekt 16 juni 2009]

⁵⁶ Zie: <<http://www.ososs.nl/noiv>> [bezoekt 10 juni 2009]

⁵⁷ Zie: <<http://www.oecd.org/dataoecd/9/61/38500813.pdf>> [bezoekt 2 mei 2009]

⁵⁸ Zie: <<http://cordis.europa.eu/esfri/home.html>> [bezoekt 2 mei 2009]

⁵⁹ Zie: <<http://www.minocw.nl/documenten/Roadmapdefinitief.pdf>> [bezoekt 2 mei 2009]

In 2008 is voor verschillende onderzoeksfaciliteiten die op de roadmap staan Europese financiering beschikbaar gekomen voor de voorbereidende fase om de infrastructuren op termijn te kunnen realiseren. Daarnaast is er door een commissie vastgesteld welke onderzoeksfaciliteiten van belang zijn voor het Nederlandse wetenschapsysteem en financiële steun dienen te ontvangen. De onderzoeksfaciliteiten worden opgezet binnen alle wetenschapsdomeinen. Bij alle onderzoeksfaciliteiten speelt uiteraard informatietechnologie in meer of mindere mate een rol en daarmee dus ook de toekomstige toegang en duurzaamheid van de onderzoeksresultaten. Binnen het domein van de geestes- en maatschappijwetenschappen heeft DARIAH ("*Digital Research Infrastructure for the Arts and Humanities*"⁶⁰), het expliciete doel om langetermijntoegang tot onderzoeksdata op Europees niveau te realiseren. Maar ook voor CESSDA⁶¹ dat beoogt de toegang tot Europese onderzoeksdata te faciliteren en CLARIN dat "*language resources*" wil ontsluiten, is de duurzaamheid van onderzoeksobjecten een belangrijk onderwerp⁶².

Binnen het domein van de natuur- en technische wetenschappen ligt de nadruk in eerste instantie op het bouwen van apparaten als lasers en telescopen. Maar ook hier is natuurlijk de duurzaamheid van de data-infrastructuur een aandachtspunt, met name omdat deze apparaten zeer grote hoeveelheden ruwe onderzoeksdata genereren waarvoor dataopslag en -verwerkingscapaciteit vereist is. Het KM3NET⁶³ project bouwt een neutrino-telescoop op de bodem van de Middellandse Zee. De onderzoeksdata worden opgenomen in bestaande GRID computing infrastructuren. Binnen het domein milieuwetenschappen en energie kan de LIFE WATCH onderzoeksinfrastructuur worden genoemd met als doel de biodiversiteit te bestuderen, waarbij data-integratie en interoperabiliteit een rol speelt⁶⁴. Binnen het domein levens- en medische wetenschappen wil het "European Biobanking and Biomolecular Resources" project duurzame en veilige toegang geven tot biologische databronnen voor gezondheidsonderzoek op het gebied van de preventie, diagnose en behandeling van ziekten⁶⁵.

⁶⁰ zie: <<http://www.dariah.org>> [bezoekt 2 mei 2009]

⁶¹ zie: <<http://www.cessda.org/project>> [bezoekt 2 mei 2009]

⁶² zie: <<http://www.clarin.eu>> [bezoekt 2 mei 2009]

⁶³ zie: <<http://www.km3net.org>> [bezoekt 2 mei 2009]

⁶⁴ zie: <<http://www.lifewatch.eu>> [bezoekt 2 mei 2009]

⁶⁵ zie: <<http://www.bbmri.eu>> [bezoekt 2 mei 2009]

3 Infrastructuur voor duurzame opslag van en toegang tot digitale onderzoeksobjecten

In dit hoofdstuk wordt de stand van zaken beschreven betreffende de duurzame bewaring van en toegang tot onderzoeksobjecten in Nederland. De sector wetenschap is groot en heterogeen, wat het onmogelijk maakt volledig te zijn. Ook is de sector onder invloed van de mogelijkheden van de informatietechnologie sterk in beweging, waardoor bestaande inzichten en structuren aan verandering onderhevig zijn.

Voortbouwend op de informatie uit de voorgaande hoofdstukken wordt in dit hoofdstuk de huidige infrastructuur voor de opslag en toegang tot digitale onderzoeksobjecten beschreven. Het hoofdstuk bestaat uit drie onderdelen. Op de eerste plaats worden voor een aantal wetenschappelijke disciplines instellingen en organisaties genoemd die een rol spelen bij het beheer van digitale onderzoeksobjecten. Het tweede deel van dit hoofdstuk besteedt aandacht aan de veranderingen in de wetenschap onder invloed van vernieuwingen in de informatie- en communicatietechnologie. Ten slotte wordt gekeken naar financiële aspecten met betrekking tot de infrastructuur voor duurzame toegang tot onderzoeksobjecten.

3.1 Instellingen en organisaties

Er zijn niet veel instellingen in Nederland die expliciet de missie uitdragen om digitale wetenschappelijke dataobjecten te archiveren in de traditionele betekenis van het woord, waarbij gedacht moet worden aan het faciliteren van een digitaal depot, analoog aan een depot van niet-digitale objecten. Hierbij speelt de connotatie van “opbergen” een sterke rol. Wat meer voorkomt is dat instellingen in toenemende mate digitale objecten creëren, beheren en beschikbaar stellen zonder een bepaalde termijn voor ogen te hebben. Beide begrippen, “archiveren” en “beheren” groeien in de praktijk naar elkaar toe, waardoor archiefinstellingen ook zorgen voor beheer en toegang en beheersinstellingen zich geconfronteerd zien met het langetermijnbeheer van objecten die bedreigd worden door de bekende risico’s. Dit proces van naar elkaar groeien van functies en processen op het gebied van archivering en beheer maakt het van belang dat betrokken organisaties structureel gaan samenwerken en efficiënter gebruik van de beschikbare middelen maken.

