

SURFshare WP6 Datacuratie en Digitale Duurzaamheid

Project

“Waardevolle Data & Diensten”

Eindrapportage



Auteur(s): Jeroen Rombouts, Jaap de Lange, Alenka Prinčič, Judith van Peet-de Wit (TU Delft), Laurents Sesink (DANS), Erik Soonieus (Alares), Maarten van Bentum (UT), Leon Osinski (TU/e)

Datum: 14 juli 2009

Versie: 1.0
Status: G



De Creative Commons Naamsvermelding–Niet–commercieel 3.0 Nederland Licentie is van toepassing op dit werk. Ga naar <http://creativecommons.org/licenses/by-nc/3.0/nl/> of stuur een brief naar Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, VS om deze licentie te bekijken

SURF Share WP6 “Waardevolle Data & Diensten”	
Auteur	<i>TU Delft Library</i>
Opdrachtgever	<i>SURF foundation, Wilma Mossink</i>
Datum	<i>14 juli 2009</i>
Status (N/O/G/A/V)*	<i>G</i>

* (Nieuw/Ontwikkeling/Gereed/Akkoord/Vervallen)

Versiebeheer

Versie	Datum	Auteur	Aanpassingen
0.1	240609	ES, JR	Eerste versie
0.2	250609	JL, JR	Correcties H1 t/m 3
0.3	290609	AP, JP, JL, LO	Uitbreiding H4 en Bijlagen, correcties totale document
0.4	300609	LS, MB, JR	Acquisitie ervaringen ingevoegd, opmaak aangepast
1.0	140709	JP, JR	Commentaar stuurgroep verwerkt

Distributielijst

Versie	Datum	I/R/A*	Distributie
0.1	240609	R	JR, AP, JL, IA
0.3	290609	R	JR, JP, JL, IA
0.4	010709	A	Stuurgroep en projectgroep
1.0	150709	I	SURF, Stuurgroep, projectgroep

* I = ter info
R = ter review
A = ter accordering

“Datasets zijn de kroonjuwelen van wetenschappelijk onderzoek”



(http://www.flickr.com/photos/matt_lee/2058799270/)

Inhoud

1. INLEIDING	1
1.1. Aanleiding.....	1
1.2. Aanpak.....	2
1.3. Leeswijzer.....	3
2. LITERATUURSCAN EN ACQUISITIE ERVARINGEN	4
2.1. Conclusies.....	4
2.2. Samenvattingen deelaspecten.....	5
3. VELDONDERZOEK	11
3.1. Casestudies.....	11
3.2. Bevindingen uit de casestudies.....	13
3.3. Expertbijeenkomst; toetsing van de bevindingen.....	22
4. PROGRAMMA VAN WENSEN	24
5. BIJLAGEN	27
5.1. Bronnen.....	27
5.2. Overige projectresultaten.....	29

1. Inleiding

Lux aeterna luceat eis
(Laat het eeuwige licht hen beschijnen)
Motto van DARELUX

Duurzame opslag van data is zeldzaam in de technische wetenschap, zowel op nationaal als internationaal niveau. Een historische analyse van de onderzoeksdata is hierdoor niet eenvoudig. Onderzoek is bovendien vaak kostbaar. De 3TU.Federatie startte daarom met het 3TU.Datacentrum, onder initiatief van de bibliotheken van de Technische Universiteit Delft (TU Delft), Technische Universiteit Eindhoven (TU Eindhoven) en de Universiteit Twente (UT). Dit datacentrum moet zorgen voor de goed gedocumenteerde opslag en langdurige toegang tot technisch-wetenschappelijke onderzoeksdata. Hiermee is een duurzame beschikbaarheid van het hele Nederlandse technisch-wetenschappelijke erfgoed gewaarborgd.

Om inzicht te krijgen in de functionele wensen en eisen vanuit de dataproducenten aan een datacentrum startte binnen het project 3TU.Datacentrum, in samenwerking met DANS, het onderzoek 'Waardevolle Data & Diensten' als bijdrage aan het SURFshare programma. Deze rapportage vormt een beschrijving van de functionele wensen aan het datacentrum.

DANS – Data Archiving and Networked Services – is een instituut van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), dat mede wordt ondersteund door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

DANS zorgt sinds haar oprichting in 2005 voor de opslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. Daartoe ontwikkelt DANS zelf duurzame archiveringsdiensten, bevordert dat anderen dat doen, en werkt samen met databeheerders om zoveel mogelijk data vrij beschikbaar te krijgen voor gebruik in het wetenschappelijk onderzoek.

1.1. Aanleiding

Een datacentrum is een succes wanneer een brede selectie van datasets uit de technische wetenschappen is opgeslagen en wordt hergebruikt¹. Om dit te bereiken zullen onderzoekers onderzoeksdata die zich lenen voor hergebruik beschikbaar moeten stellen voor derden, door ze toe te voegen aan het datacentrum. Het is echter niet vanzelfsprekend dat onderzoekers hun zelfvergaarde data zullen archiveren en beschikbaar stellen aan het datacentrum. Het zal extra werk kosten, terwijl de baten voornamelijk voor andere onderzoekers lijken te zijn. Om opties voor het verlagen van deze en andere drempels voor het gebruik van het datacentrum enerzijds, en anderzijds de behoeften voor bepaalde typen onderzoek waarin het datacentrum zou kunnen voorzien te identificeren, is het project 'Waardevolle Data & Diensten' gestart.

Het doel van het project 'Waardevolle Data & Diensten' is daarom het volgende:

¹ Onder hergebruik wordt in dit stuk verstaan zowel het inzien van de data, het reproduceren van het onderzoek voor verificatie als het gebruiken van de data voor ander onderzoek, bijvoorbeeld voor validatie van modellen of (aanvullend) nieuw onderzoek.

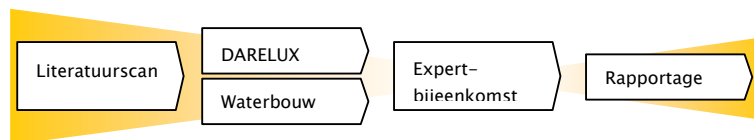
Inventariseer hoe het datacentrum dient te worden ingericht om wetenschappers te enthousiasmeren hun onderzoeksgegevens te archiveren in het datacentrum en de onderzoeksgegevens beschikbaar te stellen en geschikt te maken voor hergebruik door derden.

Doelstelling uit projectplan 'Waardevolle Data & Diensten'

De resultaten uit het onderzoek vormen de functionele wensen voor de inrichting en implementatie van het datacentrum.

1.2. Aanpak

Om te komen tot dit programma van functionele wensen voerde het projectteam een onderzoek uit. Daarvoor werd allereerst een literatuuronderzoek verricht. Vanuit dit literatuuronderzoek stelde het projectteam interviewvragen op voor het uitvoeren van expertinterviews als onderdeel van twee casestudies.

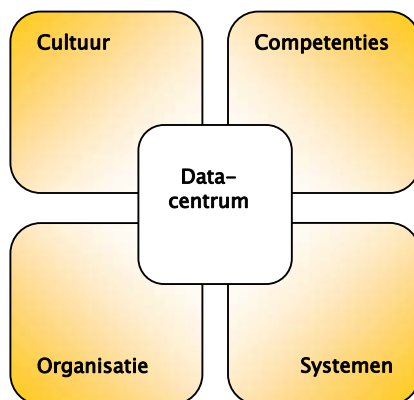


Figuur 1: Opbouw van het onderzoek

De informatie uit de casestudies vormde de input voor een expertbijeenkomst, waarbij technische onderzoekers van de TU Delft, TU Eindhoven en de UT (Twente) aanwezig waren. Tijdens de bijeenkomst vond een laatste inventarisatie van functionele wensen plaats. Daarnaast diende de bijeenkomst voor het toetsen van de opgestelde wensen en werd met de deelnemers bekeken of deze wensen generiek waren. De rapportage vormde het eindproduct van het onderzoek waarin alle elementen bijeenkwamen. Voor een uitgebreidere beschrijving van de projectaanpak wordt verwezen naar het projectplan 'Waardevolle Data & Diensten' zoals ingediend bij SURFfoundation in december 2008.

1.2.1. Integrale analyse

Het succes van een datacentrum zal voor een deel liggen in het systeem dat wordt aangeboden. Zoals bij veel technologische innovaties is het succes echter ook afhankelijk van de bereidheid van de gebruikers, de vaardigheden om er gebruik van te maken en de sturing door de organisatie.



Figuur 2: Integrale aanpak voor de analyse

In de analyse die binnen dit onderzoek is gedaan, wordt verder gekeken dan alleen naar de technische oplossingen en de inrichting van het systeem.

Organisatie

Een belangrijk onderdeel binnen bij het realiseren van een datacentrum is het vaststellen van een toekomstbeeld. Zowel de organisatie rondom het datacentrum als de organisaties die gebruik willen maken van het datacentrum zullen een heldere visie op dit gebruik moeten opstellen. Deze 'voortschrijdende' visie wordt vertaald in een strategie om de gewenste positie te bereiken. Een duidelijke visie op dit vlak vergroot het uiteindelijke succes.

Cultuur

De cultuur binnen een onderzoeksgroep of discipline is mede bepalend voor de mate waarin kennis en onderzoeksgegevens gedeeld worden. Het doen van onderzoek is kostbaar en in veel gevallen een passie. Dit kan leiden tot barrières voor onderzoekers om bij te dragen aan een open cultuur. Tegelijkertijd blijkt behoefte te bestaan aan efficiëntere en effectievere opslag van gegevens. Het beschikbaar stellen van een centraal datacentrum zal invloed hebben op de cultuur binnen de wetenschappelijke wereld. De vraag is hoe het datacentrum het best aan kan sluiten op de verschillende heersende en toekomstige culturen.

Systemen

Belangrijk binnen het ontwikkelen van het datacentrum is het technisch inrichten van het systeem. Naast het inrichten van een technisch datacentrum zijn er meer systemen die het juiste gebruik van het datacentrum kunnen stimuleren. Het is noodzakelijk inzichtelijk te krijgen welke systemen een bijdrage kunnen leveren. Het technisch functioneren alleen is niet bepalend voor het succes; de inbedding in de organisatie zorgt voor succesvol gebruik.

Competenties

De competenties en vaardigheden door de hele wetenschappelijke keten zijn ook bepalend voor het succes van collectieopbouw en het hergebruik van onderzoeksdata in een datacentrum. Indien noodzakelijke competenties en vaardigheden ontbreken om invulling te geven aan de nieuwe werkwijze zijn trainingen en instructies, richtlijnen en dergelijke op persoonlijk en organisatieniveau van toegevoegde waarde.

