



SURF

SHARE

TOEGANG TOT ONDERZOEKSDATA



OVER DE SURF SHARE SERIE

‘Toegang tot Onderzoeksdata’ verschijnt in de SURFshare serie boeken waarmee de kennis die wordt opgedaan in het SURFshare-programma beschikbaar wordt gemaakt.

De snelle veranderingen op het gebied van ICT bieden een breed scala aan mogelijkheden voor het vinden, creëren, verspreiden en openbaar toegankelijk maken van onderzoeksresultaten. ICT versnelt niet alleen de klassieke onderzoeksprocessen en onderzoekscommunicatie, maar verandert ook de onderzoekscyclus van aard. Door de toegenomen mogelijkheden van kennisdeling en -verspreiding raken traditionele publicaties, onderzoeksgegevens en instrumenten zoals modellen, algoritmes en visualisaties steeds meer met elkaar verweven.

Met het SURFshare-programma wil SURFfoundation een gemeenschappelijke infrastructuur realiseren die de toegankelijkheid en de uitwisseling van onderzoeksinformatie bevordert.

Dit programma loopt van 2008 tot 2011 en heeft tot doel om samen met alle Nederlandse universiteiten en hogescholen, NWO en KNAW te werken aan het optimaal delen van onderzoeksresultaten, gebruik makend van de nieuwste ICT mogelijkheden.

De SURFshare serie biedt kennis en ervaring die binnen het SURFshare-programma wordt opgedaan. Het bevat handzame en gebruiksklare informatie over nieuwe ontwikkelingen en resultaten met duidelijke uitleg, voorbeelden en good practices.

De SURFshare serie is in eerste instantie bedoeld voor mensen die onderzoekers bij universiteiten en hogescholen ondersteunen bij het toegankelijk maken van hun onderzoeksresultaten. Met behulp van de beschreven ervaringen en mogelijke oplossingen kunnen zij de eindgebruiker informatie en oplossingen op maat aanbieden.

Andere delen in de serie

- Nederland loopt warm voor Open Access, januari 2010
- SURFshare nulmeting: Nederlandse Academische repositories 2008, februari 2010
- Verrijkte publicaties: onderzoeksresultaten in samenhang, maart 2010
- Stromen van kennis tussen hogescholen en praktijk, januari 2011

INHOUDSOPGAVE

	INLEIDING	4
1	DE CIRKEL VAN ONDERZOEKSDATA	6
	Door: Eric Sieverts (Universiteit Utrecht)	
2	EERLIJK DE DATA DELEN	12
	Door: Werna Oosterbaan (NRC Handelsblad)	
3	ONDERZOEKSDATA FORUM VOOR BETERE TOEGANG EN MANAGEMENT	16
	Door: Wilma Mossink (SURFfoundation)	
4	DE EERSTE STAPPEN VOOR ONDERSTEUNING VAN DATAMANAGEMENT	20
	Door: Jeroen Rombouts (3TU.Datacentrum)	
5	DATASTEWARDS MOETEN ORDE BRENGEN IN DE CHAOS VAN ONDERZOEKSDATA	24
	Interview met Rob Grim (Universiteit Tilburg) en Marianne van der Heijden (NIOO-KNAW) Door: Aad van de Wijngaart	
6	VERKORTE JURIDISCHE WEGWIJZER VOOR RUWE DATA: DE STOPLICHT-ROUTE	28
	Door: Madeleine de Cock Buning, Allard Ringnalda, Tina van der Linden (CIER-Universiteit Utrecht)	
7	DATA IN AUSTRALIË EN NEDERLAND: EEN ZAAK VAN LANGE ADEM	32
	Interview met Andrew Treloar (ANDS) en Jeroen Rombouts (3TU.Datacentrum) Door: Lilian van der Vaart	
8	CHECKLIST HELPT BIJ SELECTIE TE BEWAREN ONDERZOEKSDATA	36
	Door: Annemiek van der Kuil (SURFfoundation)	
9	ORGANISATIE VAN OPSLAG EN BESCHIKBAARSTELLING ONDERZOEKSDATA	38
	Door: Maurits van der Graaf (Pleiade Management en Consultancy)	
10	KWALITEIT VAN ONDERZOEKSDATA: WAT ONDERZOEKERS WILLEN	42
	Door: Maurits van der Graaf en Leo Waijers	
11	DE VISIE VAN NWO: STOKKEN EN WORTELS	46
	Interview met Ron Dekker (NWO) Door: Aad van de Wijngaart	
12	INTERNATIONALE WORKSHOP: ONDERZOEKERS STIMULEREN TOT HET DELEN VAN ONDERZOEKSDATA	50
	Door: Christoph Kümmel (Deutsche Forschungs Gemeinschaft)	
13	VOOR ONDERZOEKSDATA IS HET ESSENTIEEL DAT ZE ALTIJD GELINKED ZIJN AAN PUBLICATIES	54
	Interview met Eefke Smit (STM) Door: Lilian van der Vaart	
14	HET WERELDWIJDE LABJOURNAAL	60
	Door: Jan Top (Vrije Universiteit)	
15	CONCLUSIE EN VERDER LEZEN	64
	Door: Wilma Mossink en Keith Russell	

INLEIDING



De belofte van het internet is dat bijna alle wetenschappelijke kennis direct beschikbaar kan komen voor iedereen die daar behoefte aan heeft. Dit biedt ongekennde mogelijkheden aan de wetenschap!

Voorheen konden wetenschappers alleen tekstuele bronnen verspreiden en delen, maar dankzij het internet zijn daar immens veel mogelijkheden bijgekomen. Denk aan gedigitaliseerde afbeeldingen van door archeologen opgegraven artefacten, of een enorme database waarin alle menselijke genen beschreven staan. Met toevoeging van beeld, video of audio zal niet alleen de communicatie tussen wetenschappers veranderen, maar zal de wetenschap zelf een ingrijpende verandering – een paradigmashift – kunnen ondergaan.

De digitale toegankelijkheid van bronnen opent namelijk ook de deur voor het laten analyseren van deze bronnen door computers. Een dergelijke benadering van wetenschap wordt ook wel 'data intensive science' genoemd¹. Wetenschappelijke inzichten zullen door het inzetten van zoek- en rekenkracht van computers sneller ontstaan omdat er meer informatie in de analyses kan worden meegenomen dan menselijkerwijs behapbaar is.

Hiernaast kan onderzoek via het internet door meerdere onderzoekers tegelijk verricht worden. Het is niet meer zo dat een onderzoeker reageert op of refereert aan een kant en klaar artikel van zijn collega, maar nu kunnen ze samen aan teksten én aan databestanden werken. Want soms hebben data méér waarde dan het artikel dat op die data is gebaseerd. Het is wel van belang dat de onderzoeker ook credits zou moeten krijgen voor het verzamelen en delen van data in plaats van, zoals nu nog vaak het geval is, alleen op basis van zijn artikelen in de vorm van tekst.

Een methodologisch voordeel van 'open onderzoeksdata' is tenslotte dat onderzoek erdoor veel gemakkelijker reproduceerbaar en verifieerbaar zal zijn, en dat het mogelijk is de data te hergebruiken. Het is niet meer zo dat voor ieder nieuw onderzoek ook opnieuw data gegenereerd moeten worden. De verzamelde data ten behoeve van een onderzoek naar de oorzaken van ziekte A zijn wellicht ook heel goed te gebruiken voor een onderzoek naar ziekte B.

Over die vergaande invloed van het internet op de onderzoekspraktijk gaat deze publicatie in de SURFshare-serie. Het zijn zegenrijke beloftes die 'open data' bieden, maar zoals uit het voorgaande blijkt, heeft de wetenschap nog een stap te zetten – een paradigmashift te ondergaan – alvorens de zegeningen geteld kunnen worden. Dat komt mede omdat er nog enige haken en ogen zitten aan het realiseren van open data.

Welke data bewaar je wel, welke niet? Hoe toegankelijk, hoe 'open' maak je die data? Hoe richt je een datarepository in? En wat is de juridische status van je database? Mag een onderzoeker bijvoorbeeld data zo maar hergebruiken? Hoe zit het met het auteursrecht op data? Om nog maar te zwijgen over de rol die onderzoeksfinanciers en uitgevers spelen.

Er liggen in het internet mooie beloftes voor de wetenschap, maar voor het inwilligen ervan moeten praktische, juridische maar ook mentale hobbels genomen worden. Dit is de reden waarom dit boek is samengesteld: SURFfoundation hoopt, samen met vele partners in het veld die aan deze publicatie hebben bijgedragen, een bijdrage te kunnen leveren aan het wegnemen van de hobbels en het inwilligen van de beloftes.

John Doove
SURFfoundation

¹ Tony Hey e.a. (2009) *The Fourth Paradigm, Data-Intensive Scientific Discovery*
<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

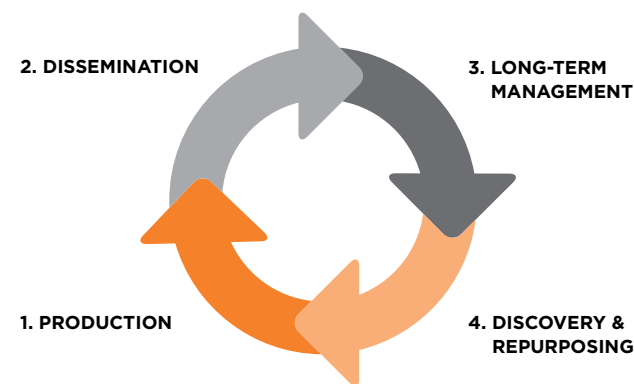
1 DE CIRKEL VAN ONDERZOEKSDATA

Door: Eric Sieverts
Universiteit Utrecht

In toenemende mate worden onderzoeksdata opgeslagen, bewaard en weer beschikbaar gesteld, bijvoorbeeld voor vervolgonderzoek. Daarin doorlopen die data doorgaans een bepaalde levenscyclus: in het begin zijn ze het bezit van de onderzoeker die ze verzamelt, vervolgens komen ze beschikbaar voor bijvoorbeeld collega onderzoekers en tenslotte kunnen ze openbaar worden gemaakt in het publieke domein. Die cyclus roept vragen op: welke data bewaar je, hoe bewaar je ze, hoe maak je ze toegankelijk en wat te doen met hele grote data-bestanden?

Wetenschappelijke vooruitgang kent vaak kringlopen. Resultaten van onderzoek worden gepubliceerd of op een andere wijze gecommuniceerd en leiden weer tot nieuwe interpretatie, nieuw onderzoek en nieuwe voortgang. Het is niet voor niets dat de wetenschappelijke zoekmachine Google Scholar, in een variant op een uitspraak van Newton, als motto hanteert: Stand on the shoulders of giants. Dat die kringloopgedachte ook van toepassing is op de onderzoeksgegevens die aan wetenschappelijke publicaties ten grondslag liggen, wordt nog niet zo lang onderkend. Ook daarvoor kun je mooie cirkeltjes tekenen. Een Canadees onderzoek naar de Data Lifecycle doet dat simpel in vier stappen: van productie via disseminatie, permanente opslag naar ontdekking (figuur 1).

Het "Curation Lifecycle Model" van het Britse DCC (Data Curation Center, figuur 2) maakt die cirkelgang wat complexer door ook tussenstappen en terugkoppelingen te benoemen, maar is toch niet wezenlijk anders. In Nederland wordt in het kader van het Onderzoeksdataforum nu ook aandacht geschonken aan de verschillende aspecten die te maken hebben met opslaan, bewaren en weer beschikbaar stellen van onderzoeksdata. De bijdragen in deze bundel vormen daar een weerslag van.



Figuur 1: Data Lifecycle.

Uit: Stewardship of Research Data in Canada: A Gap Analysis (October 2008)

<http://www.library.carleton.ca/documents/researchdataseminar/GapAnalysis.pdf>

Soorten datacollecties

Als we het over datacollecties hebben, dan kunnen die van allerlei aard zijn. In deze bundel gaat het vooral om onderzoeksdata, de gegevens die de primaire resultaten zijn van allerlei soorten onderzoeken. Maar ook die heb je nog in allerlei soorten, zoals de getallen die rechtstreeks uit een meetopstelling komen, de ruwe uitkomsten van een enquête,



Figuur 2: Curation Lifecycle Model van het DCC.

Uit: <http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf>

geluids- of video-opnamen van dier- of mensgedrag, of door menselijk toedoen al bewerkte en geaggregeerde gegevens. Zulke onderzoeksdata worden door de makers altijd wel enige tijd ergens opgeslagen en bewaard. Maar vragen omtrent het langduriger, veiliger en duurzamer opslaan en het zorgvuldig beheren en toegankelijk houden (of maken) zijn pas recenter in de belangstelling komen te staan.

Een andere interessante categorie data is die van de Linked Data, datacollecties die vrij op internet beschikbaar zijn. Door op gestandaardiseerde wijze betekenis aan die gegevens toe te kennen, kunnen ze ook autonoom door computerprogramma's gebruikt worden. Dit wordt verder uitgelegd in hoofdstuk 15.

Primaire onderzoeksdata

Universiteiten, onderzoeksinstituten en hogescholen richten al lang repositories in voor het blijvend opslaan, het digitaal beschikbaar stellen en het uitwisselen van publicaties uit de eigen instelling. In toenemende mate realiseert men zich dat het belangrijk is ook de aan die publicaties ten grondslag liggende primaire onderzoeksdata te bewaren en eventueel beschikbaar te stellen.