In deze paragraaf worden instellingen en organisaties beschreven die in Nederland wetenschappelijke dataobjecten beheren en toegankelijk maken. Dit zijn achtereenvolgens landelijke voorzieningen voor digitale archivering, institutionele digitale bewaarplaatsen, universitaire repositories, instellingen die registerdata beheren. Tenslotte worden enkele onderzoeksinfrastructuren beschreven.

3.1.1 Landelijke voorzieningen voor digitale archivering

Op landelijk niveau kunnen DANS, het Nationaal Archief, het Instituut voor Beeld en Geluid en het e-Depot van de Koninklijke Bibliotheek gezien worden als organisaties die een rol spelen waar het gaat om de digitale archivering en beschikbaarstelling van onderzoeksdata en digitale objecten die van belang zijn voor wetenschappelijk onderzoek. Deze instellingen worden hieronder nader beschreven.

DANS⁶⁶ (Data Archiving and Networked Services) zorgt voor de opslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. Met behulp van het EASY informatiesysteem kunnen onderzoekers data deponeren en terugvinden⁶⁷. DANS werkt samen met databeheerders om ervoor te zorgen dat zo veel mogelijk onderzoeksdata vrij toegankelijk zijn, rekening houdend met de licentievoorwaarden die eraan gekoppeld zijn. Naast het archiveren en beschikbaar stellen van door derden aangeleverde datasets is DANS ook betrokken bij het opzetten van datacollecties, zoals de gedigitaliseerde historische volkstellingen⁶⁸. Naast het zelf archiveren en toegankelijk maken van onderzoeksdata heeft DANS als taak het bevorderen en ontwikkelen van een infrastructuur voor het bewaren, vinden en uitwisselen van data. In samenwerking met anderen wordt onderzoek verricht en worden nieuwe technieken ontwikkeld⁶⁹. DANS heeft een methode ontwikkeld om oude databestanden die nog aanwezig zijn bij onderzoekinstellingen te inventariseren, te selecteren, (indien noodzakelijk) te reconstrueren en op een toegankelijke manier te archiveren. Deze ADA methode (Academisch Digitaal Archiveren) is modulair van opzet, waardoor deze flexibel kan worden ingezet. DANS biedt onderdak aan het “elektronisch depot voor de Nederlandse archeologie” (EDNA)⁷⁰. EDNA bevat digitale bestanden met onderzoeksgegevens van Nederlandse archeologen. De “kwaliteitsnorm Nederlandse archeologie⁷¹” bevat de belangrijkste voorwaarden waaraan een organisatie moet voldoen bij het verrichten van archeologisch onderzoek. Documentatie en archivering zijn hiervan belangrijke onderdelen.

Het Nationaal Archief⁷² in Den Haag is een belangrijke instelling voor wetenschappelijk onderzoek, met name voor historisch onderzoek. Hoewel er op dit moment nog weinig digitale archiefbescheiden zijn, is het de verwachting dat het digitale aanbod zal groeien, waardoor de bouw van een digitaal archiefdepot noodzakelijk is. Momenteel is de bouw van dit digitaal depot in volle gang. In het verleden heeft het Nationaal Archief ervaring opgedaan met aspecten van digitale archivering middels het project “testbed digitale archivering”⁷³.

Het Instituut voor Beeld en Geluid⁷⁴ beheert en ontsluit een groot deel van het Nederlands audiovisueel erfgoed. PROARCHIVE is de naam van de dienst die Beeld en Geluid heeft ontwikkeld voor een veilige en duurzame opslag van audiovisuele archieven⁷⁵.

⁶⁶ Zie: <<http://www.dans.knaw.nl>> [bezocht 20 mei 2009]

⁶⁷ Zie: <<http://easy.dans.knaw.nl>> [bezocht 20 mei 2009]

⁶⁸ Zie: <<http://www.volkstellingen.nl>> [bezocht 19 juni 2009]

⁶⁹ DANS is bijvoorbeeld betrokken bij het ontwikkelen van een infrastructuur voor “persistent identifiers”. Ook wordt in het kader van het Mixed project software ontwikkeld om data formaten te kunnen formateren naar een duurzaam formaat. (zie: <<http://mixed.dans.knaw.nl>> [bezocht 19 juni 2009])

⁷⁰ Zie: <<http://www.edna.nl>> [bezocht 2 mei 2009]

⁷¹ Zie: <<http://www.erfgoedinspectie.nl/page/archeologie/kna>> [bezocht 2 juni 2009]

⁷² Zie: <<http://www.nationaalarchief.nl>> [bezocht 12 juni 2009]

⁷³ Zie: <<http://www.digitaalduurzaamheid.nl>> [bezocht 12 juni 2009]

⁷⁴ Zie: <<http://www.beeldengeluid.nl>> [bezocht 12 juni 2009]

⁷⁵ Zie: <<http://instituut.beeldengeluid.nl/index.aspx?ChapterID=8633>> [bezocht 15 juni 2009]

Het e-Depot van de Koninklijke Bibliotheek is een digitale archiefomgeving waarin permanente toegang tot digitale informatiebronnen wordt nagestreefd. Het e-Depot is ingericht om het Depot van Nederlandse (elektronische) Publicaties te bewaren. Daarnaast zal het plaats bieden aan het Nederlands webarchief en masters van gedigitaliseerd materiaal⁷⁶.