Door de analyse integraal te benaderen kan een programma van functionele wensen opgesteld worden dat leidt tot succesvol gebruik van het datacentrum.

1.3. Leeswijzer

In de rapportage wordt de totstandkoming van het programma van functionele eisen beschreven. Hoofdstuk 2 geeft de resultaten uit de literatuurscan. De literatuurscan vormt de input voor de casestudies en de expertbijeenkomst die beschreven staan in hoofdstuk 3. Hoofdstuk 4 geeft een overzicht van de aanbevelingen voor het datacentrum. Dit hoofdstuk vormt het concrete programma van functionele wensen.

2. Literatuurscan en acquisitie ervaringen

Hieronder volgt een beschrijving van de resultaten van het literatuuronderzoek. De resultaten van de literatuurscan zijn in eerste instantie gebruikt als basis voor de interviewvragen. Zowel de resultaten van de literatuurscan als de interviewvragen kunt u terugvinden in de documenten genoemd in bijlage 5.2. De aanbevelingen en kenmerken die het literatuuronderzoek opleverde zijn tevens beschreven om ook later in het project dienst te kunnen doen bij de ideegeneratie. De resultaten na de samenvatting zijn gegroepeerd in vier categorieën: organisatie, cultuur, competenties en systemen. Deze categorieën komen overeen met de categorieën voor de aanbevelingen in de case beschrijving.

2.1. Conclusies

Het bestaansrecht van een datacentrum wordt vooral ontleend aan de mate waarin de opgeslagen datasets worden hergebruikt.

De mate van hergebruik is afhankelijk van

- de toegankelijkheid voor verschillende groepen dataconsumenten;
- de ondersteuning die het datacentrum kan bieden aan zowel dataleveranciers als dataconsumenten;
- de 'waarde' van de datacollectie, die o.a. wordt bepaald door de herhaalbaarheid van de processen waarmee de data tot stand zijn gekomen, de kwaliteit van de data, en voor longitudinaal onderzoek bijvoorbeeld de regelmaat waarmee nieuwe datasets worden toegevoegd.

De kwaliteit van de datasets moet door de dataproducent kunnen worden gegarandeerd. Ter ondersteuning van de dataproducent zou een datacentrum de mogelijkheid moeten bieden fouten te herkennen via een geautomatiseerde datacontrole.

Afhankelijk van de 'wensen' van de partij die de data aanlevert zijn de data vrij toegankelijk of alleen onder bepaalde voorwaarden.

De beschikbaarstelling van datasets aan onderzoekers moet leiden tot verbeterde communicatie tussen onderzoekers en tot efficiëntere onderzoeksmethoden waarbij onnodige duplicatie van experimenten wordt vermeden.

De leveranciers van datasets moeten voor de geleverde inspanning een incentive ontvangen.

De financiers van wetenschappelijk onderzoek moeten worden overtuigd van de meerwaarde van datahergebruik voor de wetenschap en voor henzelf.

Voor duurzame opslag en hergebruik van datasets is het nodig dat de onderzoekers die de data aanleveren enige kennis hebben van de (metadata) formaten die geschikt zijn voor opslag en hergebruik van datasets.

In veel gevallen moet de aangeleverde data worden geconverteerd naar een formaat dat betrouwbaar is op te slaan en opnieuw te gebruiken. Door het gebruik van aangepaste (gestandaardiseerde) formaten wordt het daarnaast mogelijk om datasets met elkaar te vergelijken en nieuwe verbanden te leggen, waardoor een hogere meerwaarde wordt gecreëerd.

2.2. Samenvattingen deelaspecten

De samenvattingen van het literatuuronderzoek en een verslag van de acquisitie-ervaringen van DANS en het 3TU.Datacentrum worden hieronder weergegeven voor elk van de vier deelaspecten zoals beschreven in paragraaf 1.2.1: Organisatie, Cultuur, Competenties en Systemen.

Organisatie – Literatuur

Om een indeling in de verscheidenheid aan mogelijkheden van *toegang* tot wetenschappelijke datasets te krijgen, maakt [Jacobs, 2004] onderscheid in de volgende groepen:

1. Vrije toegang tot data voor iedereen ongeacht de relatie tot het behorend instituut en gebruiksdoel;
- 2a. Toegang voor personen die hetzelfde vak uitoefenen, gebruik na bemiddeling of tegenprestatie (ruilmarkt);
- 2b. Toegang onmiddellijk of enkele jaren na afsluiting van het project of de publicatie;
- 2c. Toegang tegen betaling (commerciële markt);
3. Geen toegang.

In het geval datasets niet vrij toegankelijk zijn bestaat de mogelijkheid een vorm van gebruikersbeheer toe te passen, waarbij aan diverse gebruikers(groepen) verschillende rechten worden toegekend.

Behalve tegen misbruik, moeten tevens maatregelen worden getroffen tegen verlies van data door storingen en calamiteiten. Daarom wordt de aanbeveling gedaan om de verkregen datasets op meerdere plaatsen op te slaan [Kramer, 2006].

Belangrijke eisen aan dienstverlening met betrekking tot dataopslag en richtlijnen voor het bewaren van digitale onderzoeksdata zijn volgens [Kuula 2008]: advies over praktische kwesties, langdurig datamanagement, veilige en gebruiksvriendelijke opslag van grote hoeveelheden data, controle mechanisme voor verspreiding, en duurzame infrastructuur.

De *verantwoordelijkheid voor dataopslag en -beheer* ligt in de meerderheid van gevallen bij de onderzoeker, in mindere mate bij het afdelingshoofd of de onderzoekseenheid [ASPR, 2008]. Om grootschalige opslag van datasets door centrale instituten te bevorderen is het daarom van belang dat vooral de onderzoekers worden overtuigd van de meerwaarde van hun onderzoek voor de wetenschap en voor henzelf.

Door de voortschrijdende technieken voor aardobservatie (GPS) en dataopslag en verwerking (GIS) komen grote hoeveelheden geografische data, ook bekend onder de term geodata, ter beschikking. Het *genereren van deze data is extreem kostbaar*, de bedrijven en instituten die de hiermee verbonden projecten uitvoeren voelen er daarom weinig voor om deze data kosteloos ter beschikking te stellen via openbare kanalen. Er is iets voor te zeggen om voor deze geodata, veelal gegenereerd zonder financiële ondersteuning van publieke lichamen, tegen betaling ter beschikking te stellen en duplicatie van kostbaar onderzoek te voorkomen [PANGAEA, 2007].

Een ander specifiek kenmerk betreft de *vertrouwelijkheid van gegevens* (bijvoorbeeld medische gegevens die onder het beroepsgeheim van behandelend artsen of zorginstellingen vallen). Het spreekt vanzelf dat de bescherming van deze gegevens moet worden gewaarborgd [Godard, 2003]. Alleen als de gegevens zijn geanonimiseerd kunnen deze door andere personen en instituten worden gebruikt, bijvoorbeeld als referentiemateriaal.

Voor een goede evaluatie van het beschikbaar stellen van datasets is het van belang dat er gebruik(er)sstatistieken worden bijgehouden en dat gebruikers van datasets worden geregistreerd.

Onderscheid moet worden gemaakt tussen *dynamische data* (nog in bewerking) en *statische data* (gepubliceerd en klaar voor archivering). Dynamische data zijn een onderdeel van lopend onderzoek en kunnen gedurende het onderzoek zowel in kwantitatieve als in kwalitatieve zin (gecontroleerd) muteren. Dynamische data dienen gedurende het onderzoek goed beschermd te worden. Zodra het onderzoek is afgerond ontstaat een statische dataset, die bij uitstek geschikt is voor hergebruik in andere omgevingen, met dien verstande dat altijd in overleg met de onderzoeker bepaald zal worden welke datasets inderdaad voor duurzame opslag in aanmerking komen en onder welke condities hergebruik kan plaatsvinden [DARELUX, 2007].

De United Kingdom Data Archive (UKDA) maakt de opgeslagen datasets toegankelijk voor gebruikers die daartoe een contract (End User's Agreement) ondertekenen. Toegang wordt verleend tot non-profit organisaties voor onderwijs en onderzoek. De datasets zijn niet openbaar toegankelijk [UKDA, 2008].

In een onderzoek naar de beschikbaarheid van data en informatie op het gebied van bosbouw [Schweik, 2005] wordt geconstateerd dat de benodigde data wel aanwezig zijn, maar dat deze is opgeslagen in de hoofden, boeken en computers van individuele onderzoekers en onderzoeksafdelingen, en dat het bestaan hiervan niet of nauwelijks bekend is bij collega onderzoekers. Als de aanwezigheid wel bekend is ontstaan vaak problemen door de vorm waarin de data beschikbaar zijn (aangeduid met de term '*file cabinet problem*'), waardoor de data niet zonder meer bruikbaar zijn voor nieuw onderzoek.

Niet alle datasets komen in aanmerking voor bewaren voor de lange termijn. Daarom moet onderscheid worden gemaakt in opslag van datasets voor korte termijn (tot ca. 5 jaar), middellange termijn (5 à 10 jaar) en lange termijn (meer dan 10 jaar).

Dataopslag en -beheer door een repository van een *centraal* instituut (universiteit, land, regio, branche) wordt als positief ervaren door gebruik van standaard formats en metadata [Warden], waardoor transformatie naar een ander, meer geschikt format voor het eigen onderzoek vaak kan worden vermeden. Ook heeft de inrichting van een datacentrum een grotere meerwaarde naarmate de opgenomen vakgebieden *homogener* zijn, omwille van eenvoud.

Organisatie – Acquisitie ervaring

Financiers en opdrachtgevers van onderzoek staan welwillend tegenover het implementeren van beleid om onderzoeksdata toegankelijk te maken. In toenemende mate worden richtlijnen en procedures opgesteld. Dit is een goede eerste stap, maar in de praktijk blijken deze richtlijnen te vrijblijvend om het gewenste beleid uit te voeren.

In projectvoorstellen wordt er door financiers en opdrachtgevers wel op aangedrongen om een paragraaf over datamanagement op te nemen, maar doordat er vaak geen budget of tijd aan gekoppeld is, wordt het door onderzoekers meestal als een extra administratieve last ervaren waarvoor ze ook nog eens geen wetenschappelijke waardering ontvangen.