Vanuit tijdschriftredacties en subsidiërende instanties stelt men soms al de eis dat die gegevens ergens beschikbaar moeten zijn. Onder meer ten behoeve van de verifieerbaarheid van gerapporteerde resultaten, analyses en conclusies. In hoofdstukken 11 en 12 zijn deze spelers aan het woord.

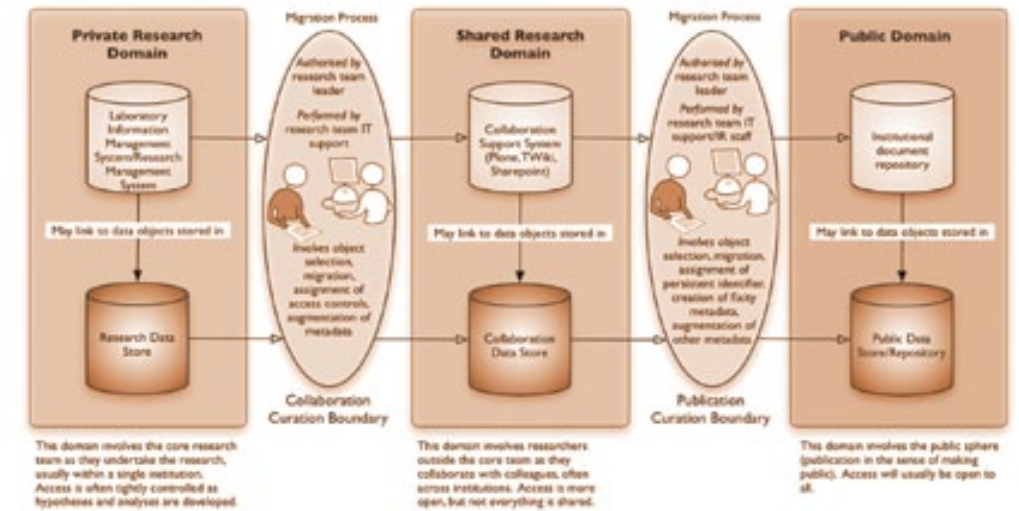
Om de link te leggen tussen publicaties en datasets moet er wel met een permanent webadres naar de dataset gelinkt kunnen worden. Het resultaat wordt dan wel een 'Verrijkte Publicatie' of een 'Complex Object' genoemd. Daarin kunnen niet alleen multimediale illustraties, maar ook datasets op een gestructureerde wijze aan teksten gekoppeld worden.

Curation continuum

In de Angelsaxische wereld wordt het beheer van datacollecties wel aangeduid als research data curation. In de loop van de levenscyclus van onderzoeksdata veranderen de rol, het gebruik en het daaruitvolgende beheer van die data geleidelijk. Men spreekt daarom wel van een curation continuum (figuur 3). Om dat wat modelmatig te kunnen beschrijven worden ook curation boundaries aangegeven, de momenten waarop data duidelijk een andere fase ingaan.

In een eerste fase zijn de gegevens nog helemaal in het privé-domein van de onderzoekers: ze beheren hun data zelf, ze werken er nog aan, vullen ze aan, analyseren ze, enzovoorts.

In een tweede fase zijn de data beschikbaar voor samenwerking met partners binnen of buiten de organisatie. Ze moeten dan dus zo worden opgeslagen en van metadata voorzien, dat die partners er toegang toe hebben en ook hun betekenis kunnen interpreteren. Wat betekenen de getalletjes, waaraan is gemeten, wat voor proefpersonen zijn



Figuur 3: The Curation Continuum

Uit: <http://ands.org.au/guides/curation.continuum.pdf>

gebruikt, wat wordt op de filmpjes waargenomen? Voor echte buitenstaanders mogen ze in dat stadium meestal nog niet toegankelijk zijn. Hoe streng men daarbij wil zijn, kan ook te maken hebben met eventuele commerciële belangen.

In een derde fase kunnen onderzoeksdata helemaal in het publieke domein terecht komen: als er geen dringende redenen zijn die dat ongewenst maken, kan iedereen er gebruik van maken. Er kan dan naar gelinkt worden en ook moeten ze duurzaam bewaard worden. In sommige gevallen zal dat zich overigens niet tot de data zelf beperken, maar moet ook iets gedaan worden om de interpretatie van de bestanden met software te kunnen garanderen.

'In allerlei verband wordt gesproken over een op handen zijnde data deluge'

Bij elke overgang naar een volgende fase zullen ook beslissingen genomen moeten worden. Zijn de data toegankelijk? Is de metadata voor die fase nog adequaat? En moeten de gegevens überhaupt nog bewaard worden? Zo zullen zeker niet alle ooit geproduceerde data in het derde stadium terecht komen. Behalve commerciële afwegingen kunnen er ook andere argumenten zijn om gegevens niet te bewaren, zoals kwaliteit en hoeveelheid van de gegevens, te sterke koppeling van gegevens aan heel specifieke meetapparatuur, of de beschikbaarheid van nuttiger geaggregeerde of bewerkte gegevens. De mogelijke criteria voor het selecteren van welke data moeten worden bewaard worden beschreven in hoofdstuk 5.

De drie genoemde fases komen we in de Nederlandse praktijk ook al tegen. Zo biedt het sinds kort gestarte Utrecht Dataverse Network⁴ een infrastructuur voor de eerste twee fases, waarbij de eigenaar van de gegevens nog helemaal bepaalt wie toegang heeft tot de gegevens en wie er gebruik van mag maken. DANS (Data Archiving & Networked Services)⁵ is in het leven geroepen om de humaniora en sociale wetenschappen te ondersteunen in de tweede en vooral de derde fase. In deze fasen staat hergebruik en digitale duurzaamheid voorop.

Data deluge

In allerlei verband wordt op dit moment gesproken over een op handen zijnde gegevensstortvloed, een data deluge. Bij onderzoeksdata denk je daarbij al gauw aan de enorme hoeveelheden gegevens die grootschalige experimenten, zoals met de nieuwe deeltjesversneller van CERN, gaan opleveren. Of aan de gegevens die uit allerlei gekoppelde astronomische waarnemingssystemen komen. In dergelijke projecten wordt bij de opzet meteen al rekening gehouden met de massale datastroom die ze gaan opleveren. Grote datahoeveelheden zijn echter niet meer het alleenrecht van de bèta-wetenschappen. In de humaniora en sociale wetenschappen worden steeds vaker beelden vastgelegd op hoge resolutie en met hoge snelheid. Dit levert grote datastromen op. Men begint zich dan ook te realiseren dat data management van te verwachten onderzoeksdata een noodzakelijk onderdeel van elke projectaanvraag zou moeten zijn.

Linked data

De grote dataverzamelingen die intussen op internet beschikbaar zijn, betreffen op dit moment vaak nog geen onderzoeksdata, maar allerlei algemene soorten gestructureerde gegevens. Opdat computers deze gegevens op standaardwijze kunnen lezen, interpreteren en gebruiken, en om zo allerlei systemen aan elkaar te kunnen koppelen, worden die data steeds vaker gecodeerd volgens de RDF-standaard (het Resource Description Framework). Vanwege dat koppelen spreekt men over Linked (Open) Data om deze dataverzamelingen te omschrijven.

Voor Tim Berners Lee speelt een stimulerende rol bij de sterke toename van het aantal zo beschikbare datacollecties. Die ziet hij namelijk als de ruggengraat voor het semantisch web. Tamelijk centraal in die wolk staat de DBpedia⁶, een grote verzameling RDF-gecodeerde gegevens die uit de Wikipedia zijn afgeleid. Andere systemen linken daarheen om er geautomatiseerd extra informatie over allerlei onderwerpen aan te kunnen onttelen. De letters DB in DBpedia zijn een aanwijzing dat het hier eigenlijk om een database-achtige extensie van het web gaat. Op de site Linkeddata.org is een indrukwekkende wolk van dergelijke onderling gelinkte datacollec-

ties gevisualiseerd⁷. Ook daar is intussen sprake van een “deluge” van tientallen miljarden beschikbare RDF-statements en varianten daarop.

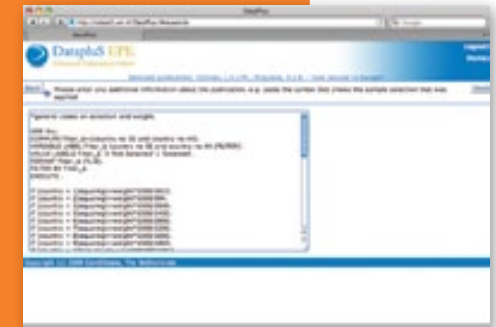
Ook wetenschappelijke onderzoeksdata zullen uiteindelijk zo beschikbaar gesteld kunnen en moeten worden. Op 6 oktober 2010 twitterde zelfs @NeelieKroesEU daarover⁸: “Scientific data it is too valuable to be locked away. #publicdata #scientificdata #opendata and #wegov <http://tiny.cc/scientificdata>”

SURFproject

DATAPLUS: ZICHTBAARDER ONDERZOEK

Het in 2009 uitgevoerde SURFshare tender project Data-plus wil inzicht geven in welke statistische data en bewerkingen van die data zijn gebruikt om tot de conclusies te komen die in wetenschappelijke artikelen worden getrokken. Het gaat om artikelen uit verkiezingsonderzoek en waardenonderzoek.

Om de koppelingen mogelijk te maken is een hulpmiddel ontwikkeld (Enhanced Publication editor) waarmee onderzoekers op relatief gemakkelijke wijze metadata aan de onderzoeksdata kunnen toevoegen en de onderzoeksdata aan de publicaties kunnen koppelen.



De zo Verrijkte Publicaties worden in twee portals aangeboden en bieden onderzoekers twee voordelen. De zichtbaarheid van het onderzoek wordt vergroot omdat het zoekproces naar bestaande literatuur wordt ondersteund door het samenbrengen van relevante publicaties en onderliggende onderzoeksdata. Bovendien wordt hergebruik gestimuleerd omdat het koppelen van data aan publicaties onderzoekers de mogelijkheid biedt onderzoek te repliceren en voort te bouwen op de gebruikte onderzoeksdata.

⁴ <http://dataverse.library.uu.nl>

⁵ <http://www.dans.knaw.nl>

⁶ <http://dbpedia.org/About>

⁷ Linked Data - Connect Distributed Data across the Web, <http://linkeddata.org/>

⁸ <http://twitter.com/#!/NeelieKroesEU/status/26568611661>

2 EERLIJK DE DATA DELEN

Door: **Warna Oosterbaan**
NRC Handelsblad



Esther Jansma

Esther Jansma laat haar dendrolab zien, een lichte ruimte in het gebouw van de Rijksdienst voor het Cultureel Erfgoed (RCE) in Amersfoort. Ze is hoogleraar dendrochronologie en paleo-ecologie van het Kwartair aan de Universiteit Utrecht en geeft op de RCE leiding aan het dendrochronologisch onderzoek. Ze is ook een gelauwerd dichter. Een rijzige vrouw, die enthousiast vertelt over bomen, hout en scheepswrakken.

We staan stil bij een onderzoekster die door een microscoop naar een houtmonster tuurt. Uit de breedte van de jaarringen en het patroon dat daarin optreedt kan een dendrochronoloog een stuk hout dateren, want de seizoenen hebben daarin voor altijd hun afdruk achtergelaten. Een warm en vochtig jaar laat in het algemeen een brede ring achter, een koud en droog jaar een smalle ring. "In 1628 voor Christus is de Thera-vulkaan in de Middellandse Zee uitgebarsten", vertelt Jansma. "Op de hele wereld werd het toen koeler en ook bij Nederlandse eiken heeft die uitbarsting tien jaar lang een depressie in de groei veroorzaakt." Die zie je in heel oud hout terug als tien smalle ringen.

Dendrochronologie is een tak van wetenschap waarin elk onderzoeksresultaat telt, want van elke meetreeks kan bij nieuw onderzoek weer met vrucht gebruik worden gemaakt. Jansma is daarom een vurig voorstander van het streven onderzoeksdata duurzaam vast te leggen en ze toegankelijk te maken via een website. In het najaar voltooit ze een groot NWO-project waarin verspreide onderzoeksdata volgens een nieuwe internationale standaard worden ontsloten. "Ik vind het extreem belangrijk dat zulke data beschikbaar

zijn voor iedereen die er belang in stelt. Er zijn onderzoekers die vinden dat de data hun persoonlijk eigendom zijn. Ik snap het wel, maar het is zó verschrikkelijk slecht."

Ook Jan Luiten van Zanden, hoogleraar economische geschiedenis aan de Universiteit Utrecht, kan zich zijn werk niet meer voorstellen zonder onderzoeksdata uit alle werelddelen. "Als je alles wilt weten over Amsterdam in de zeventiende eeuw is het misschien niet zo nodig, maar als je wilt weten hoe bijzonder Amsterdam was, kun je niet zonder de gegevens van andere onderzoekers."