3.1.2 Institutionele digitale bewaarplaatsen

Institutionele digitale bewaarplaatsen of repositories bevatten digitale objecten waarvan de inhoud aansluit bij het collectiebeleid van de organisatie. Daarvan zijn er naar schatting enige tientallen in Nederland. De digitale collectie bestaat uit gedigitaliseerde analoge bronnen, van oorsprong digitale objecten aansluitend bij de doelstelling van de organisatie en digitale toegangen (metadata). De instelling heeft vaak zowel een landelijke als een internationale functie. De mate waarin deze instellingen proactief handelen waar het gaat om het garanderen van de toekomstige bruikbaarheid van de digitale objecten verschilt van instelling tot instelling. Feit is dat collecties digitale objecten vaak binnen projecten ontstaan. De tijdelijke financiering zorgt voor problemen bij het toegankelijk houden van de collectie na afloop van het project.

Bij de instellingen zijn medewerkers werkzaam die het onderhoud en beheer van de informatie-infrastructuur verzorgen. Hieronder vallen ook werkzaamheden zoals het maken van back-ups en het vernieuwen van informatiesystemen. Voor de bedreigingen voor langetermijntoegang is over het algemeen geen oplossing beschikbaar. Het risico dat de data ontoegankelijk worden is daardoor vrij groot.

Onderstaande tabel, die niet de ambitie heeft om volledig te zijn, bevat een tiental voorbeelden van institutionele repositories. De voorbeelden tonen aan dat er een grote diversiteit bestaat op het gebied van digitale collecties, die door instellingen worden beheerd. Institutionele repositories zijn voornamelijk te vinden bij alfa- en gammawetenschappen, maar ook andere disciplines kennen institutionele repositories. De exacte wetenschappen en de levenswetenschappen kennen een traditie waarbij onderzoeksdata in een internationale context worden aangeboden en beheerd waardoor deze data ook veel meer gestandaardiseerd is.

Instituut / instelling	Voorbeeld van institutionele repository ⁷⁷	Webadres ⁷⁸
Internationaal instituut voor sociale geschiedenis	Stakingen in Nederland	http://www2.iisg.nl/databases/stakingen/index.asp
Instituut voor Nederlandse Geschiedenis	Uitslagen van de verkiezingen voor de Tweede Kamer 1848-1918	http://www.inghist.nl/Onderzoek/Projecten/Verkiezingen
Fryske Academie	Kadaster van Friesland (1832) toegankelijk via GIS systeem	http://www.hisgis.nl
Meertens instituut	Nederlandse Liederbank	http://www.liederenbank.nl
3TU.Datacentrum ⁷⁹	DARELUX, meetgegevens verzameld in een stroomgebied in Luxemburg	http://www.library.tudelft.nl/darelux/index.htm

⁷⁶ Zie: < <http://www.kb.nl/dnp/e-depot/e-depot.html> > [bezocht 12 juni 2009]

⁷⁷ Er wordt per instelling één repository genoemd. Vaak beheren de instellingen meerdere repositories

⁷⁸ De genoemde websites functioneerden op 22 juni 2009

⁷⁹ Het 3TU.Datacentrum <<http://datacentrum.3tu.nl>> heeft de ambitie om uit te groeien tot een landelijke voorziening voor data-archivering in de technische wetenschappen

Nederlands instituut voor ecologie (NIOO)	Meer dan 200 datasets met onderzoeksdata	http://data.nioo.knaw.nl/
Koninklijk instituut voor Taal- Land- en Volkenkunde	Oral history collectie over het einde van de koloniale Nederlandse aanwezigheid in Azië	http://www.kitlv.nl/interview
Nederl. Interdisciplinair Demografisch Instituut	Demografische atlas	http://www.nidi.knaw.nl/nl/atlas
Nederlands Instituut voor Oorlogsdocumentatie	Gefilmde interviews van overlevenden van kamp Buchenwald	http://www.buchenwald.nl
Rijksdienst Kunsthistorische Documentatie	Beschrijvingen en afbeeldingen van hoofdzakelijk Nederlandse en Vlaamse kunstwerken van de veertiende tot en met de negentiende eeuw.	http://website.rkd.nl/Databases

3.1.3 Universitaire repositories

Verschillende universiteiten in Nederland maken wetenschappelijke onderzoeksobjecten online toegankelijk. Uiteraard hebben deze een connectie met onderzoek en onderzoekers verbonden aan de universiteit. De universitaire repositories zijn verbonden aan de universiteitsbibliotheken en hebben vaak een band met de universitaire rekencentra.

Enkele voorbeelden van universitaire repositories zijn de afdeling Igitur⁸⁰ van de universiteitsbibliotheek Utrecht, het "Digitaal Productie Centrum (DPC)⁸¹" van de Universiteit van Amsterdam en het "Repositorium⁸²" van de Universiteit Leiden. Deze repositories bevatten publicaties van onderzoekers die verbonden zijn aan de betreffende universiteit. Voor de duurzame archivering van wetenschappelijke publicaties maken de universitaire repositories gebruik van het e-Depot van de KB. In zekere zin kan het 3TU.Datacentrum, verbonden aan de universiteitsbibliotheek van de TU Delft, ook gezien worden als een universitaire repository, maar het 3TU.Datacentrum heeft wel de ambitie om diensten te verlenen op het gebied van de digitale duurzaamheid van onderzoeksdata.

3.1.4 Registerdata

Registerdata zijn data die ontstaan in een bureaucratisch proces ten behoeve van overheidsbeleid. In de wetenschap neemt de behoefte aan toegang tot deze data toe en de instellingen zijn in toenemende mate bereid om deze toegang te verschaffen⁸³. Voorbeelden zijn het KNMI "operational data centre"⁸⁴ en de databestanden die door het Kadaster⁸⁵ en het CBS gecreëerd worden.