Wetenschappelijke organisaties zien wel het belang in van het delen van data maar er lijkt een algemene tendens te bestaan dat bestuurders niet vooruit willen lopen op hun eigen onderzoekers. Desondanks worden al wel enige initiatieven ontplooid.

Naast aandacht voor Open Access voor wetenschappelijke publicaties ontstaat een toenemende belangstelling voor hoe organisaties om moeten gaan met wetenschappelijke data. Naast een workflow met betrekking tot publicaties wordt tevens aandacht besteed aan mogelijke workflows voor data.

Veel onderzoek wordt op projectmatige basis gefinancierd. Indien bij dergelijk onderzoek databestanden worden gecreëerd die kostbaar of gecompliceerd beheer vereisen dan is het voor een databewaarplaats vaak niet mogelijk om deze bestanden in het depot op te nemen met behoud van alle functionaliteit. De meeste databewaarplaatsen hebben een vast jaarlijks budget waarmee ze in toenemende mate gecompliceerde datasets duurzaam toegankelijk moeten maken en houden. Indien teveel functionaliteit verloren gaat bij het deponeren van onderzoeksgegevens dan is het voor een dataproducent niet aantrekkelijk om onderzoeksdata voor derden beschikbaar te stellen.

Het ontbreekt op dit moment nog aan een structurele financiering voor het duurzaam toegankelijk maken van alle relevante onderzoeksdata. Onderzoekers die bijvoorbeeld datasets met audiovisueel materiaal of GIS bestanden bij een data-archief willen deponeren voelen soms een belemmering omdat het data-archief deze bestanden niet kosteloos toegankelijk kan maken of omdat een gedeelte van de functionaliteit verloren dreigt te gaan.

De recente, sterke, aanwas van het aantal datasets in DANS EASY is voor een groot deel te verklaren door de raamovereenkomsten die zijn opgesteld met bijvoorbeeld de Rijksdienst voor het Cultureel Erfgoed (voorheen de RACM), NWO en het Ministerie van VWS. De resultaten van deze raamcontracten zijn veelbelovend.

Thematische inventarisatieprojecten, zoals de Nederlandse scheepvaart in het Atlantisch gebied in de zeventiende en achttiende eeuw en Jeugdonderzoek in Nederland, leveren eveneens een aanzienlijke bijdrage aan de data-acquisitie.

Om financiële belemmeringen weg te nemen kunnen onderzoekers een beroep doen op financiering van een Klein Data Project (KDP). DANS kan door middel van KDP's een bijdrage leveren aan het beslechten van deze financiële drempels waardoor onderzoeksdata beter toegankelijk wordt. Door middel van KDP's worden veel relevante onderzoeksbestanden duurzaam toegankelijk gemaakt.

Als laatste kan geconstateerd worden dat individuele onderzoekers in toenemende mate gebruik maken van de mogelijkheid om in DANS EASY, door middel van zogenaamd *self archiving*, hun data te publiceren.

De voornaamste reden voor onderzoekers om data te publiceren lijkt te zijn het verkrijgen van wetenschappelijke waardering.

Cultuur – Literatuur

Diverse auteurs [Warden], [Kramer, 2006], [ASPR, 2008] benadrukken dat de vrije toegang tot (technisch-) wetenschappelijke datasets de communicatie tussen onderzoekers versterkt en de waarde van ieder afzonderlijk onderzoek, waarvan de data in een repository zijn opgeslagen en toegankelijk gemaakt, toeneemt.

Piwowar, Day en Fridsma vinden een correlatie tussen het open access publiceren van data en een verhoogde citatiescore. Verder constateren zij voordelen van open archiving en hergebruik van data in het versterken van de diversiteit in onderzoeksopzet en onderzoeksvragen, bevordering van de communicatie en samenwerken binnen de wetenschappelijke gemeenschap, vermindering van het dubbel verzamelen van data, financieel voordeel dankzij effectiever gebruik van data,

mogelijkheid tot kritische beschouwing van onderzoeksresultaten en herhaald uitvoeren van onderzoek, gelijkwaardiger toegang tot data voor verschillende doelgroepen en verbetering van onderzoekskwaliteit. De voordelen zijn in die zin wetenschappelijk, financieel en maatschappelijk [Piwowar, 2007].

Onderscheid wordt gemaakt tussen verschillende *type collecties* op basis van mate van standaardisering en gebruikersgroepen [NSF, 2006]:

- *Research collections*. Hierbij zijn de auteurs individuele onderzoekers of teams van onderzoekers. De collectie wordt alleen gebruikt door de deelnemers gedurende het project, ze worden beperkt verwerkt en opgeslagen en de data voldoen niet altijd aan standaarden.
- *Resource collections*. Deze datacollecties worden gevuld door een samenwerkende en samenhangende groep auteurs (community), meestal binnen één domein van natuurwetenschappen of techniek; hier worden standaarden toegepast die binnen dat domein worden gehanteerd. Data worden opgeslagen voor middellange of lange termijn.
- *Reference collections*. Deze collecties worden gevuld door grote groepen binnen een domein van natuurwetenschappen of techniek volgens breed geaccepteerde standaarden. Meestal vormt deze collectie als zodanig dé standaard.

De grenzen tussen de verschillende collecties kunnen vervagen en een researchcollectie kan evolueren naar een resource collectie of zelfs reference collectie.

De gevonden karakteristieken die bepalend zijn voor de cultuur of natuurlijke 'neiging' tot het publiceren van data zijn:

1. *Herhaalbaarheid van het onderzoek* (observations of specific phenomena at a specific time or location/scientific experiments/models or simulations [RIN, 2008]).
2. *Looptijd van het onderzoek* (eenmalige experimenten of reeks metingen over een lange periode (longitudinaal)).
3. *Onderzoeks'organisatie'*, bijv. een individu of zeer kleine groep waarbij overdracht geen belangrijke rol speelt of een grote groep (verschillende instellingen) waarbij het delen van gegevens en overdracht wel van belang is.

Cultuur – Acquisitie ervaring

De meeste onderzoekers staan positief tegenover het delen van onderzoeksdata. In de praktijk blijken er echter wel obstakels te zijn om data te deponeren of wil men de data slechts onder bepaalde voorwaarden deponeren. Een van de motieven om data niet beschikbaar te stellen aan anderen is het ontbreken van wetenschappelijke waardering.

Indien al voorwaarden gesteld worden aan een onderzoeker om onderzoeksdata na afloop van een project in een Trusted Digital Repository te archiveren, ontbreekt vaak de financiering. Een onderzoeker ziet zich dan genoodzaakt om aan het eind van een onderzoeksproject nog allerlei administratieve handelingen te verrichten terwijl het publiceren van de resultaten en het financieren van vervolgonderzoeken een hogere prioriteit hebben.

Er wordt ook wel gedacht dat andere onderzoekers weinig belang hebben bij hun data. Aan de andere kant zijn ze bang dat andere onderzoekers eerder met resultaten komen op basis van hun data of dat er fouten in hun eigen onderzoek ontdekt worden.

De ervaring leert dat naarmate de herhaalbaarheid van het dataproductieproces afneemt de belangstelling voor databeheer door een datacentrum toeneemt. Daarnaast lijkt de bereidheid tot het aanleveren van data aan het datacentrum groter bij dataproducenten die zelf data hergebruiken of zich moeten profileren.

Tevens lijkt het niveau van specialisme van invloed, hoe meer specialistisch, hoe minder belangstellenden worden verwacht voor hergebruik en hoe groter de kans dat dataconsument en dataproducent elkaar al gevonden hebben via andere kanalen.

Competenties – Literatuur

Vanuit de onderzoeksprojecten wordt op de volgende punten volgens [Valle] ondersteuning verwacht door een databeheerder:

- Geen verlies van data of kennis;
- Toegankelijk maken (zoeken en verkrijgen moet mogelijk zijn);
- Ontdekken door middel van ‘bladeren’;
- Ondersteuning bij het voorbereiden van datapublicatie;
- Terugbrengen van ‘data-entropie’;
- Ondersteunen van ‘alerting/notification’ bij nieuwe data en disseminatie resultaten.

Onderscheid wordt gemaakt in verschillende *type gebruikers* [DARELUX, 2007].

In DARELUX zijn drie gebruikersgroepen te onderscheiden:

- *Primaire gebruikers*: onderzoekers die direct betrokken waren bij het project en het verwerken van de data.
- *Secundaire gebruikers*: wetenschappers die niet direct betrokken zijn bij het project maar werkzaam zijn in de hydrologie of in een aangrenzend onderzoeksgebied en die zowel gebruiker als leverancier van data zouden kunnen worden.
- *Tertiaire gebruikers*: onderzoekers die zich uitsluitend beperken tot het gebruik van de gearchiveerde data.

Competenties – Acquisitie ervaring

Onderzoekers spelen een belangrijke rol in het duurzaam toegankelijk maken en houden van data. Om deze rol adequaat te kunnen vervullen moet in het curriculum van onderzoekers en studenten aandacht zijn voor zaken als datamanagement. Geleidelijk wordt binnen wetenschappelijke organisaties aandacht besteed aan dit soort kennisoverdracht aan wetenschappelijk medewerkers.

Daarnaast zijn veel te archiveren databestanden niet op orde. Het ontbreekt aan degelijke documentatie en vragen naar aanvullende informatie benodigd voor hergebruik kunnen vaak niet eenvoudig beantwoord worden. Onderzoekers bezitten veelal niet de nodige kennis om bestanden volgens gangbare standaarden te documenteren en vaak hebben ze hiervoor geen tijd.

Systemen – Literatuur

Aangeleverde en opgeslagen data zijn vaak niet zonder meer geschikt voor hergebruik. Daarom wordt aangedrongen op conversie van opgeslagen data in een format dat geschikt is voor de gebruiker [Warden]. Een “Checklist/handleiding op het gebied van databeheer en/of metadata ten behoeve van datasets met aanbevelingen en richtlijnen” moet bijdragen aan het gemak waarmee gebruikers over datasets kunnen beschikken [RIN, 2008].

Een interface die data-integratie ondersteunt en automatisch ‘anomalies’ detecteert draagt in hoge mate bij aan de bruikbaarheid en betrouwbaarheid van datasets [Helly], [CESSDA, 2008].