Voortbouwen

De openbaarheid van onderzoeksdata is een onderdeel van het streven naar open access in de wetenschap. De vrije toegang tot wetenschappelijke publicaties is het bekendste doel, maar misschien wel even belangrijk is de vrije beschikbaarheid van de onderzoeksgegevens. Die vormen de basis van de publicaties, maar staan er meestal niet in. Maar met die gegevens kunnen onderzoekers elkaars conclusies kritisch beoordelen, kunnen ze voortbouwen op elkaars werk, en kunnen ze soms zelfs geheel nieuwe onderzoeksvragen beantwoorden. Nieuwe technieken maken deze oude idealen mogelijk. "Vroeger stopte je de oude data in een oude schoendoos en dat was het dan. Maar nu kun je data gemakkelijk opslaan, ze via internet uitwisselen, ze met de computer analyseren", zegt Van Zanden. En daarbij komt dan nog het argument dat die data in verreweg de meeste gevallen verzameld zijn met het belastinggeld van de burgers. Ze zijn niet het persoonlijk eigendom van de onderzoekers, ze zijn eigenlijk van ons allemaal.

Open data is daarom een beginsel dat wordt bepleit door de belangrijkste wetenschappelijke instellingen en onderzoeksfinanciers. KNAW, NWO en ook het ministerie van OCW vragen op gezette tijden aandacht voor het belang van openbaarheid. Vooralsnog zonder dat onderzoekers daartoe verplicht worden. De langverwachte ministeriele nota Informatiebeleid die daar helderheid over had kunnen verschaffen is net weer uitgesteld. Er is al wel een instituut dat zich voor de belangen van open data inspant en dat zijn best doet onderzoeksdata te archiveren en openbaar toegankelijk te maken: het door KNAW en NWO opgezette DANS.

Het lijkt de essentie van de wetenschap, de vrije beschikbaarheid van onderzoeksgegevens. Dus wie is er tegen? Jeugdonderzoeker Wim Meeus van de Universiteit Utrecht bijvoorbeeld. Hij is althans niet in alle omstandigheden voor de vrije beschikbaarstelling van onderzoeksdata. Meeus doet langlopend onderzoek naar adolescenten en familierelaties. "Vrij beschikbaar maken van alle data vind ik niet verstandig. Natuurlijk moeten CBS-datasets met demografische



Jan Luiten van Zanden

gegevens er voor iedereen zijn. Maar je hebt ook andere datasets, uitkomsten van onderzoek waar intellectueel kapitaal in zit. Dan heb je verschillende belangen. Het belang van andere onderzoekers, die graag willen dat je die data met ze deelt. Maar je hebt ook het belang van de oorspronkelijke onderzoeker, en het belang van een aantal promovendi, die nog moeten promoveren op die data.”



Wim Meeus

Dubbel Werk

“Ik ben er niet voor om alles maar op internet te zetten, ik ben voor een gecoördineerd publicatieproces. Op een goeie dataset kun je een groot aantal publicaties baseren en wat je in ieder geval moet voorkomen is dat er dubbel werk wordt gedaan. Dat iemand iets wil onderzoeken waaraan je promovendus al een jaar werkt, of waarover al een artikel is geschreven dat bij de redactie van een tijdschrift ligt.” Maar er is nog wat, zegt Meeus: “Tussen wetenschappers is concurrentie. Als onderzoekers uit mijn groep bepaalde ideeën hebben, willen wij daar als eerste over publiceren.”

Het belang van open data is dus niet overal zo vanzelfsprekend als in de dendrochronologie of in de economische geschiedenis. In het vakgebied van Van Zanden, de economische geschiedenis, is er duidelijk sprake van een gedeeld belang. “Wij willen grote vragen

beantwoorden over de oorzaken van de ongelijkheid, over armoede en rijkdom. Vroeger kon je alleen maar dromen over dat soort vragen, maar nu kun je ze stellen, de data zijn er.” Maar zijn al die data wel allemaal van dezelfde orde? Zijn de data uit het China van de 17de eeuw vergelijkbaar met de Nederlandse? “Niet altijd, je moet bekijken of ze wel kloppen. Dat gebeurt in collaboraties zoals die samenwerkingsverbanden tegenwoordig heten. Je kunt er vaak door bijschatting of bewerking nog iets aan doen. En soms moet je terug naar de oorspronkelijke bronnen, om te bekijken welke analyses erop zijn losgelaten, en of dat niet beter kan.”

En wat levert dat op? Van Zanden brandt los: “Nou, nu weten we bijvoorbeeld dat de voorsprong die Europa in de 19de eeuw door de industriële revolutie neemt, eigenlijk al van de Middeleeuwen dateert! We belandden daarna op een welvaartsniveau dat in de wereldgeschiedenis niet eerder is voorgekomen. Dat kwam door scholing, maar ook door het huwelijkspatroon. Als meisjes op hun twaalfde worden uitgehuwelijkt komen ze niet aan scholing toe en zijn ze ook zelf minder geneigd hun kinderen naar school te sturen. Het perspectief op dergelijke brede inzichten is ook het argument van Esther Jansma. Als je van een stuk hout de ouderdom kunt vaststellen en door vergelijking met andere gegevens ook de herkomst, kun je tot verrassende bevindingen komen. In 2003 werd bij Utrecht een Romeinse praam opgegraven. Jansma onderzocht het eikenhout waarvan het schip gebouwd was. “We konden zien dat de boot in

148 was gebouwd, maar ook dat er in 164 nog een bed in werd geplaatst. Dat schip is dus decennialang gebruikt. We konden ook zien dat het scheepshout uit Nederland kwam. Tot dan toe werd altijd gedacht dat die Romeinse rivierpramen die we in Nederland terugvinden in Duitsland waren gebouwd, dat ze met een lading bouwmaterialen of graan eenmalig de Rijn afzakten en dan hier voor andere doelen werden gebruikt. Maar dat bleek dus niet te kloppen.” Toch vindt Jansma dat hout nog te weinig als bron wordt gebruikt, en dat heeft veel te maken met de gebrekkige beschikbaarheid van dendrochronologische data. Niet alle instellingen zijn bereid hun data te delen. “Het Rembrandt Research Project heeft veel dendrochronologisch onderzoek laten doen naar de panelen waarop Rembrandt geschilderd heeft. Meestal door het Zentrum Holzwirtschaft van de Universiteit van Hamburg. Maar hun meetresultaten worden niet openbaar gemaakt. Heel jammer.”

Een complicatie bij het begin van haar project was dat er geen universele standaarden waren voor de beschrijving en de documentatie van het jaarringpatroon. Er waren getekende jaarringgrafieken die op kassarolletjes waren vastgelegd en die in laden lagen te verteren. Er was een veelheid aan digitale formaten. Haar eigen metingen uit de jaren tachtig stonden op floppy's van een antieke Apple II-computer. Het computermuseum van de universiteit van Amsterdam slaagde erin ze weer leesbaar te maken.

Standaard

In het grote dataproject dat door NWO wordt gefinancierd maken Jansma en haar medewerkers samen met DANS een doorzoekbaar digitaal depot voor dendrodata uit allerlei landen. Ze werkte met onderzoekers aan de Amerikaanse Cornell universiteit en vele Europese collega's aan een universele standaard voor het vastleggen van de data en aan een standaard-vocabulaire voor het beschrijven ervan. Het is voorzichtig manoeuvreren, vertelt ze. “Ik heb opnieuw gemerkt dat het heel belangrijk is dat je niets opdringt, dat je iedereen laat meepraten over systemen en standaarden, anders lukt het niet.”

Van Zanden wil best verder gaan. “Ik heb meegemaakt dat ik een van de toppers in mijn vak vroeg of ik zijn data kon krijgen. Ik kon ze heel goed gebruiken. Maar hij had ze net weggegooid. Dat deed hij altijd, zei hij. Daar zijn we natuurlijk heel erg tegen. Ik ben er voor dat tijdschriften een artikel alleen maar publiceren als de data voor alle onderzoekers beschikbaar zijn. Ze zouden er een Data Availability Policy op na moeten houden. Dat heeft ook een disciplinerende functie. Het verplicht je te verantwoorden wat je gedaan hebt, dat je transparant met je data omgaat. Zodat je niet een prachtig resultaat krijgt dat door niemand anders kan worden nagedaan.”

Dit is een enigszins ingekorte versie van een op 10 april 2010 in NRC Handelsblad verschenen artikel.

3 ONDERZOEKSDATAFORUM VOOR BETERE TOEGANG EN MANAGEMENT

Door: Wilma Mossink
SURFfoundation

Onder leiding van SURFfoundation is op 19 mei 2009 het Onderzoeksdataforum opgericht. Dit forum streeft naar beter management van en betere toegang tot onderzoeksdata voor de wetenschap en maatschappij.

De samenwerking in het Onderzoeksdataforum voorkomt dubbel werk en het forum kan activiteiten goed op elkaar afstemmen. In deze samenwerking worden de initiatieven binnen diverse organisaties gebundeld. Onder meer DANS, 3TU, de Universiteit van Tilburg, de NCDD (Nationale Coalitie Digitale Duurzaamheid) en SURFfoundation nemen deel.



Onderzoeksdataforum op 18 november 2010

Samen verbeteren

De deelnemende partijen van het Onderzoeksdataforum werken ieder aan betere toegang tot onderzoeksdata. Dit doen zij allen vanuit een andere invalshoek of aandachtsgebied. Het beter toegankelijk maken van onderzoeksdata wil men bereiken door het uitwisselen van kennis en ervaring en het volgen van (inter)nationale ontwikkelingen. Ze willen samen bijdragen aan visie en beleid bijvoorbeeld door het schrijven van handreikingen voor onderzoekers en bibliotheken.

Activiteiten

Het Onderzoeksdataforum concentreert zich op de technische, infrastructurele, juridische en organisatorische aspecten van de opslag en toegang tot onderzoeksdata. Het vergaart kennis over het beheer van onderzoeksgegevens, en creëert bewustzijn voor de waarde van data en goed beheer van deze data.

Het Onderzoeksdataforum organiseert bijeenkomsten om bepaalde aspecten van onderzoeksgegevens breder te bespreken. Ook draagt het Forum bij aan de visie en strategie rondom de verschillende aspecten van onderzoeksdata. Het forum identificeert, initieert en participeert in gemeenschappelijke (Europese) projecten. Het heeft een aantal projecten en studies uitgezet die in de komende hoofdstukken worden besproken.

Data Archiving and Networked Services (DANS)

DANS biedt met het online archiveringssysteem EASY geïnteresseerden toegang tot duizenden databestanden voor wetenschappelijk onderzoek. Bovendien adviseert DANS onderzoekers over het duurzaam archiveren en toegankelijk houden van data, bemiddelt het in het verkrijgen van toegang tot andere databestanden en stimuleert het het toepassen van het door DANS ontwikkelde internationaal erkend data-keurmerk.

DANS is in het Onderzoeksdataforum van SURF vooral actief in de Werkgroep Juridische en Organisatorische Aspecten van Onderzoeksdata. Daarnaast is door DANS, samen met 3TU.Datacentrum, het rapport 'Selection of Research Data' geschreven.

Gedreven door data zorgt DANS er met zijn (inter)nationale projecten en diensten voor dat het digitale wetenschappelijke geheugen van Nederland in stand is en blijft.

DANS is een instituut van de landelijke onderzoeksorganisaties KNAW en NWO.

www.dans.knaw.nl



3TU.Datacentrum

Het plan voor het 3TU.Datacentrum is ontstaan in 2007 en opgesteld op initiatief van de bibliotheken van de drie Nederlandse Technische Universiteiten met als doel het delen van onderzoeksgegevens op technisch-wetenschappelijk gebied te faciliteren.

Het 3TU.Datacentrum bestaat uit drie onderdelen:

- Het data-archief of de data-repository waarin afgeronde datasets beschikbaar zijn gemaakt voor nieuw onderzoek. Wij zorgen voor een gestandaardiseerde en goed gedocumenteerde opslag, zodat de vindbaarheid en de herbruikbaarheid van de data zijn gewaarborgd. Ons digitale archief garandeert de leesbaarheid van de gegevens, zelfs als de gebruikte hardware en software in onbruik is geraakt.
- De data-labs. Deze bieden een al dan niet afgeschermd omgeving en faciliteren het samenwerken aan datasets. Dit is een aantrekkelijke optie in nationale en internationale samenwerkingsverbanden. Immers, alle medewerkers aan het onderzoek kunnen zo anytime anywhere aan de data werken. Wij leveren de digitale gereedschappen om na afloop van het onderzoek definitieve datasets moeiteloos in het data-archief te kunnen opnemen.
- Onder data-services verstaan we het bieden van ondersteuning bij het voorbereiden van data voor opname in het data-archief, het vergroten van de zichtbaarheid en citeerbaarheid van eigen datasets en het benaderen en hergebruiken van technisch-wetenschappelijke datasets die elders in de wereld zijn opgeslagen.

Bij het bedienen van een breed scala aan vakgebieden zijn we voor grote uitdagingen komen te staan. Binnen één discipline bestaan al verschillen in werkwijze, gebruikte tools en cultuur. Tussen disciplines zijn deze verschillen echter nog groter.