⁸⁰ Zie: <<http://www.uu.nl/NL/Bibliotheek/igitur>> [bezoekt 18 maart 2009]

⁸¹ Zie: <<http://www.uba.uva.nl/dpc/>> [bezoekt 18 maart 2009]

⁸² Zie: <<http://www.bibliotheek.leidenuniv.nl/docenten-onderzoekers/publiceren/leids-repositorium/>> [bezoekt 18 maart 2009]

⁸³ De overheid streeft naar betere toegang tot door de overheid gecreëerde data. De Gideon beleidsnota is een goed voorbeeld van dit streven. Zie: <<http://www.geonovum.nl/dossiers/gideon>> [bezoekt 18 maart 2009].

⁸⁴ Zie: <<http://kodac.knmi.nl>> [bezoekt 18 maart]

⁸⁵ Een aantal bestanden van het Kadaster is voor gebruik door wetenschappers via DANS verkrijgbaar.

3.1.5 Enkele onderzoeksinfrastructuren

Binnen de exacte en levenswetenschappen zijn in de loop der tijd omvangrijke onderzoeksinfrastructuren ontstaan die bestaan uit grote dataverzamelingen en geavanceerde geautomatiseerde analysetools. Hieronder worden er enkele nader bekeken. Deze voorbeelden maken duidelijk dat permanente, duurzame aandacht voor deze onderzoeksdata vereist is om optimaal gebruik van deze bestanden mogelijk te maken. De doorgaans hoge investeringen die nodig zijn om deze databestanden aan te leggen kunnen hiermee mede gerechtvaardigd worden.

Opvallend is dat duurzame data-archivering bij geen enkel initiatief een hoge prioriteit heeft. Er is uiteraard wel aandacht voor databeheer ten behoeve van het dagelijkse gebruik van de onderzoeksdata, ook in de toekomst. Onderstaand citaat uit het BIG GRID projectvoorstel is wat dit betreft veelzeggend: *For storing archive data in the social and cultural science domains, the required reliability level of the persistent storage is higher than in the physics and astronomy domains, where copies of the data will be available in other locations as well*⁸⁶. Opslag van de data op meerdere plaatsen, de essentie van het idee achter het grid, wordt dus beschouwd als een maatregel om de duurzaamheid van de data te verbeteren.

Op het gebied van oceanografie is het "National Oceanographic Data Committee (NODC⁸⁷) van belang als platform voor de uitwisseling van oceanografische en marine onderzoeksdata en informatie. Daarnaast verschaft het NODC adviesdiensten op het gebied van data management.

Binnen de natuurwetenschappen worden de grootste dataopslag en -verwerkingssystemen ontwikkeld. Een goed voorbeeld hiervan is de infrastructuur die ontwikkeld wordt in het kader van het LOFAR-project⁸⁸ dat een budget heeft van 148 miljoen euro. In een gebied met een diameter van 100 kilometer worden 15.000 "low frequency array" antennes geplaatst. Deze antennes leveren, afhankelijk van de interval en dichtheid waarmee de observaties worden ingesteld, zeer grote hoeveelheden onderzoeksdata op. Een observatie van 4 uur, waarbij 8 sensoren elke seconde een bandbreedte van 10 kHz registreren, levert 6 terabyte aan data op. In de projectplanning van LOFAR is opgenomen dat de dataopslagcapaciteit in de loop der tijd zal toenemen. Er is uiteindelijk voorzien in een opslagcapaciteit van 20 petabyte, maar zelfs dat zal niet genoeg zijn om alle metingen op te slaan. Langetermijnarchivering van deze grote datavolumes zal extreem kostbaar zijn. Daarom heeft LOFAR een dataverwerkingsmodel opgesteld waarin de uiteindelijke geaggregeerde gegevensproducten op geautomatiseerde wijze tot stand komen. De archivering van de data is in principe de verantwoordelijkheid van de gebruiker of het onderzoekscentrum dat behoefte heeft aan bepaalde data die door de antennes worden geleverd. Een beschrijving van het LOFAR data-archief en de dataverwerkingsmethoden is gepubliceerd in het document "*LOFAR Archive and Reprocessing Requirements*⁸⁹". Het interessante aan de data-archiefbenadering van LOFAR is dat er rekening wordt

⁸⁶ Afkomstig uit het BIG GRID projectvoorstel p33 Zie: <www.biggrid.nl/about/BIGGRID-proposal.pdf> [bezoekt 20 maart 2009]

⁸⁷ Zie: <<http://www.nodc.nl>> [bezoekt 17 maart 2009]

⁸⁸ Zie: <www.lofar.org> [bezoekt 29 maart 2009]

⁸⁹ Zie: <http://www.lofar.org/operations/lib/exe/fetch.php?media=public:documents:lofar_archive_requirements_rev0.9.pdf> Het betreft hier versie 0.9 met de status "draft" [bezoekt 9 maart 2009]

gehouden met een aantal onbekenden, zoals de mogelijkheid dat gebruikers bepaalde gegevens opnieuw willen verwerken, waarvoor dus opslag- en verwerkingscapaciteit beschikbaar moet zijn, en het volume en dichtheid van de data die aangeleverd worden door de antennes. Op basis van aannames schat men tussen de 1 en 4 petabyte per jaar aan dataopslag nodig te hebben. Opvallend is dat in het document geen aandacht is voor de manier waarop de onderzoeksdata duurzaam dienen te worden gearhiveerd en te worden gebruikt. Dit kan betekenen dat men het maken van back-ups en reservekopieën als vanzelfsprekend beschouwt en er op vertrouwt dat de ontwikkelde dataopslaginfrastructuur robuust genoeg is om verlies van data te voorkomen.