Pfaltz benadrukt dat wetenschappelijke datasets wezenlijk verschillen van, bijvoorbeeld, bedrijfsadministratie. Bij laatstgenoemde is meestal sprake van een eenduidige (of eendimensionale) relatie tussen twee zaken, die eenvoudig in een tabel of grafiek kunnen worden weergegeven. In wetenschappelijk onderzoek is vaak sprake van een *veelheid aan onafhankelijke variabelen waardoor hogere eisen worden gesteld aan de opslag van de gegevens* [Pfaltz]. Door het *verbinden van datasets* van verschillende disciplines, zo wordt door het Nederlandse instituut DANS aangevoerd, kunnen bovendien nieuwe datasets worden gegenereerd met nieuwe inzichten, correlaties en similarities [Kramer, 2006].

Systemen – Acquisitie ervaring

Een van de belangrijkste redenen voor een onderzoeker om data met anderen te delen is het wel of niet verkrijgen van wetenschappelijke waardering. De meeste wetenschappelijke organisaties zien wel in dat wetenschappelijke output zich niet alleen beperkt tot publicaties in "belangrijke" tijdschriften maar dat ook software tools en data daar onder zouden kunnen vallen. Het aanpassen van dit waarderingssysteem is echter een gecompliceerd en tijdrovend proces.

De data–infrastructuur is (inter)nationaal bijzonder gefragmenteerd. Nog niet voor alle wetenschappelijke disciplines bestaan data–archieven. Er is geen eenduidige wijze waarop gerefereerd kan of moet worden naar onderzoeksdata. Er is een gebrek aan beleid en communicatie over Persistent Identifiers. Door het ontbreken van wetenschappelijke waardering met betrekking tot het beschikbaarstellen van onderzoeksdata is er nog weinig belangstelling van onderzoekers voor het publiceren in Journals met een zogenaamde Data Availability Policy. Daarbij vergt het voor uitgevers hoge investeringen in infrastructuur om onderzoeksdata aan publicaties te koppelen. Het probleem van de lange termijn duurzaamheid van verrijkte publicaties is daar dan veelal nog niet in meegenomen.

Aan de ene kant zijn (inter)nationaal samenwerkende onderzoeksgroepen vaak meer geïnteresseerd in de mogelijkheden van een datacentrum, in het bijzonder als een vorm van 'restricted access' mogelijk wordt gemaakt. Aan de andere kant ontstaan bij groepen waar met (zeer) grote bestanden wordt gewerkt juist bedenkingen die te maken hebben met bandbreedte of het aanbieden van rekencapaciteit door het datacentrum.

Opmerkingen

De hierboven beschreven onderdelen vormen slechts een deel van de bestudeerde literatuur. In de bijlage van deze rapportage vindt u een volledig overzicht.

3. Veldonderzoek

Om de functionele eisen in kaart te brengen voor succesvolle digitale data-archivering is een onderzoek gestart door middel van een tweetal (beknopte) casestudies. De literatuurscan uit hoofdstuk 2 vormt het uitgangspunt voor deze casestudies.

De casestudies verschillen van elkaar op een aantal punten, zie hieronder een korte omschrijving:

a) DARELUX. Dit is een casestudie naar dataopslag door een sectie die al een paar jaar gebruik maakt van centrale data-archivering en die meetgegevens verzamelt van observaties van natuurlijke verschijnselen, zoals hoeveelheid neerslag, hoeveelheid waterafvoer door beken en rivieren, grondwaterniveau, etc. en derhalve niet herhaalbaar. De gegevens worden op verschillende locaties verzameld en zijn vooral interessant als ze beschikbaar zijn over een langere periode.

b) Waterlab. Dit is een casestudie bij een sectie die nog geen gebruik maakt van centrale data-archivering maar al wel ervaring heeft met de problemen van niet duurzaam databeheer zoals verouderde media en gebrekkige metadata. Deze onderzoeksgroep heeft bovendien enige ervaring met het delen van onderzoeksgegevens. De meetdata worden verzameld door middel van experimenten in een gecontroleerde omgeving en zijn in principe herhaalbaar.

De volgende paragrafen beschrijven de resultaten van de interviews met de verschillende onderzoeksgroepen. Tijdens de interviews werd de nadruk altijd gelegd op de aspecten die van belang zijn voor de dataproducenten en binnen de invloedssfeer liggen van een datacentrum. Alle aanbevelingen die tijdens de interviews naar voren zijn gekomen zijn echter gebruikt voor het opstellen van de functionele wensen, daarom vindt u in Hoofdstuk 4 tevens aanbevelingen die betrekking hebben op bijvoorbeeld het hergebruik en aanbevelingen die gericht zijn op de onderzoeksfinanciers.

De aanbevelingen zijn in de vorm van stellingen getoetst tijdens een bijeenkomst met een brede groep experts. Paragraaf 3.3. geeft een overzicht van de bevindingen tijdens deze bijeenkomst.

3.1. Casestudies

Voor de eerste casestudie is het project Data Archiving River Environment Luxemburg (DARELUX) geselecteerd. Binnen de betrokken onderzoeksgroepen is ervaring opgedaan met het duurzaam digitaal opslaan van technisch-wetenschappelijke onderzoeksgegevens. De interviews gaven inzicht in de werkwijze van DARELUX bij dataverzameling en de behoeften voor het online bijhouden en archiveren van data. Binnen deze casestudie wordt de nadruk gelegd op de leerervaringen in het gebruik van centrale data-archivering. De tweede casestudie vormt een analyse van de werkprocessen van de sectie Waterbouwkunde van Civiele Techniek en Geowetenschappen (CiTG). Deze sectie maakt geen gebruik van centrale data-archivering. Door de ervaring van een goed practise case te combineren met een case waarin geen ervaring is, ontstaat inzicht in de toegevoegde waarde van het datacentrum voor dataproducenten.

Hieronder worden eerst de achtergronden van de onderzoeksgroepen beschreven, vervolgens zal dieper worden ingegaan op de bevindingen.

3.1.1. DARELUX: centrale data-archivering

In oktober 2004 is onder de naam DARELUX² een project gestart voor het duurzaam digitaal opslaan van hydrologische meetgegevens. Het project bouwde voort op de resultaten van het in 2002 succesvol afgeronde e-Archive project en liep tot 2006.³ Hoewel het DARELUX-project officieel in 2006 is afgerond loopt het onderzoek nog steeds en wordt het destijds ontwikkelde data-archief nog steeds gebruikt voor opslag en toegankelijk maken van nieuwe onderzoeksgegevens.

In het DARELUX project werkten faculteiten Civiele Techniek en Geowetenschappen (CiTG, Leerstoel Hydrologie), en Elektrotechniek, Wiskunde en Informatica (EWI, Informatiesystemen en algoritmiek) van de TU Delft samen met de faculteit Geowetenschappen (Fysische Geografie) van de Universiteit Utrecht en de TU Delft Library.

Binnen de Geowetenschappen hebben onderzoeksgegevens een lange levenscyclus, zeker wanneer het gegevens ten behoeve van klimaatonderzoek betreffen. Het DARELUX project is opgestart om te waarborgen dat de technisch-wetenschappelijke onderzoeksgegevens ook voor de zeer lange termijn (meer dan 50 jaar) beschikbaar zijn. DARELUX bevat zowel ruwe meetdata als verwerkte data.

Een van de deelnemende partijen aan het DARELUX project – een partij die nog altijd gebruik maakt van de data – is de onderzoeksgroep Hydrologie van de faculteit CiTG. Deze onderzoeksgroep verzamelt meetgegevens in onder andere het stroomgebied van de Maisbich in Luxemburg.

3.1.2. Sectie Waterbouwkunde

Om de functionele eisen en wensen in kaart te brengen voert het projectteam naast een casestudie naar een good practise case tevens onderzoek uit bij een onderzoeksgroep die geen gebruik maakt van centrale data-archivering. Hiervoor is de sectie Waterbouwkunde van CiTG van de TU Delft benaderd.

Een belangrijk deel van het onderzoek uitgevoerd door de sectie Waterbouwkunde betreft golfstromen. In tegenstelling tot DARELUX doet deze sectie geen lange termijn onderzoek maar onderzoek op basis van experimenten. Deze experimenten kunnen zowel meetgegevens als beeldmateriaal opleveren. De data worden tijdens het uitvoeren van het onderzoek opgeslagen op een server. De analyse van de data gebeurt echter lokaal op het werkstation van de onderzoeker. De archivering van de onderzoeksdata gebeurt op initiatief van de onderzoeker. In het verleden werd dit voornamelijk gedaan op tapes, cd's en dvd's. Door de toename van de omvang van de meetdata worden momenteel vaak externe harde schijven gebruikt.

Door de werkprocessen van deze sectie nauwkeurig met de onderzoekers te doorlopen worden de verbeterpunten voor de archivering en het hergebruik van data blootgelegd. De bevindingen van beide casestudies beschrijven we in de volgende paragraaf.

² Website DARELUX: www.library.tudelft.nl/darelux/

³ <http://www.library.tudelft.nl/darelux/projectinformatie/index.htm>

3.2. Bevindingen uit de casestudies

De casestudies leiden tot een aantal directe aanbevelingen, zowel naar aanleiding van praktijkervaring als gebaseerd op wensen met betrekking tot het archiveren van data. De resultaten uit de casestudie zijn onderverdeeld in de onderdelen organisatie, cultuur, competenties en systemen.

3.2.1. Organisatie

Strategie

Het datacentrum moet een duidelijke strategie formuleren en communiceren. Het datacentrum heeft de doelstelling om datasets toegankelijk te houden op lange termijn en hergebruik van datasets te stimuleren. In de basis streeft het datacentrum naar open access. De periode die data beschikbaar zouden moeten blijven dient vooraf gedefinieerd te worden, vooral voor onderzoek dat onderdeel uitmaakt van een historische analyse. De periode waarvoor toegankelijkheid gegarandeerd kan worden is mede afhankelijk van de businesscase.

Het datacentrum moet een businesscase ontwikkelen om kosten en baten in kaart te brengen. Op basis van de businesscase kan het datacentrum beslissen of uitbreiding van de dienstverlening (waaronder opslag van commerciële datasets, rekencapaciteit, duur van de opslag) wenselijk is. Vanuit de wens om open access te realiseren moet toegang tot wetenschappelijke datasets gratis zijn. Dit verlaagt de drempel om data opnieuw te gebruiken voor onderzoek en daarmee worden de investeringen voor het genereren van de data beter benut. Bovendien zou de opslag van onderzoeksdata voor de dataproducenten geen (extra) kosten met zich mee moeten brengen om de drempel voor het aanleveren en delen van data laag te houden. Om financiering voor het datacentrum te garanderen zal de (duurzame) opslag van data onderdeel kunnen worden van onderzoeksopzet en -budget. In het onderzoeksvoorstel dient de onderzoeker zich de vraag te stellen of opslag van onderzoeksdata in een datacentrum zinvol en gewenst is. Is dit het geval, dan moet de onderzoeker de duurzame opslag van meetdata in zijn onderzoeksplanning opnemen.