Essentieel bij het aangaan van deze uitdagingen is de samenwerking tussen meerdere instellingen vanwege de benodigde capaciteit en competenties en ook de financiering. Een missie zoals die van het 3TU.Datacentrum kan al snel de mogelijkheden van een enkel instituut te boven gaan terwijl juist het lange-termijn aspect een duurzame basis vereist. De samenwerking met wetenschappers is niet alleen een middel om op een efficiënte manier onderzoeksgegevens voor te bereiden op archivering maar ook een ondersteuning van onderzoek. Voor het efficiënt opbouwen van kennis is het ten slotte ook noodzakelijk om afstemming te realiseren met wetenschapsdisciplines buiten de eigen focusgebieden - in het geval van het 3TU.Datacentrum de niet technische wetenschappen, en zelfs andere sectoren zoals overheid en cultureel erfgoed.

Met betrekking tot de technologie is aandacht nodig voor de toegankelijkheid en de lange-termijn bruikbaarheid. Vooral dat laatste vraagt om nieuwe en specifieke kennis die verder gaat dan het maken van back-ups.

www.datacentrum.3tu.nl



Jeroen Rombouts presenteert tijdens onderzoeksdataforum

3TU.Datacentrum

De Nationale Coalitie Digitale Duurzaamheid (NCDD)

De NCDD is een samenwerkingsverband tussen een aantal grote organisaties dat betrokken is bij het langdurig beheer van digitale gegevens. De leden komen uit alle sectoren: de wetenschap, de archieven, de media en het cultureel erfgoed. Aanleiding voor de oprichting van de NCDD in 2008 was het besef dat digitale gegevens traditionele grenzen overschrijden en dat daarom nieuwe afspraken geboden zijn. Ook zijn de investeringen die nodig zijn om digitale gegevens duurzaam toegankelijk te houden zo groot dat samenwerking de enige optie is.

De NCDD stelt zich ten doel om een infrastructuur te bouwen voor duurzame toegankelijkheid tot (geselecteerde) digitale objecten in de publieke sector. Een dergelijke infrastructuur bevat niet alleen opslagfaciliteiten en software, maar ook financiering, onderlinge afspraken,

kennis en opleidingen. Vijf voortrekkersorganisaties zijn door de NCDD aangewezen om binnen hun eigen sector te werken aan deze infrastructuur: de Koninklijke Bibliotheek en Data Archiving and Networked Services (wetenschap), het Nederlands Instituut voor Beeld en Geluid (media), het Nationaal Archief (archieven) en de Culturele Coalitie Digitale Duurzaamheid (cultureel erfgoed).

De NCDD coördineert de activiteiten in de diverse sectoren en dat is ook de achtergrond voor de deelname aan het Onderzoeksdataforum van SURF: de lijnen tussen de diverse sectoren openhouden en zorgen dat kennis gedeeld kan worden over de grenzen van de eigen sector heen.

www.ncdd.nl



Het Parelsnoer Initiatief

Het Parelsnoer Initiatief (PSI) is een samenwerking tussen de acht universitaire medische centra (UMC's). Het initiatief verzamelt op interuniversitair niveau klinische data en biomaterialen. Door het bundelen van deze gegevens en materialen kan de wetenschap zich ontwikkelen, wordt de patiënt beter behandeld en kan nieuwe productontwikkeling plaatsvinden.

PSI richt zich op negen ziektebeelden (parels): beroerte, aangeboren hartafwijkingen, diabetes mellitus, erfelijke darmkanker, inflammatoire darmziekten (IBD), leukemie, neurodegeneratieve ziekten (waaronder Alzheimer), nierfalen en reumatoïde artritis/artrose. In de toekomst zullen de activiteiten van PSI uitgebreid worden naar andere ziektebeelden.

In alle UMC's wordt aan patiënten, die vanwege hun ziekte in aanmerking komen, gevraagd of hun klinische gegevens en lichaamsmateriaal (o.a. bloed, urine, faeces, biopsies) opgenomen mogen worden in deze data- en biobank.

Voor het verzamelen en bewaren van gegevens en lichaamsmaterialen gelden de reguliere medisch-ethische normen.

De data- en biobank wordt gebruikt voor wetenschappelijk onderzoek. Onderzoekers hebben geen directe toegang tot de data- en biobank, maar kunnen in een online catalogus vinden wat en hoeveel er is verzameld binnen een ziektebeeld. Uitgifte gebeurt alleen na medisch-ethische toetsing en na beoordeling van de wetenschappelijke relevantie.

Om ervoor te zorgen dat alle UMC's op dezelfde wijze klinische gegevens verzamelen en registreren, is het Parelsnoer Informatiemodel ontwikkeld. Daarin staan gestandaardiseerde instructies waarmee ICT'ers van de UMC's applicaties kunnen programmeren voor de registratie van gegevens.

www.parelsnoer.org

4 DE EERSTE STAPPEN VOOR ONDERSTEUNING VAN DATAMANAGEMENT...

Door: Jeroen Rombouts
3TU.Datacentrum

Het 3TU.Datacentrum is een van de organisaties in Nederland die zich bezig houden met de archivering en het beschikbaar stellen van data, kortom met de ondersteuning van datamanagement. Dit hoofdstuk poogt een indruk te geven met welke zaken je rekening dient te houden bij het opzetten van een datacentrum. Dit inzicht komt voort uit de lessen die zijn geleerd bij het opzetten van het 3TU.Datacentrum.

Elke data repository rust op drie 'poten'. Dit zijn organisatie (beleid en strategie), techniek (toegankelijk en doorzoekbaar) en resources (financieel en personeel). Met het houvast van deze drie poten kunnen we de eerste voorzichtige stapjes zetten op weg naar goed datamanagement.

Organisatie

Bij het opzetten van een datacentrum is de eerste vraag die zich aandient: waar wil ik naartoe? Is het doel om de gegevens van afgerond onderzoek veilig te stellen en te etaleren? Of is het doel om ondersteuning te bieden op het moment dat het onderzoek nog loopt?

Het datacentrum dat is opgezet door de drie Nederlandse Technische Universiteiten heeft gekozen om zowel lopend onderzoek te ondersteunen, in data-labs, als data uit afgerond onderzoek te bewaren, in een data-archief. Daarnaast is er gekozen om een koppeling tussen de onderdelen te realiseren die het 'publiceren' van data vanuit data-labs naar het data-archief zo eenvoudig mogelijk maakt.

De belangrijkste verschillen tussen het data-lab en het data-archief betreffen de toegankelijkheid en beschikbaarheid van de beheerde onderzoeksgegevens. De gebruikersgroep van een data-lab kan beperkt zijn tot één of enkele onderzoekers en de datasets kunnen vervangen en verwijderd worden door de producenten.



Het data-archief van het 3TU.Datacentrum houdt zich aan de richtlijnen voor data-repositories van het Data Seal of Approval¹ en is dus verantwoordelijk voor de kwaliteit van opslag en beschikbaarheid van de data: data management. De data repository kan dus ook verantwoordelijk gesteld worden voor de beschikbaarheid op lange termijn. Het is echter niet ondenkbaar dat voor de lange termijn opslag archieven zullen ontstaan die zich specialiseren in bepaalde datatypes en bestandstypes. Dit kan noodzakelijk zijn omdat in vergelijking met publicaties er veel meer verschillende, en meer complexe bestandstypes bestaan.

Een nieuw op te richten datacentrum zal ook keuzes moeten maken met betrekking tot de disciplines en data types waarop het zich wil richten.

De antwoorden op deze vragen zullen moeten worden vastgelegd in een datamanagementbeleid en -strategie. Daarnaast is het van belang om samenwerkingsverbanden op te zetten, onder andere om duurzaamheid van investeringen te borgen. Bovendien moet er worden gezorgd voor de acquisitie van datasets, het regelen van toegang tot deze sets, het vaststellen van criteria voor de selectie van data en het plannen van de lange termijn beschikbaarheid. Ook de communicatie en promotie van datamanagement en de mogelijkheden van het data-repository zijn van groot belang.

Techniek

De tweede noodzakelijke poot is de techniek. Kort geleden is voor data-archivering het OAIS (Open Archival Information System) referentiemodel² tot standaard verheven. Het is een uitgebreid model waarin alle functies van een archiefsysteem worden uitgewerkt.

¹ <http://www.datasealofapproval.org/>

² <http://public.ccsds.org/publications/archive/650x0b1.PDF>

Bij het bruikbaar houden van meetgegevens op de lange termijn is het behouden van de originele 'bit stream' (het bestand dat werd ingeleverd) in vrijwel alle gevallen onvoldoende.

- De techniek moet toelaten dat de bestanden over bijvoorbeeld 50 jaar nog geopend en geïnterpreteerd kunnen worden.
- Vanwege de omvang van beschikbare gegevens zal het archief bovendien goed vindbaar en doorzoekbaar moeten zijn.
- En omdat hergebruik het bestaansrecht van de repository³ vormt zal het eenvoudig toegankelijk houden van de gevonden data ook een essentieel onderdeel zijn.

Resources

Om de robuustheid van de opgezette organisatie te garanderen zijn voor de derde poot onder een datacentrum meerdere inkomstenbronnen een noodzaak. Bovendien is voldoende capaciteit nodig voor het opnemen van data in het repository.

Het genereren van inkomsten en beschikbaar hebben van voldoende geschikt personeel is echter niet eenvoudig. Het belang van datamanagement begint pas net door te dringen. Fondsen zijn nog niet voldoende beschikbaar en personeel moet veelal nog opgeleid worden. Ook bij de bestuurders en leveranciers van de data, de dataproducenten, is de bewustwording nog in volle gang en door de beperkte ervaring vraagt dit veel ondersteuning. Waar mogelijk vraagt dit om een nauwe samenwerking tussen het datacentrum en de dataproducent en soms het verrichten van pionierswerk.

Een advies voor de eerste stappen op het gebied van datamanagement ondersteuning: neem eerst de gemakkelijke paden. Werk samen met onderzoeksgroepen die ondersteuning zoeken. Er liggen daar genoeg behoeftes om aan te werken; de uitdaging is met name deze behoeftes te vinden.

Hier volgen enkele tips voor het vinden van geïnteresseerde dataproducenten. Ga op zoek naar:

- groepen met een open cultuur, indicatoren zijn: vooral publiek gefinancierd en niet verwickeld in een publicatiewedloop;
- groepen die onderzoek doen aan langlopende meetreeksen;
- samenwerkingsverbanden waarin behoefte bestaat aan het kunnen delen van gegevens en aan normalisatie van gegevens;
- kostbare meetgegevens, bijvoorbeeld afkomstig uit dure experimenten maar ook waarnemingen die moeilijk, of niet, te herhalen zijn;
- groepen die zichtbaar willen zijn, bijvoorbeeld jonge onderzoeksgroepen die veel gegevens produceren als bijdrage aan een groter onderzoek.

³ Zie de uitkomsten van het project *Waardevolle Data en Diensten*, <http://www.surffoundation.nl/nl/projecten/Pages/Waardevolledataendiensten.aspx>

DataCite

DataCite is een internationaal samenwerkingsverband dat is opgericht eind 2009. Het doel van de organisatie is:



- de toegang tot wetenschappelijke onderzoeks-informatie op internet te vereenvoudigen
- de acceptatie te bevorderen van onderzoeksdata als volwaardige bijdragen in de wetenschappelijke communicatie, meewegend op een wetenschappelijk cv
- de ondersteuning van data-archivering voor verificatie en hergebruik in toekomstig onderzoek.

DataCite maakt zich sterk voor het delen van onderzoeksgegevens. Daarvoor is het noodzakelijk om meer dan voorheen datasets citeerbaar, vindbaar en toegankelijk te maken in publiek toegankelijke data-archieven. Daarnaast levert DataCite inspanningen voor een betere bescherming van onderzoeks-investeringen en zet het zich, samen met data-archieven en uitgevers, in om een infrastructuur op te zetten die alle bovengenoemde doelen helpt te realiseren.

Aan DataCite nemen deel: VS, Canada, VK, Nederland, Denemarken, Zweden, Duitsland, Frankrijk, Zwitserland, Korea en Australië. Vanuit Nederland neemt de TU Delft Library deel aan het initiatief. Andere landen kunnen zich aansluiten.

DataCite biedt een voorziening waarin datasets geregistreerd kunnen worden en van een Digital Object Identifier (DOI) worden voorzien. Deze unieke identificatie-'nummers' worden in wetenschappelijke publicaties gebruikt om naar gepubliceerd onderzoek te verwijzen of dat te citeren. Bovendien maken DOI's het terugvinden van de bijbehorende objecten eenvoudig en betrouwbaar. Inmiddels is aan ruim 30 miljoen wetenschappelijke artikelen een DOI toegewezen en heeft DataCite ruim 800.000 DOI's toegewezen aan datasets. Gebruikers van DataCite DOI's zijn o.a. Dryad, PANGAEA, EURORAD en het Deutsche GeoForschungsZentrum (o.a. de Scientific Drilling Database).

www.datacite.org

5

DATASTEWARDS MOETEN ORDE BRENGEN IN DE CHAOS VAN ONDERZOEKSDATA

Interview met Rob Grim (Universiteit Tilburg)
en Marianne van der Heijden (NIOO-KNAW)
Door: Aad van de Wijngaart

Naarmate de vloedgolf van onderzoeksgegevens stijgt, verschijnen de contouren van een nieuw vakgebied: het beheer van wetenschappelijke onderzoeksgegevens. Twee 'datastewards' vertellen over de uitdagingen waarvoor zij zich gesteld zien.