Nauw verbonden aan het LOFAR project is het Target programma dat in april 2009 16 miljoen euro subsidie ontving en een budget heeft van 32 miljoen euro⁹⁰. Target beoogt een duurzaam economisch cluster van intelligente informatiesystemen op te bouwen in Noord-Nederland, in het bijzonder gericht op databeheer van zeer grote hoeveelheden data die afkomstig zijn uit sensornetwerken. Binnen Target zullen de Rijksuniversiteit Groningen, de sterrenkundige onderzoeksschool NOVA, de stichting Astron en het bedrijfsleven nieuwe complexe en schaalbare datasystemen ontwikkelen en bestaande verbeteren. Target verschaft een proeftuin die het doorontwikkelen naar concrete markttoepassingen mogelijk maakt en deelnemers in staat stelt om in vervolgtrajecten verdere producten en diensten te ontwikkelen.

Ook binnen de medische wetenschappen is een omvangrijke infrastructuur ontstaan voor digitale dataverzamelingen. Het rapport "Van gegevens verzekerd. Kennis over volksgezondheid in Nederland"⁹¹ uit eind 2008 bevat een advies van de Gezondheidsraad aan het ministerie van Volksgezondheid, Welzijn en Sport over maatregelen die nodig zijn om de beschikbaarheid te verzekeren van gegevens over de Nederlandse volksgezondheid ten behoeve van beleid en wetenschap. Het is niet goed gesteld met de digitale duurzaamheid van epidemiologische gegevensverzamelingen. Het rapport stelt: *Lang lopende gegevensverzamelingen van hoge kwaliteit zijn dan ook zowel voor volksgezondheidsbeleid en gezondheidszorg als voor wetenschappelijk onderzoek goud waard. De onderzoeksgroepen die in staat en bereid zijn dit soort onderzoek uit voeren blijken in de praktijk al jaren moeilijkheden ondervinden bij de instandhouding van de infrastructuur van hun onderzoek*⁹². Het rapport adviseert onder andere om gegevens uit verschillende bestanden aan elkaar te koppelen (bij voorkeur op basis van het burger service nummer) en stelt dat structurele financiering van longitudinaal epidemiologisch onderzoek gewenst is. Bij deze aanbevelingen speelt de langetermijnduurzaamheid van de onderzoeksbestanden een grote rol. Het rapport geeft aan welke juridische, technische en organisatorische maatregelen genomen dienen te worden om de duurzaamheid van deze medische onderzoeksdata te waarborgen. Het rapport bevat beschrijvingen van ruim 35 Nederlandse populatiecohorten. Dit zijn bestaande informatiesystemen die medische gegevens bevatten over bepaalde groepen Nederlanders⁹³.

⁹⁰ Zie: <<http://www.rug.nl/target/nieuws/actueel/persbericht>> [bezocht 3 mei 2009]

⁹¹ Het rapport "Van gegevens verzekerd. Kennis over Volksgezondheid in Nederland" is te vinden op de website van de Gezondheidsraad, zie: <<http://www.gr.nl/pdf.php?ID=1765&p=1>> [bezocht 23 juni 2009]

⁹² "Van gegevens verzekerd" p27.

⁹³ Een voorbeeld van een populatiecohort is het "Hongerwintercohort" dat sinds 1994 gegevens verzameld van tussen 1 november 1943 en 28 februari 1947 in het Wilhemina gasthuis

Het feit dat delen van onderzoeksdata, waarvoor een duurzame informatie-infrastructuur vereist is, van belang is voor de wetenschap wordt duidelijk aan de hand van het "Parelsnoer"⁹⁴ initiatief. In dit project, opgezet in 2007 door de acht Nederlandse universitaire medische centra, worden klinische data en biomaterialen verzameld. Hierdoor kan de patiënt beter behandeld worden, kan de wetenschap zich beter ontwikkelen en kan er nieuwe productontwikkeling plaatsvinden. Het initiatief richt zich in eerste instantie op acht ziektebeelden (de parels). Het koppelen en delen van onderzoeksgegevens is een belangrijke functie van het Parelsnoer initiatief. Om dit te kunnen doen is het belangrijk dat de onderzoeksdata van goede kwaliteit zijn en goed worden beheerd.

3.2 Wetenschap en ICT

De wetenschap verandert onder invloed van het toegankelijk worden van grote hoeveelheden data. De wetenschap is veel meer "data driven" geworden. Dit vereist maatregelen voor de beschikbaarheid en verwerkingsmogelijkheden van onderzoeksdata en daarmee de duurzame toegang.

De toepassing van ICT in het wetenschappelijk onderzoek heeft geleid tot de ontwikkeling van wat e-science (of e-research, ref. het model op pagina 18) wordt genoemd. Dit is een nieuw onderzoeksparadigma dat veel invloed heeft op de manier waarop met onderzoeksdata wordt omgegaan, nu en in de toekomst. e-Science omvat samenwerkingsverbanden waarbij gebruik wordt gemaakt van gedistribueerde rekenkracht en databronnen. e-Science zorgt voor een verandering van de manier waarop wetenschap bedreven wordt. Nederland heeft een goede netwerkinfrastructuur (de SURF6 infrastructuur gerealiseerd in het Gigaport project) en vernieuwde supercomputers met namen als Huygens en Blue Gene.

Momenteel vormen de schaalgrootte en complexiteit van gedistribueerde data een belangrijke uitdaging. De natuurkunde en chemie lopen voorop waar het gaat om het gebruik van "large-scale" ICT infrastructuren. Wetenschappelijk onderzoek wordt steeds meer internationaal. De trend is dat er grootschalige virtuele onderzoeksfaciliteiten ontstaan, bijvoorbeeld bij bio-informatica. Grootschalige onderzoeksfaciliteiten zijn belangrijk voor de Nederlandse kenniseconomie. Componenten van deze infrastructuur zijn: computers, dataopslag, visualisatie-tools en instrumenten.