Aanbeveling

- Het datacentrum moet data zoveel mogelijk open access aanbieden om hergebruik te bevorderen.
- Kosten voor opslag moeten geen belemmering zijn voor het aanleveren van wetenschappelijk data.
- Er moet een businessmodel opgesteld worden voor duurzame financiering van het datacentrum.
- Bij de start van een onderzoek moet al een inschatting worden gemaakt van het belang van lange termijn toegankelijkheid van de data.
- Voor relevante datasets moet in de onderzoeksplanning tijd gereserveerd worden voor opslag van data in een datacentrum.

Doelgroepen

Het veiligstellen van de verzamelde onderzoeksgegevens is van groot algemeen belang voor de wetenschap. Archiveren in een betrouwbaar datacentrum biedt hiertoe de mogelijkheden. De voornaamste leveranciers van het datacentrum zullen promovendi en post docs zijn. Zij gebruiken het datacentrum voor de opslag van door hen verzamelde gegevens in het kader van het eigen onderzoek. In toenemende mate kunnen daarnaast docenten en afstudeerders gebruik maken van de opgeslagen onderzoeksdata voor het onderwijs en afstuderen.

Een andere doelgroep die het datacentrum kan benaderen zijn de commerciële onderzoeksinstellingen. Voor hen dienen in principe dezelfde regels te gelden als voor de wetenschappelijke onderzoekers. Voor een 'klant' kunnen voor extra diensten aparte afspraken worden ontwikkeld. Dit soort afspraken is uiteraard niet voorbehouden aan commerciële klanten van het datacentrum maar is mogelijk voor elke partij.



Aanbeveling

- Het datacentrum moet de focus leggen op de dataproducenten (met name post docs en promovendi) voor het verzamelen van data. Voor hergebruik bestaat de doelgroep uit zowel de hierboven genoemde dataproducenten als betrokkenen bij het wetenschappelijk onderwijs zoals docenten en studenten.
- Het datacentrum kan extra diensten tegen betaling aanbieden maar biedt bij voorkeur zonder kosten toegang tot de datasets.

Schaal

Een zo breed mogelijk opgezette digitale database, internationaal of nationaal, biedt voordelen boven een beperktere database. Het streven moet zijn een internationale verzameling onderzoeksgegevens. Hiervoor dient het datacentrum afspraken te maken met andere partijen. Standaardisatie vormt hierbij zowel een mogelijkheid als een moeilijkheid. Standaarden zijn momenteel nog niet uitontwikkeld en er zijn geen universele standaarden op het gebied van formats en metadata die aan alle eisen voldoen. Ook het bewaken van de kwaliteit vormt een aandachtspunt indien gekozen wordt voor een internationale open omgeving, aangezien het niveau van onderzoekers en faciliteiten over de wereld verschilt.



Aanbeveling

- Het datacentrum moet streven naar een internationale scope, waarbij aandacht is voor het adopteren van meerdere standaarden en faciliteiten voor kwaliteitscontrole van de datasets.

Kwaliteit

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

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

Een bepaald minimum kwaliteitsniveau van de onderzoeksdata lijkt een voorwaarde voor hergebruik en hoe hoger de kwaliteit hoe vaker de gegevens zullen worden gebruikt ten opzichte van ‘concurrerende’ (vergelijkbare) datasets.



Aanbeveling

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

Hergebruik

Onderzoeksdata van anderen worden in veel vakgebieden nog op beperkte schaal hergebruikt. Belangrijke oorzaak hiervan is waarschijnlijk de ontoegankelijkheid van onderzoeksdata. Bovendien is in veel gevallen het onderzoek niet te reproduceren door o.a. het ontbreken van gegevens, hierdoor worden de onderzoeksdata als onbetrouwbaar gezien. Voor het datacentrum is het daarom van belang de opgeslagen data inzichtelijk en vindbaar te maken. Daarnaast moeten de data voorzien zijn van metadata waarin het onderzoek en de onderzoekopstelling staan beschreven.

Onderzoekers doen vaak veel aan hergebruik van hun eigen onderzoeksgegevens. Tijdens de uitvoering van het onderzoek vinden verschillende iteratieslagen plaats waarbij de data geanalyseerd worden. In veel gevallen worden de verschillende versies in een lokale omgeving opgeslagen wat kwetsbaarheid en in sommige gevallen problemen met opslagcapaciteit met zich meebrengt. Het datacentrum kan in een behoefte voorzien door tijdens het onderzoek verschillende versies, eventueel afgeschermd voor anderen, centraal op te slaan en op een andere locatie back-ups te bewaren.

De toegang tot grote datasets, in het bijzonder als deze nog onder bewerking zijn, vraagt om aandacht wat betreft bandbreedte voor het verplaatsen van de data dan wel voor het bieden van reken capaciteit bij de opslaglocatie.

Meer aandacht moet worden gevestigd op de nieuwe mogelijkheden die geboden worden doordat onderzoeksdata van hoge kwaliteit geschikt zijn gemaakt voor hergebruik en open access beschikbaar worden gesteld. Het gaat dan bijvoorbeeld om het gebruik van digitale data voor ‘verrijkte publicaties’⁴ of voor onderwijs. Een andere voorwaarde voor hergebruik, en in sommige gevallen een incentive, is de ‘zichtbaarheid’ van het datacentrum en de daarin opgeslagen datasets.

Andere veelgenoemde voorwaarden zijn het ‘regelen’ van het eigendom en de gebruiksvoorwaarden van de datasets. Voor het vertrouwen in datasets is bovendien een bronvermelding een minimum vereiste. Daarnaast moet het mogelijk zijn gegevens af te schermen totdat het onderzoek is afgerond. Onderzoekers willen daarna inzicht in wie de door hen verzamelde data inziet en gebruikt.

⁴ ‘Verrijkte publicatie’ of *enhanced publication* in het Engels staat voor een publicatie die behalve uit tekst en afbeeldingen ook bestaat uit andere vormen, bijvoorbeeld de onderliggende onderzoeksgegevens of zelfs ruwe data waarop het artikel gebaseerd is. Door het toegankelijk maken van dergelijke gegevens kunnen de bewerkingen en conclusies geverifieerd worden.

**Aanbeveling**

- De metadata van datasets moeten doorzoekbaar zijn (zie ook 'Systemen').
- Binnen grote datasets moeten zoek- en selectiemogelijkheden zijn (zie ook 'Systemen').
- In de metadata die bij de dataset wordt opgeslagen moet het onderzoek goed beschreven zijn.
- Het datacentrum moet de opslag van data al tijdens het onderzoek mogelijk maken (incl. versiebeheer), tijdelijke afscherming van onderzoekgegevens moet daarom mogelijk zijn (zie ook 'Cultuur').
- Het datacentrum moet het uitbrengen van verrijkte publicaties door wetenschappers faciliteren.
- Voor datasets in het datacentrum moeten de eigendomsrechten en gebruiksvoorwaarden duidelijk beschreven worden (zie ook 'Cultuur').
- Alle data moet een authentieke bron hebben die herleidbaar is.

Rolverdeling

Voor het succesvol realiseren van het datacentrum moet een aantal rollen ingevuld zijn. Het systeem moet onderhouden en aangepast worden door een technisch beheerder. De technisch beheerder speelt voornamelijk op de achtergrond een belangrijke rol om de continuïteit van het datacentrum te waarborgen. Hij wordt inhoudelijk gevoed door functioneel beheerders binnen de onderzoeksinstellingen. De functioneel beheerder ziet toe op de juiste uitvoer van het opslagproces en vormt het eerste aanspreekpunt binnen de gebruikende organisaties. Eventuele verbeterpunten voor het systeem worden verzameld door de functioneel beheerder en doorgegeven aan de technisch beheerder.

Naast een goede operationele rolverdeling is de bekendheid van het datacentrum van groot belang voor succesvol (en breed) gebruik. Om die bekendheid te genereren zal het datacentrum binnen de onderzoeksinstellingen gepromoot moeten worden.

Het verdient aanbeveling om naast een rolverdeling binnen de organisatie van het datacentrum, een rolverdeling te stimuleren onder de gebruikersgroep.

In de casestudie van DARELUX hebben de onderzoekers de rol van dataverzamelaar. Zij hebben speciaal een aparte, centrale functie ingericht voor het kwaliteitsbeheer en het invoeren van de data in het datacentrum. Voor het borgen van de kwaliteit lijkt, zeker bij grotere onderzoeksgroepen, aandacht voor het opzetten van de organisatie voor data aanlevering van belang.

**Aanbeveling**

- Zowel technisch als functioneel beheer rondom het datacentrum moet geregeld zijn.
- Het datacentrum moet de voordelen van gecentraliseerde opslag en ontsluiting duidelijk kenbaar maken aan de doelgroepen.
- Het datacentrum kan de dataproductanten begeleiden bij het inrichten van een gebruikersorganisatie en definiëren van workflows, bijvoorbeeld door middel van aanbevelingen.

3.2.2. Cultuur

De cultuur van de onderzoekers is bepalend voor het gebruik van het datacentrum. Indien de onderzoekers niet openstaan voor het gebruik van het datacentrum, is de slagingskans beperkt.

Delen van data

Het delen van onderzoeksdata is nog geen standaard binnen de wetenschappelijke wereld. Deels heeft dit te maken met de beschikbaarheid van goede voorzieningen om deze onderzoeksdata te delen waardoor de drempel als te hoog wordt ervaren. Daarnaast bestaat er bij sommige wetenschappers terughoudendheid voor het delen van data. Deze terughoudendheid blijkt soms ingegeven door voorwaarden van de financier, soms door ‘concurrentie’ overwegingen tussen onderzoeksgroepen en soms door persoonlijke voorkeuren of historisch gebruikelijke praktijken. De maatschappelijke ontwikkeling laat echter steeds meer openbaar gedeelde informatie en kennis zien. Indien de juiste middelen beschikbaar zijn is de verwachting dat ook in de wetenschap de onderzoeksdata gedeeld zullen worden. Dit dient op vrijwillige basis te gebeuren. Het dwingen van wetenschappers lijkt geen goede methode. Dit kan weerstand oproepen en een negatieve uitwerking hebben op de kwaliteit van de opslag van de onderzoeksdata. Het belang van duurzame archivering en de meerwaarde die de diensten van een datacentrum bieden moeten voldoende reden zijn om data in het datacentrum op te slaan. Stimulering vanuit de onderzoeksfinanciers en de leidinggevenden draagt wel bij aan het gebruik en de snelheid waarmee een cultuurverandering kan plaatsvinden.