"Toen Lehman Brothers omviel in 2008, was er onmiddellijk 43 miljard zoek", zo vertelt Rob Grim. "Na drie maanden was nog steeds 23 miljard onvindbaar. Er waren allerlei contracten, maar wie er precies verantwoordelijk was voor de data..."

Bij de Universiteit van Tilburg staat Grim voor de uitdaging om te zorgen dat de onderzoeksresultaten van zijn instelling wél goed vindbaar worden opgeslagen. Hetzelfde geldt voor Marianne van der Heijden van het Nederlands Instituut voor Ecologie (NIOO-KNAW). Zij zijn beiden pioniers in een beroep dat nog geen duidelijk profiel heeft.

"Als ik naar mijn eigen situatie kijk", zegt Van der Heijden, "dan gaat het om iemand die binnen een wetenschappelijk instituut probeert een voortrekkersrol te vervullen in de organisatie rond opslag, archivering en toegankelijk maken van wetenschappelijke onderzoeksdata." Grim: "Het beroep is pionieren en heeft dus niet de bestaande status van een vakreferent of gewone informatiespecialist. Er is echt een functieprofiel bedacht toen ze me aanstelden bij de universiteitsbibliotheek."

Nieuw vakgebied

De behoefte aan datastewards is er niet minder om. Grim: "De huidige wetenschap levert een vloedgolf aan data op; tegelijk hebben we ongekende mogelijkheden om die data te delen. Dat zie je terug in allerlei samenwerkingsomgevingen, die niet meer gebonden zijn aan



Rob Grim

tijd en plaats. Maar de infrastructuur is veel minder ontwikkeld dan die voor publicaties, dus er is nog heel veel te doen om die bronnen toegankelijk te maken."

In de Angelsaksische landen is op dat gebied al wat meer geregeld: een universiteit als Oxford of Berkeley heeft vaak een datacenter, dat zorgt voor licenties op databanken en onderzoekers helpt bij het zoeken naar relevante gegevens.

Maar de zorg voor de onderzoeksgegevens die binnen de eigen instellingen worden gegenereerd, is een ander verhaal. Grim: "Databeheer is echt een nieuw vakgebied aan het worden." Echter wel een vakgebied waarmee het treurig is gesteld, vindt Marianne van der Heijden: "Er is geen brede overtuiging dat het noodzakelijk is om onderzoeksdata goed op te slaan. Er worden regelmatig dikke rapporten over gepubliceerd, maar ik heb niet het idee dat er forse vooruitgang is geboekt sinds ik er zelf in 2007 mee begon."

"We hebben bij ons instituut in Yerseke een inventarisatie laten doen door onze Vlaamse zusterorganisatie", licht ze toe. "Er werden zo'n 300.000 records aangeleverd en daarvan bleek nog geen tien procent bruikbaar. De rest was onleesbaar of onbegrijpelijk. Dat krijg je wanneer je niet aan de bron al begint om onderzoeksdata op een bewuste manier op te slaan."

'Je moet met name sturen op gedrag en bewustzijn van onderzoekers'

Veelkoppig monster

Er moet heel wat gebeuren om de gegevens die bij een onderzoek worden verzameld ook voor anderen bruikbaar te maken. Essentieel zijn metadata, die bijvoorbeeld vertellen waar de gegevens over gaan en hoe ze zijn verzameld. En dan zijn er beleidskeuzes: waar moeten de data worden bewaard? En welke data wil je juist niet openbaar maken? Denk bijvoorbeeld aan onderzoek in opdracht van bedrijven.

Op al deze terreinen, zowel technologische, wetenschappelijke als organisatorische, moet een datasteward resultaten zien te boeken. Dat is een schier onmogelijke opgave. Grim onderscheidt dan ook data librarians en data scientists. De eerste groep verleent data-diensten vanuit de bibliotheek, terwijl de tweede veel meer direct betrokken is bij het werk van de onderzoekers, vaak binnen één of enkele vakgebieden.

Mede gezien de internationale ontwikkelingen blijven de data volgens Grim een veelkoppig monster. “Het is niet waarschijnlijk dat je één model kunt opleggen dat zal voldoen voor alle data en disciplines. De alfa-, bèta- en gamma-wetenschappen zijn zo verschillend. Ik houd me nu bezig met de economische wetenschappen. Daarbinnen heb je al zulke verschillende gebruiken en soorten van data dat je moet oppassen dat je geen verkeerde verwachtingen schept.”

Eerst de AIO's

Dat standpunt wordt bevestigd door Marianne van der Heijden: “Een van onze senior onderzoekers is een groot voorstander van goed databeheer. Maar hij waarschuwde me: je weet van tevoren nooit wat voor data er uit een onderzoek komen, dus je kunt geen al te specifieke richtlijnen geven. Je moet met name sturen op gedrag en bewustzijn van onderzoekers.”

‘We beginnen met de AIO's, zodat die begrijpen dat ze gestructureerd moeten werken’

Dat doet ze dus. “De meeste gevestigde onderzoekers willen verdergaan zoals ze het altijd gedaan hebben. Dat is meteen het pijnpunt. Daarom heeft de directie gezegd dat we gewoon beginnen met de AIO's, zodat die begrijpen dat ze gestructureerd moeten werken. Dichter kan ik niet bij de onderzoeksdata komen, want ik weet niet hoe je er in concreto mee moet omgaan. Daar moet je zelf onderzoeker voor zijn.”

Daarom heeft het NIOO een partner in de arm genomen, het Vlaams Instituut voor de Zee, dat een aantal cursussen verzorgt. Van der Heijden: “Die zijn erg goed ontvangen, met name bij de jonge medewerkers. We proberen nu KNAW-breed een cursus op te zetten, bestemd voor alle Nederlandse AIO's. Die presenteert een aantal basisprincipes voor de omgang met onderzoeksdata.”



Marianne van der Heijden

Maar voor een breed en blijvend succes zal meer nodig zijn, denkt Van der Heijden. “Het bewustzijn kan versterkt worden door meer bekendheid te geven aan de noodzaak van goed databeheer, maar ook door verplichting vanuit een subsidiegever of een werkgever.

De cursus die we gaan doen ontleen we aan de Universiteit van Minneapolis: daar zijn alle onderzoekers verplicht zo'n datacursus te volgen. Als ze dat niet binnen een bepaalde termijn doen, wordt hun onderzoeksopdracht ingetrokken. Dat is een vrij sterke motivator.”

Meer informatie

- Voor het vakblad InformatieProfessional (nummer 11/12/2010) heeft Rob Grim geïnventariseerd wat datastewards doen, en wat ze daarvoor moeten weten en kunnen. Lees het artikel hier: http://epub01.publitas.nl/ottocramwinkeluitgeverij/informatie_professional_11-12_2010/magazine.php?spread=36
- Het Data and Information Portal van NIOO-KNAW (<http://data.nioo.knaw.nl>) geeft een indruk van wat er mogelijk is met gestructureerd databeheer.

6 VERKORTE JURIDISCHE WEGWIJZER VOOR RUWE DATA: DE STOPLICHT-ROUTE

Gebaseerd op een studie door Madeleine de Cock Buning, Allard Ringnalda, Tina van der Linden
Centrum voor Intellectueel Eigendoms Recht, Universiteit Utrecht

De mogelijkheden van ICT en Internet maken de toegankelijkheid van onderzoeksresultaten en -data steeds groter. De mogelijkheden groeien om aanvullend aan tekst, publicaties te verrijken met de onderzoeksdata. Door deze toegankelijk te maken, wordt het gemakkelijker om de resultaten van onderzoek te verifiëren, te reproduceren en te hergebruiken. Maar met hergebruik van de (ruwe) onderzoeksdata is het belangrijk te weten wat de juridische status is van het materiaal. Soms is toestemming van de maker vereist maar sommige handelingen kunnen ook zonder toestemming worden verricht.

Om dit te achterhalen heeft SURFshare aan CIER (Centrum voor Intellectueel Eigendomsrecht) gevraagd de stand van zaken aan de hand van de belangrijkste wetgeving en rechtspraak met betrekking tot datasets in kaart te brengen. Om als onderzoeker snel te kunnen bepalen welke toestemming nodig is om onderzoeksgegevens van anderen te kunnen hergebruiken is er een verkorte versie hieruit afgeleid. Deze is bedoeld als hulpmiddel aan de hand waarvan men snel de weg naar de toelichtingen in de meer uitgebreide, juridische wegwijzer¹ kan vinden.

¹ <http://www.surfoundation.nl/nl/publicaties/Pages/Juridischestatusvanruwedata.aspx>

• Wilt u een kopie maken voor eigen gebruik?



Dan hoeft u geen toestemming te vragen.

• Wilt u de data inlezen in een eigen wetenschappelijke databank, zonder dat de data met anderen dan de eigen onderzoekers worden gedeeld?



Dan hoeft u geen toestemming te vragen.

Toelichting: Zelfs als de data door het auteurs- of databankenrecht beschermd zijn, valt dit soort gebruik onder de toegestane uitzonderingen: het maken van een kopie voor eigen gebruik (paragraaf 2.4.5) van een auteursrechtelijk beschermd werk, of het opvragen van substantiële delen van een databank in het kader van wetenschappelijk onderzoek (paragraaf 3.2.5); voor het opvragen van niet-substantiële delen is hoe dan ook geen toestemming vereist (paragraaf 3.2.3).

• Wilt u de over te nemen data (voor een deel) publiceren of met anderen dan de eigen onderzoekers delen?



Dan moet u waarschijnlijk wel toestemming vragen.

>> Controleer of het auteursrecht van toepassing is:

• Heeft de onderzoeksdata een eigen, oorspronkelijk karakter en het persoonlijk stempel van de maker?

(tip: denk hierbij of het denkbaar is dat twee makers, afzonderlijk van elkaar, tot precies dezelfde vormgeving hadden kunnen komen)



Dan is het auteursrecht van toepassing en moet u toestemming vragen aan de maker(s).

Toelichting: Bepalend is aldus of de maker een keuzeruimte had bij het vormgeven aan de data, en of hij deze (bewust of onbewust) op basis van persoonlijke voorkeuren heeft ingevuld (paragraaf 2.3.2). Ook een selectie van louter blote feiten kan auteursrechtelijk beschermd zijn (paragraaf 2.3.3).

>> Controleer of het databankenrecht van toepassing is:

- Is er sprake van een verzameling zelfstandige elementen (onderzoeksdata)?
 - Is de dataverzameling systematisch geordend?
 - Is er substantieel geïnvesteerd in de verkrijging, controle of presentatie van de elementen?
- Hierbij telt de investering in het doen van het onderzoek waaruit de gegevens voortvloeien niet mee.



Dan is het databankenrecht van toepassing en moet u toestemming vragen aan de maker(s).

Toelichting: bovenstaand wordt verder uitgewerkt in de paragrafen 3.1.1, 3.1.2 en 3.1.3.

- Gaat u de overgenomen beschermde data op een of andere wijze openbaar maken, bijvoorbeeld door ze te publiceren of op andere wijze aan anderen ter beschikking te stellen?



Dan is het auteursrecht van toepassing en moet u toestemming vragen aan de maker(s).

Toelichting: De wet spreekt in deze gevallen van openbaarmaking (paragraaf 2.4.2). Data van anderen mogen uitsluitend voor eigen onderzoekgebruik wel worden gekopieerd of overgenomen zonder toestemming van de maker, mits geen openbaarmaking plaatsvindt (paragraaf 2.4.2). Voor het voorziene gebruik is toestemming van de maker vereist (paragraaf 4.3).

- Is er sprake van: het citaatrecht, werken van de overheid, de integrale vorm, een selectie blote feiten of andersoortige data (foto's e.d.)?



Dan hoeft u te maken met een uitzondering op de auteurswet en hoeft u geen toestemming te vragen.

Toelichting: bovenstaande uitzonderingen worden beschreven in de paragrafen 2.4 en 2.1.2, 2.3.3 en 2.3.4.

- Gaat u gegevens uit de databank niet alleen voor eigen onderzoek overnemen, maar ook verder openbaar maken, bijvoorbeeld in een publicatie of deelt u het eigen materiaal (met daarin de overgenomen data) met anderen?



Dan is het databankenrecht van toepassing en moet u toestemming vragen aan de maker(s).

Toelichting: De wet spreekt in dergelijke gevallen van hergebruik.

- Gaat u een substantieel deel van de databank hergebruiken?



Duidelijke criteria voor wat substantieel is, ontbreken. Er moet zowel gekeken worden naar de relatie tussen het hergebruikte deel en de gehele databank, als de technische en economische waarde van het hergebruikte deel.

Toelichting: zie paragraaf 3.2.3; zie ook 3.2.8.

- Maakt u gebruik van een databank van de overheid?



Dan hoeft u geen toestemming te vragen.

Toelichting: Op databanken van de overheid rust in beginsel geen databankenrecht, tenzij anders bepaald (paragraaf 3.2.6). Voor het opvragen van substantiële delen van een databank in het kader van wetenschappelijk onderzoek is geen toestemming van de rechthebbende op de databank vereist (paragraaf 3.2.5)

- Gaat u een geschrift dat gepubliceerd is of voor publicatie is bestemd, maar dat niet door het auteursrecht of het databankenrecht wordt beschermd, integraal overnemen en openbaar maken?