Voorbeelden van projecten waarbij grote hoeveelheden data worden geproduceerd zijn het LOFAR project en het internationale LHC (Large Hadron Collider) project. De Nationale Computer Faciliteiten (Blue Gene/Huygens) en het Big Grid project zijn belangrijke componenten voor de opslag en analyse van deze data. Bij levenswetenschappen zijn DNA-sequenties grote consumenten van dataopslag.

De KNAW wil dat er meer wordt samengewerkt tussen beoefenaars van de geesteswetenschappen en informatie- en communicatietechnici uit de beta-wetenschappen. Deze "computational humanities" vereist onder andere een duurzame data-infrastructuur waarbinnen het aanleveren, bewaren en hergebruik van onderzoeksbestanden mogelijk is.

We staan pas aan het begin van grote veranderingen die het gevolg zijn van het beschikbaar komen van grote hoeveelheden onderzoeksdata. We staan ook pas

levendgeborenen plus hun moeders en kinderen, zie <<http://www.hongerwinter.nl>> [bezocht 28 juni 2009]

⁹⁴ Zie: <<http://www.parelsnoer.org>> [bezocht 25 juni 2009]

aan het begin van het inrichten van duurzame dataopslag en -beheerfaciliteiten ten behoeve van dit wetenschappelijk onderzoek. Een interessant initiatief om te onderzoeken op welke manier de wetenschap kan profiteren van de toenemende beschikbaarheid van onderzoeksdata is het "digging into data"⁹⁵ initiatief, dat middelen beschikbaar stelt voor initiatieven om op innovatieve manier om te gaan met het groeiend aanbod van digitale onderzoeksdata. Hiermee wordt ook de noodzaak van aandacht voor langetermijnarchivering van de onderzoeksdata aangetoond.

3.3 Financiële aspecten

De financiering van publiek wetenschappelijk onderzoek in Nederland is afkomstig uit drie geldstromen. De eerste geldstroom is direct afkomstig van de overheid zonder tussenkomst van intermediairs als NWO en KNAW en wordt bijvoorbeeld gebruikt om universiteiten te financieren. De tweede geldstroom wordt door NWO, KNAW en internationale onderzoeksorganisaties rechtstreeks uitgekeerd aan onderzoekers, onderzoeksprojecten en onderzoeksprogramma's. Zowel de kwaliteit van concurrerende onderzoeksvoorstellen als de reputatie van onderzoekers is van invloed op de toekenning van subsidies. Het proces van *peer review*, de beoordeling van onderzoeksoutput en onderzoeksvoorstellen door collega-onderzoekers, speelt daarbij een belangrijke rol. De derde geldstroom is projectgebonden financiering, vaak van private instellingen maar ook van ministeries.

In de publicatie "Dertig jaar publieke onderzoeksfinanciering in Nederland 1975 – 2005"⁹⁶ wordt geconcludeerd dat het onderzoekssysteem en de financieringsstromen in de loop der tijd onmiskenbaar complexer zijn geworden. Het aantal organisaties, programma's, regelingen en organen is sterk gegroeid. Ook stelt het rapport dat de groei van de publieke financiering van onderzoek in Nederland achterblijft ten opzichte van de private financiering, ten opzichte van het nationaal inkomen, en ten opzichte van het buitenland.

Hoewel er in Nederland enkele organisaties bestaan die als taak hebben om onderzoeksdata te archiveren en toegankelijk te houden, bestaat er geen consensus over de financieringsmodellen die toegepast dienen te worden voor de langetermijntoegang tot onderzoeksdata. Dit komt omdat de meeste onderzoeksprojecten een eenmalige projectfinanciering hebben, waarbij documentatie en overdracht van de digitale resultaten naar een data-archief niet opgenomen zijn in de begroting. Daarnaast zijn de rollen en verantwoordelijkheden met betrekking tot toegang tot data en conservering niet duidelijk. Onderzoekers verwachten dat ze gratis toegang hebben tot wetenschappelijke data en zij niet hoeven op te draaien voor de kosten van opslag en beheer. Ook ontbreekt bij dataproducenten vaak de bereidheid om structureel samen te werken met alle belanghebbenden om interoperabiliteit te stimuleren op het gebied van documentatiestandaarden, bestandsformaten, hardware en software.

De financiering van de ICT-infrastructuur komt voornamelijk van specifieke investeringsprogramma's, bijvoorbeeld van NWO⁹⁷ en stimuleringsprogramma's,

⁹⁵ Zie: <<http://www.diggingintodata.org>> [bezocht 20 juni 2009]

⁹⁶ Anouschka Versleijen (red.) *Dertig jaar publieke onderzoeksfinanciering in Nederland 1975-2005* (Den Haag 2007), online beschikbaar op: <<http://www.rathenau.nl/downloadfile.asp?ID=1214>> [bezocht 28 mei 2009]

⁹⁷ Via de subsidiewijzer van NWO kan informatie verkregen worden over de bestaande subsidieprogramma's. Zie: <http://www.nwo.nl/nwohome.nsf/pages/SPPD_7AK3L2> [bezocht 20 juni 2009]

zoals BSIK (Besluit Subsidies Investerings Kennisinfrastructuur⁹⁸). De organisatie van de ICT-infrastructuur is versnipperd tussen organisaties als NWO, de Stichting Nationale Computer Faciliteiten (NCF) en SARA reken- en netwerkdiensten, alsmede BSIK en NWO programma's. De economische levensduur van een ICT-systeem is slechts vijf tot zeven jaar, maar de ICT infrastructuur voor wetenschappelijk onderzoek bevat elementen die langer relevant zijn dan de levensduur van het project. Ook is de verantwoordelijkheid gedistribueerd over verschillende partijen. ICTRegie adviseert daarom om al het ontwikkel- en uitvoeringswerk voor de ICT-infrastructuur voor het wetenschappelijk onderzoek onder te brengen onder de paraplu van SURF. Tevens pleit ICTRegie voor het oprichten van een e-science research center dat multidisciplinair onderzoek verricht⁹⁹.