Aanbeveling

- Stimuleer de open houding van onderzoekers ten opzichte van het delen van onderzoeksdata.
- Deponeren van data bij het datacentrum moet geen verplichting zijn, de motivatie van dataproducenten moet voornamelijk worden ingegeven door het bieden van meerwaarde door het datacentrum.
- Stimuleer beheer van data door datacentra middels beloning van dataproducenten door financiers en leidinggevenden.

Vrije toegang, tijdelijke afscherming

De houding ten aanzien van vrije toegang tot de data (Open Access) varieert. Onder sommige onderzoekers bestaat de vrees dat derden gegevens zullen gebruiken voor eigen onderzoek nog voordat de ‘auteur’ ze zelf gepubliceerd heeft. Vooral voor sterk concurrerende onderzoeksgroepen zal tijdelijke afscherming van gegevens daarom mogelijk gemaakt moeten worden totdat de onderzoeker de data mag en/of wil vrijgeven, bijvoorbeeld na afronding van het onderzoek. Bovendien moeten de eigendomsrechten en gebruiksvoorwaarden, van data die zich in het datacentrum bevinden, bepaald en gecommuniceerd worden. Daarnaast kan angst bestaan voor controle van hun werkwijze door anderen. De verwachting is dat beschermingsdrang voor het eigen onderzoek zal toenemen naarmate op grotere schaal digitale gegevens van anderen gebruikt gaan worden.



Aanbeveling

- Voor datasets in het datacentrum moeten de eigendomsrechten en gebruiksvoorwaarden duidelijk beschreven worden (zie ook ‘Organisatie’).
- Het datacentrum moet tijdelijke afscherming van onderzoeksdata toestaan (zie ook ‘Organisatie’).

Open voor innovatie

Nieuwe vormen van data(her)gebruik moeten binnen de onderzoekswereld actiever gepromoot worden. Op deze manier kan een open mindset voor nieuwe werkwijzen ontstaan. Voorbeelden zijn het samenwerken aan datasets, het gebruiken van datasets van anderen voor eigen onderzoek, het verrijken van publicaties met data, etc.. Datamanagement moet daarom onderdeel gaan uitmaken van het onderwijs zodat het in een vroeg stadium gewoonte wordt onderzoeksdata

zorgvuldig te beschrijven en te bewaren. Daarom zal beleid ontwikkeld moeten worden gericht op het stimuleren van (her)gebruik in het onderwijs en vooral door de wetenschappers.



Aanbeveling

- Stimuleer het hergebruik van datasets door datamanagement op te nemen in het (wetenschappelijk) onderwijs.

Incentives

Het toekennen van incentives aan het opslaan van data in het datacentrum kan onderzoekers over de drempel trekken om gebruik te maken van het datacentrum. De meest relevante incentives voor een datacentrum zijn:

- Het makkelijker maken van de opslag van gegevens in het datacentrum (zie tevens hierboven onder "Delen van data"). Iedere onderzoeker moet zijn gegevens hoe dan ook opslaan. Wanneer dit gemakkelijk en duurzaam kan – met kwaliteitscontrole – dan zullen veel onderzoekers de diensten van het datacentrum willen gebruiken in plaats van lokale opslagmogelijkheden op te zetten.
- Het citeerbaar maken van onderzoeksgegevens.

Naast te beïnvloeden incentives zijn er andere interessante stimuleringsmogelijkheden voor het gebruik van het datacentrum. Deze vallen echter buiten de directe invloedssfeer van het datacentrum.

- Beloning van onderzoeker bij opslag data direct of bijv. door middel van een puntensysteem zoals bij publicaties.
- Opname van gepubliceerde datasets in de CV's van onderzoekers, eventueel met een waardering.
- Het toevoegen van feedback op opgeslagen datasets (bv. in reviews).



Aanbeveling

- Het datacentrum moet 'gebruiks'incentives bieden aan onderzoekers, onder andere efficiënt databeheer en citeerbaar maken van datasets.
- Stimuleer wetenschappelijke waardering voor het delen van datasets.

Tijdsinvestering

Een belangrijke drempel voor onderzoekers is de benodigde tijd die het kost om de data voor te bereiden voor opslag in het datacentrum. Voor de acceptatie van het datacentrum zal daarom onderzocht moeten worden of de efficiëntie van aanlevering en beheer vergroot kunnen worden.

De gewenste situatie is dat onderzoekers zelf hun data opslaan en bewerken. Een alternatief is dat het servicecentrum van het datacentrum zorgt voor de invoer, ordening en kwaliteitscontroles. Algemene kwaliteitscontrole op de datasets blijft nodig, ook wanneer de onderzoekers de invoer doen. Bekeken moet worden of enkele taken door studenten kunnen worden uitgevoerd. Scheiding van de rollen functioneel en technisch beheer lijkt nodig wanneer op grote schaal van het datacentrum gebruik gemaakt gaat worden.

Door het waarborgen van de kwaliteit zal de bereidheid tot het investeren door dataproducenten naar verwachting toenemen. Kwaliteit wordt over het algemeen dermate belangrijk gevonden dat de gevraagde inspanning minder als een probleem wordt ervaren.

**Aanbeveling**

- Het opslaan van data in het datacentrum moet eenvoudig en efficiënt kunnen gebeuren.
- De onderzoeker moet data zelf (kunnen) invoeren in het datacentrum.
- Het datacentrum moet de kwaliteit van opgenomen datasets bewaken.

3.2.3. Competenties

Het werken met en verwerken van onderzoeksdata binnen het datacentrum vereist bepaalde competenties en vaardigheden. In deze paragraaf wordt beschreven welke competenties volgens de casestudies nodig zijn.

Onderzoekplan

Bij de start van het onderzoek schrijft de onderzoeker een onderzoeksvoorstel. Dit voorstel geeft een beschrijving van de wijze waarop het onderzoek uitgevoerd wordt. In het onderzoeksvoorstel dient de onderzoeker al plannen te beschrijven voor datamanagement. Momenteel wordt dit in veel gevallen niet meegenomen in het onderzoeksvoorstel.

**Aanbeveling**

- Dataproductenten moeten plannen kunnen opstellen voor datamanagement.

Opslag data

De uitgangssituatie voor het datacentrum is dat de onderzoekers hun onderzoeksgegevens zelf opslaan in het datacentrum. De opslag moet daarom zo eenvoudig mogelijk gemaakt worden. Dit is mogelijk door automatisering van een deel van het werk. Van de onderzoekers vraagt het dan nog een eenmalig leerproces voor het opslaan van data. Daarnaast zal een richtlijn opgesteld moeten worden voor het invullen van metadata waardoor hergebruik mogelijk is.

Omdat het gebruik van een datacentrum enige kennis vereist, moeten gebruikers de mogelijkheid krijgen een korte (online) training te volgen op het gebied van datamanagement (lieftst als onderdeel van de opleiding). Bovendien moet een goede handleiding aangeboden worden middels een helpfunctie in het systeem maar tevens als hardcopy. Het bevordert vooral de consistentie van de opgeslagen metadata en datasets.

**Aanbeveling**

- Het datacentrum moet een korte (online) training aanbieden om het correct opslaan van de onderzoeksgegevens te bevorderen.
- Er moeten richtlijnen beschikbaar zijn voor het invullen van metadata.
- Er moet een goede handleiding beschikbaar zijn voor het gebruik van het datacentrum.

3.2.4. Systemen

De onderzoekers worden door verschillende personen en tools ondersteund bij het opslaan van de data. Door een goede afstemming tussen deze systemen kan het gebruik gestimuleerd en gemakkelijk gemaakt worden.

User interface

Een ander cruciaal punt is de kwaliteit en gebruiksvriendelijkheid van de user interface. De interface moet enerzijds dataproducenten faciliteren zodat deze zelf de gegevens snel en eenvoudig kunnen opslaan. Anderzijds is voor dataconsumenten een visuele interface gewenst, die het zoeken naar data vergemakkelijkt. Wanneer de interface goed aansluit bij de wensen van de gebruikers zullen de onderzoekers eerder bereid zijn het datacentrum te gebruiken als primaire opslagplaats voor hun gegevens.



Aanbeveling

- De user interfaces voor zowel aanleveren als hergebruik moeten eenvoudig en intuïtief functioneren.
- De user interface voor hergebruik moet visualisatie van (meta)data mogelijk maken ten behoeve van selectie (bijv. geografische posities op een kaart, grafieken/plots, picture galleries, etc.)

Metadata

Onderzoekers willen meer inzicht in de wijze waarop de invoer van gegevens plaatsvindt. Hiervoor is een 'flexibele standaard' voor metadata nodig. Er moet een standaard template komen voor het beschrijven van de onderzoeksdata. In de template moet ten behoeve van hergebruik een minimale beschrijving plaatsvinden. De onderzoeker moet echter binnen de template de ruimte krijgen om een verdere beschrijving van het onderzoek te doen.



Aanbeveling

- Er moet een (flexibele) standaard komen voor (beschrijvende) metadata bestaande uit een kleine minimum set met voldoende optionele velden voor een beschrijving die hergebruik van de data mogelijk maakt.

Vindbaarheid

Naast een goede beschrijving van het onderzoek in de metadata is de vindbaarheid van de datasets belangrijk. Hiervoor dient het datacentrum over een goede zoekfunctie te beschikken. De zoekfunctie moet de mogelijkheid bieden om te zoeken op combinaties van zowel metadata inhoud (zoals locatie (visueel), dataproducent of woorden in de beschrijving) als op gegevens zelf (zoals tijden en andere waardes). Het is aan te raden om de metadata van de datasets bovendien indexeerbaar te maken voor Google. Dit vergroot de vindbaarheid van de onderzoeksdata en draagt daarmee bij aan het hergebruik van de gegevens.



Aanbeveling

- De metadata van datasets moeten doorzoekbaar zijn (zie ook 'Organisatie').
- Binnen grote datasets moeten zoek en selectiemogelijkheden zijn (zie ook 'Organisatie').
- Metadata van datasets moeten indexeerbaar zijn door externe zoekmachines als Google.