Dat is alleen toegestaan als u het in gewijzigde vorm publiceert.

Toelichting: zie hoofdstuk 4.

7 DATA IN AUSTRALIË EN NEDERLAND: EEN ZAAK VAN LANGE ADEM

Interview met Andrew Treloar (ANDS) en Jeroen Rombouts (3TU.Datacentrum)
Door: Lilian van der Vaart

Andrew Treloar en Jeroen Rombouts moeten mannen van lange adem zijn. Treloar heeft in het verleden Germaanse talen gestudeerd en spreekt nog altijd 'een beetje Nederlands'. Sinds midden jaren negentig houdt hij zich bezig met elektronisch publiceren, het ontwerpen van de technische architectuur voor grote repository-projecten, virtuele onderzoeksomgevingen en dataprojecten. Sinds begin 2009, toen de ANDS (de Australian National Data Service) werd opgericht, is hij Director of Technology bij deze organisatie, en hij hoopt de visie voor de ANDS voor 2018 te kunnen realiseren. Hij weet alles van groei, want hij kweekt groenten en fruit en houdt kippen. Rombouts is ooit begonnen in het industrieel- en milieuonderzoek, is vervolgens overgestapt naar het managen van innovatieprojecten bij de TU Delft Library en is nu directeur van het 3TU.Datacentrum. Geen zee is hem te hoog – hij zeilt en hij duikt.

De situatie in Australië

Het hoge subsidiebedrag en alle aandacht voor databeheer zijn benijdenswaardig, beaamt Treloar. De laatste twee regeringen hebben zich hard gemaakt voor de totstandkoming van een uitstekende onderzoeksinfrastructuur, overigens deels naar aanleiding van de crisis, om Australische onderzoekers internationaal tot aantrekkelijke partners te maken. Er wordt veel geld gestoken in de ontwikkeling van instrumenten en informatie-infrastructuur en inmiddels is Australië het land met de grootste investeringen in databeheer ter wereld. Daarbij heeft ANDS het geluk te kunnen samenwerken met ambtenaren die zijn doordrongen van het belang van databeheer en -opslag. Er wordt bewust ingezet op het nationale niveau, omdat dat betere garanties biedt voor de toekomst dan wanneer per discipline wordt gewerkt.

Verschillen in datalandschappen

Als meest opvallende verschil tussen het Australische en het Nederlandse databeheer noemt Rombouts – even afgezien van de hoogte van de subsidies – de Nederlandse 'scheidslijn' tussen de alfa- en gammawetenschappen enerzijds (DANS) en de bètawetenschappen anderzijds (3TU.Datacentrum). Hoewel deze situatie vanuit historisch oogpunt te begrijpen is, wordt toch gehoopt op meer samenwerking in de toekomst. Ondanks de verschillen in schaalgrootte vinden Rombouts en Treloar dat het in Nederland net zo moeilijk is om het hele datalandschap helder in beeld te krijgen als in Australië. Er gebeurt – vooral in de exacte wetenschappen – van alles op lokaal niveau, bij afzonderlijke instellingen of onderzoeksgroepen die eigen 'infrastructuur' aanschaffen.



Andrew Treloar

Wat ook grotendeels overeenkomt, is het verschil in houding en gedrag ten opzichte van databeheer tussen de diverse disciplines. Treloar wijst er op dat het belang van data als zodanig steeds meer erkend wordt. Onderzoeksfinanciers beginnen de publicatie van data te zien als onderzoeksoutput van gefinancierde projecten. Ook is er een toenemende trend naar open overheidsdata. Hij heeft het idee dat Australië in dat opzicht een inhaalslag aan het maken is ten opzichte van Europa. De overheid erkent het belang van hoogwaardig databeheer nu vermoedelijk voldoende om de subsidie in de een of andere vorm voort te zetten, en Treloar hoopt dat universiteiten de voordelen zelf zullen ondervinden en zullen doorgaan. Rombouts twijfelt of men in Nederland al zo ver is. Een samenwerkingsverband als DataCite, waarvan de ANDS en TU Delft Library allebei medeoprichters zijn, zal helpen om de zichtbaarheid van data te verbeteren en het idee te laten doordringen dat data 'er echt toe doen'. Onderzoekers moeten erkenning krijgen voor hun data als zodanig en niet alleen voor het onderzoek waarvoor deze zijn gebruikt.

Opheffen van de scheidslijn

Het opheffen van de scheidslijn tussen disciplines heeft voordelen, bijvoorbeeld omdat daarmee multi- en interdisciplinair onderzoek kan worden gestimuleerd. Treloar en Rombouts zijn het erover eens dat het aanpakken van databeheer organisatorische uitdagingen met zich meebrengt, of de aanpak nu nationaal, institutioneel of (inter)disciplinair is.

Feiten over de ANDS

- Databeheer op nationaal niveau
- Gericht op: universiteiten en onderzoeksinstellingen
- Budget: 122 miljoen Australische dollars voor databeheer, waarvan 72 miljoen voor de ANDS en 50 miljoen voor dataopslag (bij een afzonderlijk centrum)
- Zeven programma's ten behoeve van databeheerbeleid en plannen en infrastructuurprojecten op nationaal niveau, zoals software voor het vastleggen van data en metadata in alfa-, bèta- en gammawetenschappen
- Veel werkzaamheden worden uitgevoerd door onderzoeksinstellingen die volledig door de ANDS worden gefinancierd

www.ands.org.au

Treloar denkt dat het beter is om contact met de onderzoekers te organiseren bij de instellingen. Nauw overleg tussen onderzoekers en metadataspecialisten is namelijk een voorwaarde om tot meer begrip te komen, en fysieke nabijheid helpt daarbij. Een nadeel is wel dat bij iedere instelling allerlei verschillende bestandsformaten moeten worden ondersteund.

Rombouts wijst erop dat onderzoekers ook vragen om andere soorten ondersteuning, zoals hulp bij het verzamelen van data of het verbeteren van onderzoeksmethoden, of met betrekking tot presentatie (visualisering en zichtbaarheid). Wellicht is daarvoor een nieuw soort professional nodig, die over een combinatie van informatiebeheer/IT-vaardigheden en sociale vaardigheden beschikt. Maar zulke mensen zijn nu nog moeilijk te vinden.

Toepassing van het datacuratie-continuüm-model

“O ja, het model...” Als hem wordt gevraagd naar het model dat hij de afgelopen jaren heeft ontwikkeld (zie figuur 3 in hoofdstuk 1), toont Treloar zich vooral verheugd over het feit dat het vrij goed schijnt aan te sluiten bij wat mensen doen. Hij benadrukt dat onderzoeksdata niet per se alle drie de fasen hoeft te doorlopen: het is een model, geen voorschrift. Het onderzoek dat door de ANDS wordt gesubsidieerd, wordt grotendeels uitgevoerd in de persoonlijke werkruimte en voor een klein deel in het openbare domein. Er moet meer werk worden gedaan aan de gezamenlijke werkruimte. Onderzoekers willen dat ook. Maar pas als er daar veel betere en eenvoudiger tools beschikbaar zijn, zullen deze de voorkeur krijgen boven de huidige, vertrouwde, maar verre van robuuste tools. “Er zijn mensen die Facebook gebruiken voor hun data”.

‘... er zijn mensen die Facebook gebruiken voor hun data’

Er zijn in dat verband twee aandachtsgebieden: de software en de dienstverlening. De gedeelde werkruimte is misschien het moeilijkste domein om te faciliteren: mogelijke problemen bij onderzoekssamenwerking zijn immers internationaal van aard, maar oplossingen moeten op lokaal of nationaal niveau worden gevonden. Rombouts sluit zich daarbij aan. Het dilemma voor IT-managers bij instellingen is vast te stellen welke oplossingen alleen door de instelling kunnen worden geboden en welke problemen beter buitenshuis kunnen worden opgelost, gezien de financiële, juridische en veiligheidsaspecten waar de instelling mee te maken heeft. Op dit terrein is er behoefte aan samenwerking tussen instellingen, en zelfs met het bedrijfsleven; er is een tendens om oplossingen (en financiering) op Europees niveau te gaan zoeken.

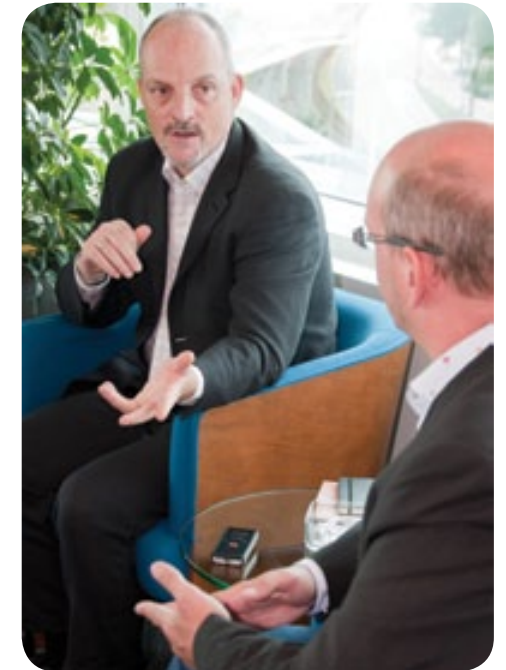
Wat het model betreft, zegt Rombouts, werkt het 3TU.Datacentrum voornamelijk in het openbare domein en een beetje aan de gedeelde werkruimte. Op een recente rondetafelbijeenkomst werd geconclu-

deerd dat steun voor de persoonlijke werkruimte wenselijk is. De vraag is dan hoe je daar, gezien de relatief grote afstand in een gecentraliseerde organisatie, in een vroeg stadium bij betrokken kunt worden. Treloar citeert een collega die heeft gezegd: “Ik wil werken met mensen die ‘datapijn’ hebben.” Bij de ANDS werken ze met capabele en gemotiveerde mensen, en daar hebben ze hun handen al aan vol...

‘Ik wil werken met mensen die ‘datapijn’ hebben’

Nog een laatste boodschap?

“Al die willen data ontsluiten, moeten mannen van lange adem zijn...” Voor databasebeheer moet je een doorzetter zijn. Treloar: “Het kost tijd om met onderzoekers aan de slag te gaan, en je moet goed omgaan met hun verwachtingen.” Rombouts vult aan: “De aard van de problemen waar ze mee kampen, is de afgelopen tien jaar niet veranderd.” Het vereist het geduld van een tuinier om dingen tot bloei te brengen, en het uithoudingsvermogen van een duiker om langdurige problemen op te lossen...



Treloar (links) en Rombouts

Wat Treloar benijdenswaardig vindt aan Nederland:

- klein land maar zeer gefocust, wat grote voordelen heeft; in Australië krijg je niet zo gemakkelijk iedereen rond de tafel...
- SURF, DANS en 3TU.Datacentrum hebben een aantal geweldige mogelijkheden; het equivalent daarvan ontbreekt in Australië

en wat hij ons adviseert:

- doorgaan met het idee om een nationale, interdisciplinaire laag te creëren; de onderverdeling per discipline is niet altijd handig; er is tegenwoordig tenslotte ook meer interdisciplinair onderzoek.

Wat Rombouts bevalt in Australië:

- de schaalgrootte: de ideeën, het grote aantal betrokkenen
- steun aan de persoonlijke werkruimte

en wat hij hun zou adviseren:

- financiering en ondersteuning geven aan samenwerkingsverbanden, omdat dit goede ambassadeurs zijn.

8

CHECKLIST HELPT BIJ SELECTIE TE BEWAREN ONDERZOEKSDATA

Door: Annemiek van der Kuil
SURFfoundation

De hoeveelheid onderzoeksdata groeit continu. Daarom is het belangrijk te weten hoe, waar en hoe lang onderzoekers en instellingen onderzoeksdata willen bewaren. En minstens zo belangrijk is om te weten welke data de moeite van het bewaren waard zijn. Om onderzoekers hierbij te helpen zijn algemene richtlijnen opgesteld voor het selecteren van informatie.

Deze richtlijnen zijn gebaseerd op drie studies die SURFfoundation heeft laten uitvoeren door DANS samen met 3TU. Datacentrum¹, kunstgeschiedenis & mediastudies aan de Universiteit Leiden² en het Internationaal Instituut voor Sociale Geschiedenis (IISG) van de KNAW³. Vanuit drie wetenschappelijke disciplines is gekeken hoe onderzoekers met onderzoeksdata omgaan, hoe ze data opslaan en hoe men die onderzoeksdata met elkaar deelt. Dat alles blijkt op verschillende manieren ervaren te worden, niet alleen tussen disciplines, maar ook binnen vakgroepen. Maar ondanks die verschillen hebben de onderzoekers algemene uitgangspunten kunnen formuleren over welke onderzoeksgegevens bewaard moeten worden, hetzij om de resultaten te valideren, hetzij voor hergebruik.

Deze uitgangspunten zijn verzameld in een checklist: 'Algemene richtlijnen voor het selecteren van onderzoeksdata om te bewaren'⁴. Deze bevat de belangrijkste redenen voor het opslaan van onderzoeksgegevens voor de lange termijn. De checklist kan gebruikt worden door individuele onderzoekers, onderzoeksgroepen, instellingen en door financiers. Maar ook door beheerders van archieven, onderzoeksrepositories en erfgoedinstellingen.