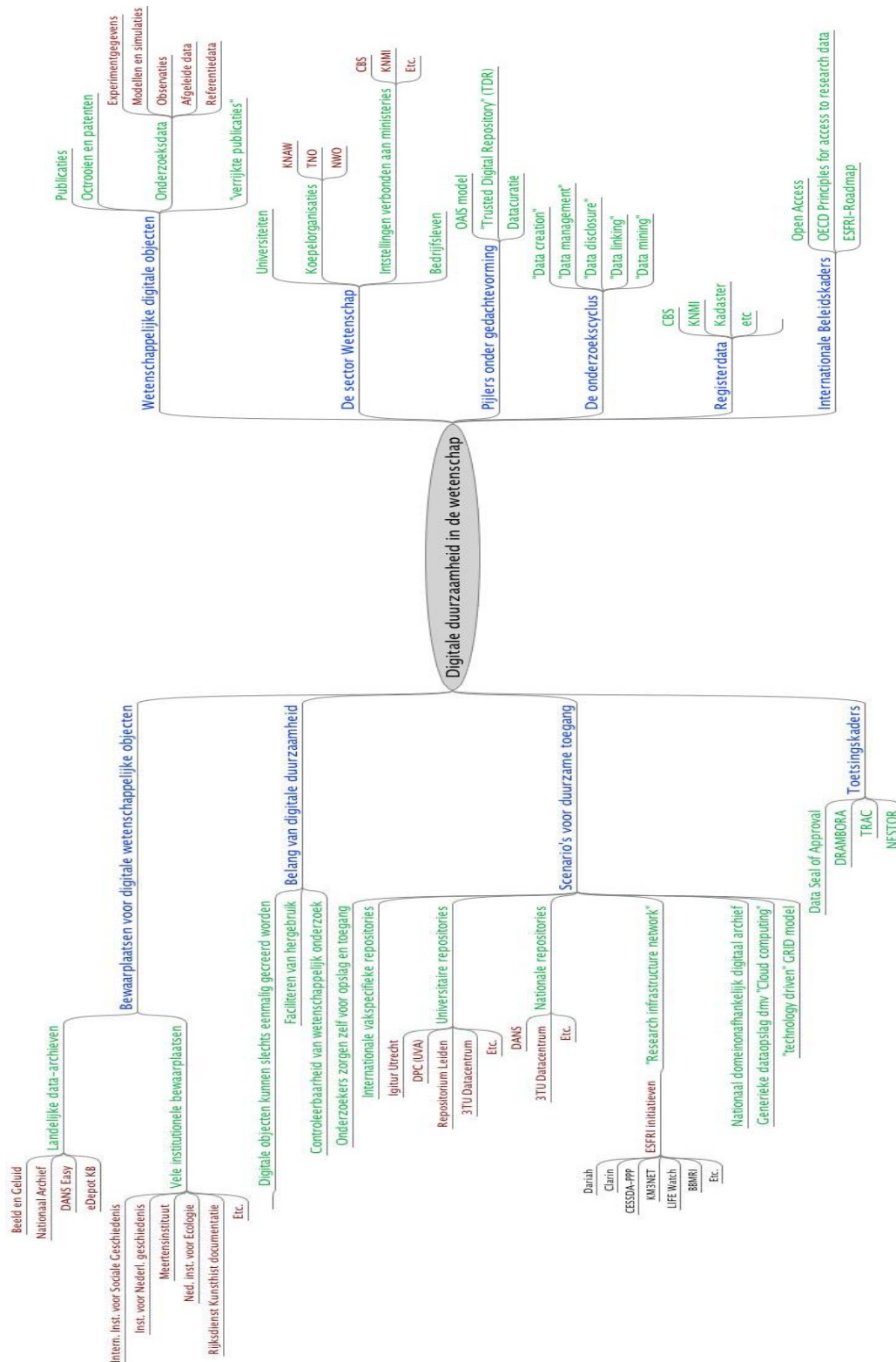
Over de kosten van digitale archivering is nog niet veel bekend. Het project LIFE (Life Cycle Information for E-Literature) heeft een model ontwikkeld voor de "digitale levenscyclus" van elektronische bestanden en rekenmodellen ontworpen die momenteel alleen nog retrospectief werken. In fase 3 gaat men proberen om ook modellen te ontwerpen voor het begroten van kosten. Deze kosten bestaan uit de volgende onderdelen: datacreatie, data acquisitie, "ingest", bitstream-preservation, "content preservation" en toegang.

⁹⁸ Zie: <<http://www.senternovem.nl/BSIK/>> [bezoekt 12 mei 2009]

⁹⁹ Zie: Prof. Dr. Ir. L.J.M. Nieuwenhuis, *Towards a competitive ICT infrastructure for scientific research in the Netherlands* (december 2008), te vinden op: <http://www.ictregie.nl/publicaties/nl_08-NROI-258_Advies_ICT_infrastructuur_vdef.pdf> [bezoekt 20 maart 2009]

4 Conclusies en aanbevelingen

Onderstaande figuur toont schematisch de uitkomsten van deze verkenning naar de stand van zaken op het gebied van de duurzame opslag van en toegang tot digitale wetenschappelijke onderzoeksobjecten.



Het doel van deze verkenning was om de situatie in Nederland te beschrijven ten aanzien van de duurzaamheid van wetenschappelijke digitale onderzoeksobjecten. Dit levert in eerste instantie een tegenstrijdig beeld op. Feit is dat er in Nederland maar weinig organisaties en initiatieven zijn op het gebied van de digitale archivering van onderzoeksdata die ervoor zorgen dat na afloop van een project de onderzoeksdata duurzaam worden gearhiveerd. Daar staat tegenover dat er een grote heterogene data-infrastructuur is ontstaan die vele vormen van dataverwerking en -beheer mogelijk maken. De inrichting en aanpak van deze infrastructuur is enerzijds sterk disciplinegericht en anderzijds gebaseerd op bestaande organisaties die een rol spelen in het wetenschapsbedrijf, zoals bibliotheken en archieven.