Functionaliteiten

Het datacentrum heeft in beginsel de functionaliteit van het duurzaam opslaan van onderzoeksgegevens. Door gebruik van het datacentrum is hergebruik van de gegevens mogelijk. Om inzicht te hebben in de opgeslagen data en het gebruik hiervan kan het datacentrum enkele functionaliteiten toevoegen. Geïdentificeerde functionaliteiten zijn o.a. een checklist ten behoeve van kwaliteitscontrole, gebruikersstatistieken en de mogelijkheid om commentaar te geven op datasets (recensies).

**Aanbeveling**

- Het datacentrum moet o.a. de volgende functionaliteiten bieden: checklist voor kwaliteitscontrole, gebruikersstatistieken en een mogelijkheid om feedback te geven op datasets (zie ook 'Cultuur').

Opslag

Het datacentrum moet voorzien in de opslag van verschillende soorten onderzoeksdata. Deze onderzoeksdata kunnen uiteenlopen van tekst en meetdata tot beeld- en video-materiaal. Het is wenselijk om de ruwe onderzoeksgegevens op te kunnen slaan omdat deze gegevens nu door de onderzoeker vaak op een mobiele datadrager worden opgeslagen (zoals tape, dvd, etc.) en deze kwetsbaar zijn voor verlies en verval. Naast voldoende opslagcapaciteit moet het datacentrum kunnen garanderen dat de opslag veilig beheerd wordt. Het is daarom wenselijk om o.a. te zorgen voor redundantie: er moet een kopie van de opgeslagen data beschikbaar zijn en teruggezet kunnen worden.

**Aanbeveling**

- Het datacentrum moet de opslag van verschillende soorten onderzoeksdata faciliteren.
- Het datacentrum moet minimaal 'betrouwbare' opslag (dat wil zeggen veiligstellen en toegankelijk houden van data) kunnen garanderen.

Standaarden

Standaarden helpen om gegevens uitwisselbaar te maken, ook op lange termijn. Er zijn verschillende standaarden in omloop die nog in ontwikkeling zijn. Het datacentrum moet een set standaarden kiezen voor de opslag die zorgt voor duurzame opslag. Daarnaast moeten de gekozen standaarden data-uitwisseling mogelijk maken. Op deze manier kunnen koppelingen worden gelegd tussen datasets, waardoor grootschaliger en multidisciplinair onderzoek eenvoudiger wordt. De datasets moeten in verschillende formaten gedownload kunnen worden. De verhoogde compatibiliteit zal het hergebruik vereenvoudigen.

**Aanbeveling**

- De opslag van data in het datacentrum moet gebeuren in een uitwisselbare standaard.
- Downloaden van datasets moet mogelijk zijn in verschillende gangbare formats.

Techniek

Het datacentrum moet toegankelijk zijn via het internet. De data moeten bovendien worden beschermd tegen het maken van wijzigingen door onbevoegden. Naast de ruwe data moeten tevens de bewerkte data beschikbaar zijn. Deze data worden over het algemeen gebruikt in de rapportages, maar tevens voor hergebruik (en controle).

**Aanbeveling**

- Het datacentrum moet toegankelijk zijn via internet.
- In het datacentrum moeten zowel ruwe als bewerkte onderzoeksdata opgeslagen kunnen worden.

3.3. Expertbijeenkomst; toetsing van de bevindingen

Om de bevindingen uit de casestudies te toetsen organiseerde de projectgroep een bijeenkomst met onderzoekers van de drie technische universiteiten (TU Delft, TU Eindhoven en Universiteit Twente). Naast toetsing van de bevindingen stond het ochtenddeel van de bijeenkomst tevens in het teken van het genereren van functionele wensen. Deze paragraaf beschrijft de aandachtspunten voor het datacentrum die tijdens de bijeenkomst naar voren kwamen.

De deelnemers benoemden zes thema's waar het datacentrum zijn aandacht op moet richten voor succesvol gebruik:

1. **Kwaliteit van de data:** De kwaliteit van de onderzoeksdata is voor de deelnemers van groot belang. Goede beschrijving van het onderzoek zorgt ervoor dat het *herhaalbaar* is, terwijl een consistente werkwijze bij databewerking binnen collecties bepalend is voor de mogelijkheden tot hergebruik. Beide aspecten dragen bij aan de *valorisatie* van kennis.
2. **Standaardisatie:** Om (her)gebruik van data te stimuleren is het van belang dat de data op een *eenduidige wijze opgeslagen* is. Hier tegenover staat dat voor het aanleveren van datasets bij voorkeur *geen vast formaat* gebruikt wordt om te voorkomen dat experimenteel onderzoek dat afwijkt van de standaard niet opgeslagen kan worden.
3. **Gebruikersgemak voor de dataproducent:** Voor de dataproducent is *betrouwbaarheid* van de dataopslag erg belangrijk, zowel tijdens als na afloop van het onderzoek. De opslag van data moet *eenvoudig* zijn, waardoor de onderzoeker de archivering zelf kan uitvoeren. De onderzoeker moet tijdens zijn onderzoek al gebruik kunnen maken van het datacentrum in een *afgeschermd* omgeving. Na afloop van het onderzoek kan de dataset vervolgens op eenvoudige wijze openbaar gemaakt worden. De onderzoeker moet ondersteund worden bij het beschrijven van de dataset. Het gebruik van *vaste templates* kan de onderzoeker hierbij faciliteren.
4. **Dataopslag en werkproces:** De centrale dataopslag moet leiden tot een *kostenbesparing* voor het uitvoeren van onderzoek. Onderzoekers kunnen vervolgonderzoek doen naar eerder behaalde resultaten. De centrale dataopslag moet daarom *onderdeel worden van het onderzoeksproces* en deel uitmaken van de onderzoeksopzet. Hiertoe moeten onderzoekers worden *aangestuurd door het management* van hun organisatie. *De onderzoeker bepaalt* echter zelf welke data hij opslaat in het datacentrum (afstemming met de sponsor). Het datacentrum moet het mogelijk maken om in ieder geval bewerkte onderzoeksdata op te slaan. Voor het opslaan van ruwe onderzoeksdata met een grote capaciteit zal een kosten-batenanalyse moeten plaatsvinden. Dit moet als onderdeel van het financieringsmodel meegenomen worden. Alternatieve geldstromen kunnen ontstaan uit *dienstverlening die rondom het datacentrum* plaatsvindt. Deze diensten kunnen op een later moment worden ontwikkeld.
5. **Vindbaarheid en metadata:** Om hergebruik van onderzoeksdata te stimuleren moeten de data goed *vindbaar en beschreven* zijn door middel van metadata. Voor het invullen van de metadata moeten onderzoekers gebruik kunnen maken van *standaard templates*. Het moet mogelijk zijn om in een publicatie te *refereren* aan de dataset en vice versa. *Online interactie* over de kwaliteit van datasets hoeft niet plaats te vinden bij het datacentrum. Hier zijn specialistische onderzoeksfora voor beschikbaar. In veel gevallen is het wenselijk om *de bron van de dataset* te kunnen achterhalen. Dit komt de betrouwbaarheid van de data ten goede.

6. Toegang en gebruikerseisen: Toegang tot het datacentrum is in beginsel *openbaar* voor wetenschappelijke instellingen. Ook openbare *commerciële datasets* kunnen in het datacentrum worden opgeslagen. Voor het gebruik van datasets door commerciële instellingen kan een vergoeding gevraagd worden. In het gebruik van de datasets is eenvoud belangrijk. Naast *vindbaarheid* moeten data gemakkelijk *over te hevelen* zijn naar de eigen omgeving.

De resultaten uit de bijeenkomst zijn vrijwel geheel in lijn met het veldonderzoek in de vorm van de twee casestudies. Met de uitkomsten van de bijeenkomst kunnen we de bevindingen uit de casestudies daarom bevestigen.

Tijdens de bijeenkomst zijn de volgende aanbevelingen toegevoegd aan de al genoemde aanbevelingen.



Aanbeveling

- Bewerkingen op de data moeten als onderdeel van de data bij de data opgeslagen kunnen worden, zodat je data kunt reviewen.
- Standaardisatie van grootheden en units (vocabularies) is noodzakelijk voor hergebruik van onderzoeksdata.
- De wetenschapper bepaalt welke data hij opslaat maar een datacentrum moet ook aanbod kunnen weigeren.

Op basis van deze bevindingen wordt in het volgende hoofdstuk het Programma van functionele wensen opgesteld.

4. Programma van wensen

Hieronder volgt een overzicht van alle tijdens het onderzoek door dataproducenten geformuleerde aanbevelingen. De aanbevelingen zijn samengevat tot een programma van wensen waarbij voor elke aanbeveling is aangegeven tot wie de aanbeveling gericht is.

Aangezien het onderzoek zich heeft gericht op de meerwaarde die een datacentrum 'C' aan dataproducenten 'P' kan bieden, hebben we ons bij de indicatie beperkt tot deze twee 'doelgroepen'.

Voor de volledigheid wordt vermeld dat sommige aanbevelingen van beide partijen inzet verwachten maar daar is altijd een keus gemaakt voor een partij die volgens de projectgroep de grootste verantwoordelijkheid heeft. Als dit geen van beide 'doelgroepen' is, is dit aangegeven met een 'X'.

Table 1: Programma van wensen.