De opvattingen over het bewaren van onderzoeksdata lopen breed uiteen, van: 'bewaar alles want je weet niet wat je er ooit nog mee kan doen' tot 'specificeer wat je wel en wat je niet bewaart'. Alle drie de studies concluderen dat het belangrijk is om tijdig na te denken over het bewaren van de onderzoeksdata.



Algemene richtlijnen voor het selecteren van onderzoeksdata om te bewaren

Hoewel onderzoeksgegevens veelal bewaard moeten worden voor (her)gebruik of om onderzoeksresultaten te valideren, geldt dit niet voor *alle* gegevens. Om te bepalen welke onderzoeksgegevens waardevol zijn als (bron)materiaal voor onderzoek zijn een aantal algemene richtlijnen opgesteld in de vorm van een checklist. De checklist bevat de belangrijkste redenen voor het opslaan van onderzoeksgegevens voor de lange termijn. De checklist kan gebruikt worden door individuele onderzoekers, onderzoeksgroepen, instellingen en door financiers. Maar ook door beheerders van archieven, onderzoeksrepositories en erfgoedinstellingen.

Redenen om onderzoeksgegevens te bewaren

1. Is het **verplicht** om onderzoeksgegevens te bewaren voor **(her)gebruik**? Voor hoelang? De onderzoeksfinancier (bijv. NWO), een wetenschappelijke uitgever of anderen kunnen deze eis stellen.

Als er geen verplichting is, is er dan toch een goede reden om de onderzoeksgegevens voor

(her)gebruik ten behoeve van onderzoek te bewaren? Voor de onderzoeksdiscipline waar de gegevens gecreëerd zijn of voor andere wetenschapsdisciplines? Deze redenen kunnen zijn:

- Belang van de gegevens: potentiële waarde voor hergebruik, (inter)nationale positionering en kwaliteit, oorspronkelijkheid, omvang, schaal, de productiekosten van de gegevens of het innovatieve karakter van het onderzoek.
- Uniceit van de gegevens: de gegevens omvatten niet-herhaalbare waarnemingen.
- Belang van de gegevens voor historisch onderzoek, in het bijzonder wetenschapshistorisch onderzoek.

2. Is het **verplicht** om de onderzoeksgegevens te bewaren voor **controle**? Dit kan onderdeel zijn van een gedragscode voor onderzoekers waarbij ruwe gegevens bijv. vijf jaar moeten worden bewaard.

3. Zijn er redenen om de onderzoeksgegevens voor **algemene (in principe niet-wetenschappelijke)** doeleinden te bewaren? Zijn de gegevens van belang voor cultureel erfgoed, musea of presentaties?

Randvoorwaarden voor het selecteren

In alle gevallen geldt dat een van hier genoemde redenen doorslaggevend moet zijn bij het nemen van de uiteindelijke selectiebeslissing. Voor of tijdens het selectieproces moet men ook rekening houden met randvoorwaarden. Over al deze punten moet duidelijkheid bestaan voordat een definitief selectiebesluit wordt genomen.

- Technisch: welke data formats, software (standaard of op maat gemaakt: onderzoeksspecifieke tools), hardware?
- Metagegevens: beschikbaar en voldoende? Technische informatie, codeboeken, informatie over de structuur van de gegevens, contextuele informatie, informatie over intellectuele eigendomsrechten, links naar publicaties of gerelateerde gegevens.
- Gegevens: uit welk moment van de digitale cyclus: ruwe gegevens, halfbewerkte gegevens, gepubliceerde gegevens?
- Duidelijkheid met betrekking tot intellectuele eigendomsrechten, zoals auteurs- en octrooirechten en/of databankrechten, bescherming van persoonsgegevens?
- Infrastructuur: is deze beschikbaar voor het bewaren van de gegevens? Ofwel een gegevensarchief, ofwel een institutionele of thematische repository.
- Kosten: is in de kosten voorzien van het selecteren, converteren, langdurig bewaren en beschikbaar stellen van de gegevens?

Meer weten?

Rapport: 'Selection of Research Data; Guidelines for appraising and selecting research data'
www.surfoundation.nl > Publicaties

1 <http://www.surfoundation.nl/nl/publicaties/Pages/StudieSelectionofResearchData.aspx>

2 <http://www.surfoundation.nl/nl/publicaties/Pages/StudieDataCurationinArtsandMediaResearch.aspx>

3 <http://www.surfoundation.nl/nl/publicaties/Pages/StudieIISHGuidelinesforpreservingresearchdata.aspx>

4 http://www.surfoundation.nl/nl/themas/openonderzoek/cris/Documents/SHAREflyer_selectionofdata_defversie_def_NL.pdf

9

ORGANISATIE VAN OPSLAG EN BESCHIKBAARSTELLING ONDERZOEKSDATA

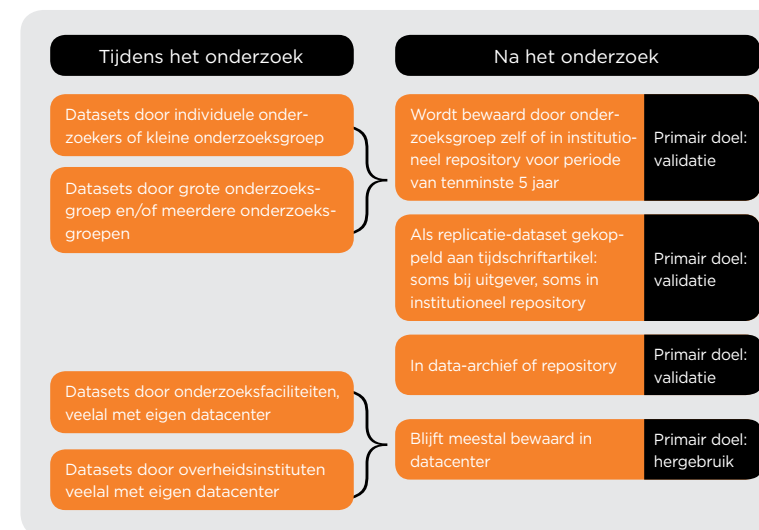
Door: Maurits van der Graaf
Pleiade Management en Consultancy

Volgens de Gedragscode Wetenschapsbeoefening van de VSNU¹ dienen onderzoeksdata tenminste 5 jaar na publicatie bewaard te worden met het oog op validatie. Daarnaast vereisen wetenschappelijke tijdschriften in toenemende mate publicatie van de onderliggende datasets bij hun artikelen, eveneens met het oog op validatie. Beschikbaarstelling van onderzoeksdatasets maakt tevens nieuwe vormen van onderzoek mogelijk zoals meta-analyses waarbij meerdere datasets worden samengevoegd en opnieuw geanalyseerd. Deze nieuwe vormen van wetenschapsbeoefening worden wel data-exploratie of het 'Fourth Paradigm of Science' genoemd. Kortom, er zijn belangrijke redenen om onderzoekdatasets duurzaam op te slaan en beschikbaar te stellen. SURF heeft dit onderkend en een studie laten verrichten naar de 'Organisatorische aspecten van duurzame opslag en beschikbaarstelling onderzoeksdata'².

Huidig landschap van organisatievormen

Figuur 1 geeft een overzicht van de huidige organisatievormen voor opslag en beschikbaarstelling van onderzoeksdatasets.

Op een aantal vakgebieden zijn grote onderzoeksfaciliteiten en/of onderzoeksinstituten met bijbehorende datacentra actief. Zo kunnen astronomen waarnemingstijd aanvragen bij de radiotelescoop LOFAR. Gedurende een periode van een jaar kunnen zij exclusief over de waarnemingsgegevens beschikken, waarna de dataset beschikbaar wordt gesteld voor alle astronomen. Meerdere vakgebieden beschikken over vergelijkbare onderzoeksfaciliteiten. De datacentra hiervan (en van vergelijkbare overheidsinstituten zoals



Figuur 1: Organisatievormen voor opslag en beschikbaarstelling van onderzoeksdatasets

het KNMI) bevatten de onderzoeksdata vanaf de creatie. Het delen van datasets is veelal vanaf een vroeg stadium goed geregeld, maar blijft beperkt tot de 'eigen community'.

Voor onderzoekers, die niet aangesloten zijn bij een dergelijk data-centrum van een grote onderzoeksfaciliteit zijn er drie mogelijkheden om hun data op te slaan:

- **Zelf-organisatie:** Onderzoekers organiseren in veel gevallen zelf de opslag en beschikbaarstelling van hun onderzoeksdatasets. Deze worden dan bijvoorbeeld intern opgeslagen of zijn via de website van de onderzoeksgroep beschikbaar. Wijdere beschikbaarstelling en een langdurige bewaartermijn zijn in deze situaties meestal niet gegarandeerd.
- **Data-archieven:** Data-archieven - meestal gericht op een of meerdere disciplines - zijn gericht op een wijde beschikbaarstelling en langdurige bewaartermijn. Onderzoekers in hun rol van dataproducent dragen na afloop van hun onderzoek de onderzoeksdatasets over. Bekende voorbeelden in Nederland zijn het data-archief van DANS voor sociale en geesteswetenschappen en het data-archief van het Max Planck instituut voor taalwetenschappers.
- **Supplementary data:** Een toenemend aantal wetenschappelijke tijdschriftredacties vereist dat de onderliggende datasets bij een artikel (replicatiedatasets) Open Access ter beschikking worden gesteld. Tijdschriften bieden daar zelf mogelijkheden toe (maar vaak met beperkingen qua omvang en qua dataformats), maar de meeste uitgevers zien geen rol voor zichzelf weggelegd in de opslag en beschikbaarstelling van onderzoeksdatasets.

¹ <http://www.vsnul.nl/Media-item/Nederlandse-Gedragscode-Wetenschapsbeoefening.htm>

² Het volledige onderzoek is na te lezen op: <http://www.surfoundation.nl/nl/publicaties/Pages/Organisatorischeaspectenduurzameopslagenbeschikbaarstellingonderzoeksdata.aspx>

Aanbevelingen

Rol voor onderzoeksinstellingen

Er is een belangrijke rol voor onderzoeksinstellingen op dit gebied weggelegd. Er heerst onduidelijkheid over de verantwoordelijkheden en het eigenaarschap van onderzoeksdatasets, die binnen een instelling gecreëerd worden. Ook het datamanagement laat in sommige gevallen te wensen over. Tenslotte zijn er op veel vakgebieden geen mogelijkheden voor duurzame opslag en beschikbaarstelling van datasets. Daarom wordt een samenhangend pakket maatregelen aanbevolen, met onder andere voorstellen voor het vaststellen van een institutioneel beleid, verbeteringen op het gebied van datamanagement (inclusief scholing voor onderzoekers), registratie van onderzoeksdatasets en de mogelijkheid van het opzetten van een institutioneel datarepository.

‘Beschikbaarstelling van onderzoeksdatasets maakt nieuwe vormen van onderzoek mogelijk’

Naar een nationale data-infrastructuur

In meerdere landen worden er acties ondernomen om de bestaande infrastructuur uit te bouwen tot een dekkend en interoperabel netwerk van datacentra, data-archieven en institutionele datarepositories. Daarbij wordt een taakverdeling voorzien tussen onderzoeksinstellingen, die de registratie en de ‘eerstelijns zorg’ van de datasets op zich nemen en een netwerk van veelal disciplinegeoriënteerde data-archieven en datacentra, nationaal en internationaal, voor de lange termijn archivering en beschikbaarstelling van geselecteerde onderzoeksdatasets. Om dit ook in Nederland te bereiken bevat het rapport een aantal aanbevelingen voor het nationale niveau. De belangrijkste aanbevelingen zijn dat de Gedragscode Wetenschapsbeoefening van de VSNU kan worden aangescherpt. In deze gedragscode wordt de verantwoordelijkheid voor, het eigenaarschap van en de toegang tot onderzoeksdatasets geregeld. Een tweede aanbeveling is dat DANS een nationale rol zou kunnen spelen als expertisecentrum voor alle disciplines.



SURFproject PODIUM PLUS

Het SURFshare project Podium Plus levert aanbevelingen voor het opzetten van een efficiënte landelijke structuur voor opslag van en toegang tot onderzoeksdata. Voor veel instellingen is dataopslag een nieuwe dienst en kostbaar om te realiseren. Met Podium Plus wordt onderzocht of instellingen in een landelijke structuur kunnen en willen samenwerken. Als blijkt dat dit het geval is, kan een ‘Dutch Dataverse network’ worden opgezet. In het project is een koppeling gezocht met DANS en 3TU.Datacentrum om langdurige opslag te garanderen.



10 KWALITEIT VAN ONDERZOEKSDATA: WAT ONDERZOEKERS WILLEN

Door: Maurits van der Graaf en Leo Waaijers

Onderzoeksdatasets worden in toenemende mate ter beschikking gesteld en gebruikt als autonome bron van informatie voor nieuw onderzoek. Dat heet de 'Fourth Paradigm of Science'. Niet alleen moeten onderzoeksdatasets daarom, net als publicaties, vindbaar en toegankelijk zijn, ook een toets op kwaliteit wordt dan steeds belangrijker.