Het is belangrijk dat wetenschappers als domeindeskundigen betrokken worden bij het inrichten van een optimale duurzame data-infrastructuur. Ook is specifieke kennis vereist op het gebied van data-archivering, bijvoorbeeld op het gebied van auteursrecht, Open Access, open standaarden en "repository" systemen. Er bestaat een aantal toetsingskaders, waaronder het "Data Seal of Approval" dat als leidraad kan dienen bij het inrichten van een bewaarplaats voor digitale wetenschappelijke objecten.

Wat de verkenning heeft duidelijk gemaakt is dat digitale data van belang voor de wetenschap vele verschijningsvormen kan hebben. Het meest duidelijk zijn de wetenschappelijke publicaties en octrooien en patenten, waarvoor inmiddels databewaarplaatsen zijn ingericht. Ook voor de archivering van datasets die het resultaat zijn van afgesloten onderzoek in de geestes- en sociale wetenschappen is een data-archief opgezet. Maar daarnaast bestaat er nog een groot aantal andere soorten wetenschappelijke dataobjecten, waarvan de aard, waarde en vereiste duurzaamheid nog niet eenduidig is vast te stellen. Nader onderzoek is nodig om een goede classificatie te verkrijgen van digitale dataobjecten die een rol spelen in het wetenschapsbedrijf. Uiteraard dient dit te gebeuren in de vorm van een samenwerking tussen wetenschappers, informatiekundigen en archiefspecialisten. Het is aan te bevelen een werkgroep op te richten die onderzoekt welke soorten digitale onderzoeksobjecten een rol spelen bij wetenschappelijk onderzoek. Op basis van deze classificatie kan bepaald worden welke objecten duurzaam bewaard moeten worden. Vervolgens dient dan vastgesteld te worden op welke wijze deze bewaring het beste georganiseerd kan worden. Er zijn inmiddels toetsingskaders beschikbaar om deze werkzaamheden uit te voeren.

Binnen de wetenschap speelt de informatietechnologie een grote rol. De wetenschapper verwacht toegang te hebben via internet tot alle voor het vakgebied relevante wetenschappelijke literatuur en databanken met essentiële onderzoeksdata. De analyse-, simulatie-, en visualisatiesoftware van onderzoeksdata wordt steeds geavanceerder en biedt de onderzoeker gereedschap om de kwaliteit van het wetenschappelijk onderzoek te verhogen. De financiers van het wetenschappelijk onderzoek realiseren zich in toenemende mate dat de duurzaamheid van de data en tools van belang is. Maar een eenduidige visie en strategie op de langetermijnbewaring van onderzoeksdata bestaat er alleen op het gebied van digitale publicaties en voor een aantal disciplinegebonden wetenschappelijke datasets.

Het vastleggen van selectiecriteria van dataobjecten die in aanmerking komen voor bewaring dient per wetenschappelijke discipline verder uitgewerkt te worden. Gezien de groei van het aantal digitale objecten is het niet mogelijk ze allemaal te archiveren.

De opslagcapaciteit van onderzoeksdata groeit, maar de kosten hiervoor bedragen een fractie van de totale kosten voor digitale duurzaamheid. Het beheer

van de onderzoeksdata is aanzienlijk duurder. Hieronder vallen kosten voor documentatie, beschikbaarstelling, kwaliteitscontrole en verwerking. De financiers van wetenschappelijk onderzoek zijn bij uitstek de partij die sturend kan optreden bij het verbeteren van de duurzaamheid van de wetenschappelijke data-infrastructuur door wetenschappers te verplichten data goed te documenteren en over te dragen aan een duurzame bewaarplaats. Ook onderzoeksprogramma's en thema's dienen aandacht te schenken aan de lange termijn bewaring van en toegang tot de onderzoeksdata die het resultaat zijn van deze projecten.

De wetenschap dient meer erkenning te geven aan activiteiten rondom het maken, beheren en de beschikbaarstelling van onderzoeksdata. Van oudsher gaat de meeste wetenschappelijke erkenning uit naar de publicatie. Door ook professionele erkenning te krijgen voor bijdragen aan de data-infrastructuur, neemt de waarde van onderzoeksdata toe, waardoor er ook meer aandacht zal zijn voor de duurzame archivering van de onderzoeksdata.

Met name binnen de exacte wetenschappen is het vaak vanzelfsprekend om onderzoeksdata te hergebruiken, maar bij een aantal wetenschapsgebieden wordt nog onvoldoende gekeken of bestaande onderzoeksbestanden hergebruikt kunnen worden. De bestaande "belonings"-initiatieven om dit hergebruik te stimuleren dienen gecontinueerd en uitgebreid te worden.

Scholing en training op het gebied van de aanleg en het (her)gebruik van onderzoeksdata is van belang om ervoor te zorgen dat wetenschappers ook daadwerkelijk meewerken aan de realisering van een duurzame data-infrastructuur. Hierbij is internationale afstemming van groot belang.

Bij alle hierboven genoemde aanbevelingen is het van belang dat deze in samenwerking met alle betrokkenen wordt uitgevoerd. Een goede vorm om dit te organiseren zou kunnen zijn in de vorm van werkgroepen die een aantal deelaspecten nader uitwerken, bijvoorbeeld op het gebied van financieringsmodellen, metadata-standaarden en datacuratie-tools. Op een aantal terreinen is al een begin gemaakt met dit overleg.