PROGRAMMA VAN WENSEN		
➔ Organisatie		
▪ Het datacentrum moet data zoveel mogelijk open access aanbieden om hergebruik te bevorderen		P
▪ Kosten voor opslag moeten geen belemmering zijn voor het aanleveren van wetenschappelijke data		C
▪ Er moet een businessmodel opgesteld worden voor duurzame financiering van het datacentrum		C
▪ Bij de start van een onderzoek moet al een inschatting worden gemaakt van het belang van lange termijn toegankelijkheid van de data		P
▪ Voor relevante datasets moet in de onderzoeksplanning tijd gereserveerd worden voor opslag van data in een datacentrum		P
▪ Het datacentrum moet de focus leggen op de dataproducenten (in het bijzonder post docs en promovendi) voor het verzamelen van data. Voor hergebruik bestaat de doelgroep uit zowel de hierboven genoemde dataproducenten als betrokkenen bij het wetenschappelijk onderwijs zoals docenten en studenten		C
▪ Het datacentrum kan extra diensten tegen betaling aanbieden maar biedt bij voorkeur zonder kosten toegang tot de datasets		C
▪ Het datacentrum moet streven naar een internationale scope, waarbij aandacht is voor het adopteren van meerdere standaarden en faciliteiten voor kwaliteitscontrole van de datasets		C
▪ Het datacentrum moet streven naar een zo hoog mogelijke kwaliteit van de opgenomen datasets (zie ook 'Cultuur')		C
▪ De onderzoeker is verantwoordelijk voor de kwaliteit van de datasets		P
▪ De in het datacentrum opgeslagen datasets moeten citeerbaar zijn		C
▪ Fouten moeten kunnen worden aangepast door (collega) onderzoekers zonder het bronbestand te veranderen (bijv. door middel van versiebeheer)		C
▪ De metadata van datasets moeten doorzoekbaar zijn (zie ook 'Systemen')		C
▪ Binnen grote datasets moeten zoek- en selectiemogelijkheden zijn (zie ook 'Systemen')		C
▪ In de metadata die bij de dataset wordt opgeslagen moet het onderzoek goed beschreven zijn		P
▪ Het datacentrum moet de opslag van data al tijdens het onderzoek mogelijk maken (incl. versiebeheer), tijdelijke afscherming van onderzoekgegevens moet daarom mogelijk zijn (zie ook 'Cultuur')		C
▪ Het datacentrum moet het uitbrengen van verrijkte publicaties door wetenschappers faciliteren		C
▪ Voor datasets in het datacentrum moeten de eigendomsrechten en gebruiksvoorwaarden duidelijk beschreven worden (zie ook 'Cultuur')		C
▪ Alle data moet een authentieke bron hebben die herleidbaar is		P

- | | |
|--|---|
| ▪ Zowel technisch als functioneel beheer rondom het datacentrum moet geregeld zijn | C |
| ▪ Het datacentrum moet de voordelen van gecentraliseerde opslag, beheer en ontsluiting duidelijk kenbaar maken aan de doelgroepen | C |
| ▪ Het datacentrum kan de dataproducenten begeleiden bij het inrichten van een gebruikersorganisatie en definiëren van werkprocessen, bijvoorbeeld door middel van aanbevelingen. | C |
| ▪ Bewerkingen op de data moeten als onderdeel van de data bij de data opgeslagen kunnen worden, zodat je data kunt reviewen | C |

➔ Cultuur

- | | |
|---|---|
| ▪ Stimuleer de open houding van onderzoekers ten opzichte van het delen van onderzoeksdata | X |
| ▪ Deponeren van data bij het datacentrum moet geen verplichting zijn, de motivatie van dataproducenten moet voornamelijk worden ingegeven door het bieden van meerwaarde door het datacentrum | C |
| ▪ Stimuleer beheer van data door datacentra door middel van beloning van dataproducenten door financiers en leidinggevenden | X |
| ▪ Voor datasets in het datacentrum moeten de eigendomsrechten en gebruiksvoorwaarden duidelijk beschreven worden (zie ook 'Organisatie') | C |
| ▪ Het datacentrum moet tijdelijke afscherming van onderzoeksdata toestaan (zie ook 'Organisatie') | C |
| ▪ Stimuleer het hergebruik van datasets door datamanagement op te nemen in het (wetenschappelijk) onderwijs | X |
| ▪ Het datacentrum moet 'gebruiks'incentives bieden aan onderzoekers, onder andere efficiënt databeheer en citeerbaar maken van datasets | C |
| ▪ Stimuleer wetenschappelijke waardering voor het delen van datasets | X |
| ▪ Het opslaan van data in het datacentrum moet eenvoudig en efficiënt kunnen gebeuren | C |
| ▪ De onderzoeker moet data zelf (kunnen) invoeren in het datacentrum | P |
| ▪ Het datacentrum moet de kwaliteit van opgenomen datasets bewaken (zie ook 'Organisatie') | C |
| ▪ De wetenschapper bepaalt welke data hij opslaat maar een datacentrum moet ook aanbod kunnen weigeren | P |

➔ Competenties

- | | |
|---|---|
| ▪ Dataproducten moeten plannen kunnen opstellen voor datamanagement | P |
| ▪ Het datacentrum moet een korte (online) training aanbieden om het correct opslaan van de onderzoeksgegevens te bevorderen | C |
| ▪ Er moeten richtlijnen beschikbaar zijn voor het invullen van de metadata | C |
| ▪ Er moet een goede handleiding beschikbaar zijn voor het gebruik van het datacentrum | C |
| ▪ Standaardisatie van grootheden en units (vocabularies) is noodzakelijk voor hergebruik van onderzoeksdata | P |

Systemen	
▪ De user interfaces voor zowel aanleveren als hergebruik moeten eenvoudig en intuïtief functioneren	C
▪ De user interface voor hergebruik moet visualisatie van (meta)data mogelijk maken ten behoeve van selectie (bijv. geografische posities op een kaart, grafieken/plots, picture galleries, etc.)	C
▪ Er moet een (flexibele) standaard komen voor (beschrijvende) metadata bestaande uit een kleine minimum set met voldoende optionele velden voor een beschrijving die hergebruik van de data mogelijk maakt	C
▪ De metadata van datasets moeten doorzoekbaar zijn (zie ook 'Organisatie')	C
▪ Binnen grote datasets moeten zoek en selectiemogelijkheden zijn (zie ook 'Organisatie')	C
▪ Metadata van datasets moeten indexeerbaar zijn door externe zoekmachines als Google	C
▪ Het datacentrum moet o.a. de volgende functionaliteiten bieden: checklist voor kwaliteitscontrole, gebruikersstatistieken en een mogelijkheid om feedback te geven op datasets (zie ook 'Cultuur')	C
▪ Het datacentrum moet de opslag van verschillende soorten onderzoeksdata faciliteren	C
▪ In het datacentrum moeten zowel ruwe als bewerkte onderzoeksdata opgeslagen kunnen worden	C
▪ De opslag van data in het datacentrum moet gebeuren in een uitwisselbare standaard	C
▪ Downloaden van datasets moet mogelijk zijn in verschillende gangbare formats	C
▪ Het datacentrum moet toegankelijk zijn via internet	C
▪ Het datacentrum moet minimaal 'betrouwbare' opslag (dat wil zeggen veiligstellen en toegankelijk houden van data) kunnen garanderen	C

- P = Dataproductent
 C = Datacentrum
 X = Andere dan bovengenoemde

5. Bijlagen

5.1. Bronnen

- [APSR, 2008] Henty, Margaret, Belinda Weaver, Stephany Bradbury, Simon Porter; Investigating Data Management Practices in Australian Universities, Australian Partnership for Sustainable Repositories (APSR), 2008
- [CESSDA, 2009] <http://www.cessda.org/sharing/depositing/1/>
- [DANS, 2007] Balkestein, Marjan, Heiko Tjalsma; The ADA Approach: Retro-archiving data in an Academic Environment, DANS, 2007
- [DANS, 2008] DANS, "Datakeurmerk (Data Seal of Approval)".
- [DARELUX, 2007] DARELUX Eindrapportage, 2007
- [DART] Tsoi, Ah Chung, Dataset Acquisition, Accessibility, Annotation, e-Research Technologies (DART) Project
- [ECDL, 2007] Rice, Robin, Peter Burnhill, Christine Rees, Anne Robertson; Repository Junction and Beyond at the EDINA (UK) National Data Centre, ECDL 2007
- [Fish, 2003] Fish, Mark, Effective Access and Retrieval of Laboratory Data to Enable Knowledge Management, American Laboratory, 2003
- [Godard, 2003] Godard, Beatrice, Jörg Schmidtke, Jean-Jacques Cassiman, Ségolène Aimé; Data storage and DNA Banking for Biomedical Research: informed consent, confidentiality, quality issues, ownership, return of benefits. A professional perspective, European Journal of Human Genetics, 2003
- [Hahsler] Hahsler, Michael, Stephan Koch; Discussion of a Large-Scale Open Source Data Collection Methodology, Hawaii International Conference on System Sciences, 2005
- [Helly] Helly, J. F. et al., Controlled Publication of Digital Scientific Data en_US. Communications of the ACM, 45(5): 97-101.
- [Jacobs, 2004] Jacobs, James A., Charles Humphrey; Preserving Research Data, Communications of the ACM, 2004
- [Kramer, 2006] Kramer, Rutger; Possibilities for Advanced Dissemination and Durable Storage of Scientific Data on the Grid, Data Archiving and Networked Services (DANS), TEAA 2006
- [Kuula, 2008] Arja Kuula & Sami Borg (2008). Open Access to and Reuse of Research Data
- [NSF, 2006] NSF, "Cyber infrastructure Vision for 21st century Discovery".
- [PANGAEA, 2007] Diepenbroek, Michael, Uwe Schindler, Hannes Grobe; PANGAEA, an ICSU World Data

Center as a Networked Publication and Library System for Geoscientific Data, 2007

- [Pfaltz] Pfaltz, John L.; What Constitutes a Scientific Database? University of Virginia
- [Piwowar, 2007] Piwowar, Day en Fridsma (Sharing detailed research data is associated with increased citation rate).
- [RIN, 2008] RIN Report, "To Share or not to Share: Publication and Quality Assurance of Research Data Outputs".
- [Schweik, 2005] Schweik, Charles M., Alexander Stepanov, J. Morgan Grove The Open Research System, a Web-Based Metadata and Data Repository for Collaborative Research, Computers and Electronics in Agriculture, 2005
- [UKDA, 2008] United Kingdom Data Archive (UKDA), Frequently Asked Questions: Data Management and Sharing, University of Essex, 2008
- [Valle] Scientific Data Management – an introduction (Mario Valle web:
<http://personal.cscs.ch/~mvalle/sdm/scientific-data-management.html>)
- [Vickers, 2006] Vickers, Andrew J.; Whose Data is it Anyway? Sharing Raw Data from Randomized Trials, Biomed Central, 2006
- [Warden] Warden, Sean, Arturo Sanchez, Sherif Elfayoumi; Toward a General Framework for Building Scientific Data Sharing Web Services, Univeristy of Florida

5.2. Overige projectresultaten

Behalve deze eindrapportage zijn de volgende documenten met tussenresultaten en onderliggend materiaal ook beschikbaar:

- Samenvatting Literatuur WD&D, v0.4, 19-03-2009
- Rapportage Voorbereidend onderzoek & Good practice case, v0.4, 23-04-2009
- Interview vragen Waterbouwkunde
- Werkprocesbeschrijving
- Verslag Expertbijeenkomst