Bij publicaties wordt de kwaliteit getoetst door middel van peer review en citatie indexen. Maar hoe zou dit voor onderzoeksdata kunnen gebeuren? En wat willen de onderzoekers? Om antwoord te krijgen op deze vragen liet SURFfoundation een onderzoek¹ uitvoeren met als methoden: een literatuurstudie, een serie interviews met sleutelfiguren in de wereld van onderzoeksdata en een enquête onder een representatief aantal Nederlandse hoogleraren en UHD's. Uit de literatuurstudie en de interviews blijkt dat de kwaliteit van onderzoeksdatasets de volgende drie aspecten heeft:

- Kwaliteitszorg bij het creëren van de data: het gaat dan om nauwkeurigheid van de data (afstelling van de meetapparatuur en dergelijke) en om de methodische correctheid van de data verzameling;
- Datamanagement: het gaat dan onder andere om toegankelijkheid (is de dataset goed beschreven door middel van metadata en documentatie?) en duurzaamheid (is het format van de data in de toekomst nog leesbaar?);
- De wetenschappelijke waarde ('the scholarly merit'): het gaat dan vooral om de vraag of de dataset betekenis kan hebben voor andere wetenschappers in verband met toekomstig onderzoek.

¹ www.surfoundation.nl/nl/publicaties/Pages/Verkenndonderzoek.aspx

Toetsen van de kwaliteit van onderzoeksdata

In het onderzoek is bekeken hoe de kwaliteit van onderzoeksdata getoetst en eventueel verbeterd zou kunnen worden. Ook is onderzocht hoe de beschikbaarstelling van datasets kan worden gestimuleerd.

Een eerste mogelijkheid is kwaliteitstoetsen vooraf. In dat geval is het mogelijk om de onderliggende dataset mee te nemen in het peer review van het tijdschriftartikel. In de praktijk blijkt dit momenteel echter slechts in zeer beperkte mate te gebeuren: het peer review systeem is sowieso al overbelast en peer reviewers hebben meestal geen tijd en missen soms de middelen om ook de dataset bij het artikel nog eens te beoordelen. Een oplossing zou kunnen zijn datgene waarmee het tijdschrift Earth System Science Data twee jaar geleden is begonnen: de doelstelling van dit tijdschrift is om de intellectuele inspanningen van dataproducenten en de inspanningen om deze datasets voor anderen ter beschikking te stellen tot uiting te laten komen in het 'academic record'. De eerste ervaringen met het tijdschrift zijn positief volgens de geïnterviewde hoofdredacteur: het tijdschrift blijkt in een behoefte te voorzien en bevordert het publiceren van datasets.



Presentatie door Gerben de Boer tijdens colloquium over kwaliteit van data

'Datasets verdienen een plaats in het Academic Record'

Een tweede mogelijkheid is kwaliteitstoetsen achteraf. Dit kan op verschillende manieren. Eén optie is om hergebruikers van datasets commentaren over de kwaliteit aan de dataset te laten toevoegen. In de praktijk gebeurt dit

nog niet, al is er wel een aantal aanzetten om dit te gaan uitproberen. Een andere optie is het mogelijk maken van citeren van datasets, zodat hergebruikers die een publicatie schrijven, de door hen gebruikte datasets kunnen citeren in de literatuurlijst. Veel geciteerde datasets zijn dan (waarschijnlijk) van hoge kwaliteit en bovendien kunnen de dataproducerende onderzoekers deze citaties meetellen bij hun citatiescores. In de praktijk zijn er al enkele initiatieven om dit te realiseren, maar een algemeen geaccepteerde en bekende standaard voor het citeren van een dataset bestaat nog niet.

Om kwaliteitszorg van datasets in een vroeg stadium te ondersteunen, worden in een aantal universiteiten al datamanagement-cursussen gegeven. Onderzoekers zijn in het algemeen niet onderwezen in datamanagement, dit is immers een relatief nieuw domein. Daarom wordt ook in de literatuur gepleit voor opname van datamanagement in de opleiding van onderzoekers (bijvoorbeeld in het PhD

programma). Een andere manier is het houden van zogenaamde data audits, waarbij onderzoeksgroepen doorgelicht worden op het datamanagement van hun onderzoeksdata. Dit kan lacunes in de kwaliteitszorg aan het licht brengen.

Een derde mogelijkheid om kwaliteitszorg van datasets in een vroeg stadium aan te pakken, is het opnemen van een verplichte datamanagementparagraaf in de onderzoeksvoorstellen die ingediend worden bij onderzoeksfinanciers, zoals NWO. Een dergelijke paragraaf dwingt de indiener om na te denken hoe er tijdens en na afloop van het onderzoek met de onderzoeksdata omgegaan moet worden.

‘Commentaar op datasets in alle disciplines gewenst’

Onderzoekers zijn in het algemeen gebonden aan een gedragscode, die nationaal, in instellingsverband of in disciplineverband is vastgesteld. In Nederland geldt bijvoorbeeld de Gedragscode Wetenschapsbeoefening van de VSNU², die nu al een bewaartermijn van vijf jaren na publicatie voor onderzoeksdata vereist met het oog op validatie van de onderzoeksresultaten. Een mogelijkheid is om deze gedragscode op dit punt uit te breiden en aan te scherpen. Tenslotte zou een mogelijke maatregel kunnen inhouden om Open Access beschikbaarstelling van datasets te vereisen, eventueel na een embargoperiode. Een dergelijke maatregel zou immers ook kunnen leiden tot een verbetering van de kwaliteit van onderzoeksdata.

En wat willen onderzoekers?

Onderzoekers zijn zowel de dataproducenten als in veel gevallen ook de dataconsumenten. De vraag wat zij willen is dus meer dan opportuun. In een enquête onder een representatieve steekproef van hoogleraren en UHD's zijn hen de negen hierboven geschetste mogelijkheden voorgelegd. De 392 respondenten konden aangeven welke drie opties zijn voor hun vakgebied het meest stimulerend zouden vinden, en welke drie opties bij hen op de meeste bezwaren zouden stuiten.

Daaruit valt af te leiden dat de mogelijkheden om datasets te citeren, datapublicaties te beschrijven in de daartoe opgezette tijdschriften en commentaren over kwaliteit te leveren bij al gepubliceerde datasets bij zowel technische & natuurwetenschappen, sociale & geesteswetenschappen en levenswetenschappen in een behoefte voorzien. Die mogelijkheden zouden dus met spoed ontwikkeld moeten worden. Tegelijk zijn er tussen de disciplines ook verschillen waar te nemen. In de technische & natuurwetenschappen wordt Open Access beschikbaarstelling van datasets als een goede optie gezien. In de levenswetenschappen daarentegen ligt dit van-

wege de privacygevoeligheid en andere ethische kwesties ingewikkelder. Hier zou een gedragscode voor het omgaan met de beschikbaarstelling van datasets ontwikkeld moeten worden. De sociale en geesteswetenschappen zitten tussen deze twee andere disciplines in.

Hoge percentages respondenten blijken bezwaren te hebben tegen het peer reviewen van datasets als onderdeel van peer review van het tijdschriftartikel: dat gaat hen te veel tijd kosten. Deze optie dient daarom geen prioriteit te krijgen.

De opties met betrekking tot een verplichte datamanagementparagraaf in onderzoeksvoorstellen en het instellen van periodieke data audits zijn impopulair bij veel onderzoekers en dienen daarom niet in deze vorm te worden ingevoerd.



Onderzoekers gaven op 3 februari 2011 hun reacties op rapport 'What researchers want'

² <http://www.vsnul.nl/Media-item/Nederlandse-Gedragscode-Wetenschapsbeoefening.htm>

11

DE VISIE VAN NWO: STOKKEN EN WORTELS

Interview met Ron Dekker (NWO)
Door: Aad van de Wijngaart

Als financier van wetenschappelijk onderzoek kan de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) een sleutelrol spelen bij het toegankelijk maken van de verzamelde onderzoeksgegevens. Hoe kijkt ze daar zelf tegenaan? We vroegen het aan Ron Dekker, NWO-directeur Instituten, Financiën en Infrastructuur.

"Toegang zonder drempels." Dat wil NWO volgens Dekker bereiken voor de resultaten van het gesubsidieerde onderzoek. Het middel is ook duidelijk: "Dat is het Open Access-model".

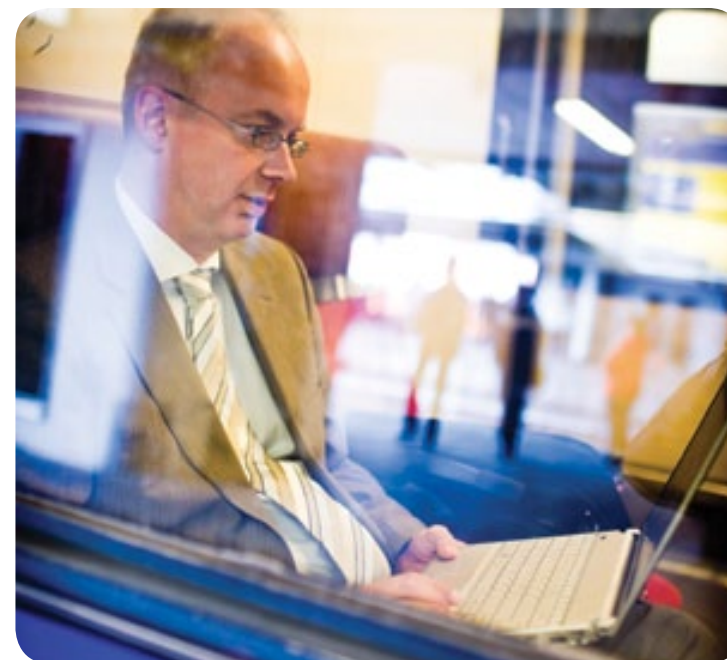
Maar vervolgens wordt het een stuk lastiger, tenminste waar het gaat om ruwe data. "Bij artikelen heeft één partij het auteursrecht, en die kan eenvoudig zelf bepalen dat het stuk openbaar toegankelijk wordt. Bij onderzoeksdata ligt dat veel ingewikkelder."

Volgens juristen kun je exclusief eigenaar zijn van bewerkte data, of van plaatjes die erop gebaseerd zijn, zo legt Dekker uit. Het eigendom van ruwe data is echter heel complex. "Volgens de Databankenwet ligt die bij de financier en de werkgever van de onderzoeker. Maar privacybescherming kan dat recht beperken. Bovendien kan de onderzoeker dwarsliggen: hij heeft immers al die tijd erin gestoken, ook al was het tijd van de baas. Gaat hij dan zijn onderzoeksgegevens weggeven aan de concurrentie?"

Duizend variabelen

Die vraag wordt nog prangender, doordat het extra werk vergt om ruwe data bruikbaar te maken voor derden. Daarvoor is namelijk een nauwkeurige beschrijving nodig. Dekker: "Vaak zijn er wel duizend variabelen. Definities veranderen soms in de loop der tijd, of je bent anders gaan meten, of de steekproef is veranderd. Het maakt ook heel veel uit

¹ Zie: http://www.nwo.nl/nwohome.nsf/pages/NWOP_86AFST



Ron Dekker

of bepaalde cijfers bruto of netto zijn. Dat zijn allemaal zaken die iemand precies moet weten om met de gegevens te kunnen werken." "Een voorbeeld is de epidemiologie. Daar werkt men met lange tijdreeksen. Dan moet je zoveel kennis hebben van de structuur van die data... Het is efficiënter om je bij het onderzoek aan te sluiten, dan te zeggen: doe mij maar een cd-rom en ik zoek het verder zelf wel uit." En dan is er nog een praktisch probleem: "Een artikel kun je gewoon lezen, of het nu op papier staat of in een PDF. Maar data moet je kunnen analyseren en bewerken. Dan word je meteen geconfronteerd met het gegeven dat elke discipline en elk type data een eigen format heeft, met bijbehorende ingewikkelde software."

'Databestanden zijn vaak vruchtbare akkers, waar in korte tijd vele malen geoogst kan worden'

Revolutie

Er zijn dus nogal wat drempels op de weg naar het ideaal van data-hergebruik. Ze maken dat ideaal echter niet minder aantrekkelijk, want databestanden zijn vaak vruchtbare akkers, waar in korte tijd vele malen geoogst kan worden. Dekker: "Het CBS had bijvoorbeeld enquêtes over koopgedrag. Het heeft een tijd geduurd voor onderzoekers doorkregen, dat die ook de beste metingen bevatten van het inkomen. Dus die data bleken goud waard."

Bovendien gaan de wetenschappelijke trends dezelfde richting uit. Dekker geeft het voorbeeld van LOFAR², het stelsel van kleine radio-telescopen dat verspreid is over een groot deel van Noord-Nederland

² <http://www.lofar.org/>

en -Duitsland. "Bij LOFAR gaat het niet meer om een schotel die je ergens op een bepaald punt richt: zowat de hele hemel wordt nu continu bemeten. Dat is een revolutie! Een onderzoeker kan nu verschillende data combineren en (nog steeds) focussen op een bepaald gebied. Die data liggen dus al klaar."

Hij moet ze alleen nog kunnen vinden tussen al die petabytes van meetgegevens. Maar dat komt wel goed, denkt Dekker. "Daar kunnen de disciplines zelf het beste voor zorgen. In de astronomie heb je al goede standaarden, zodat je snel kunt zoeken. In de sociale wetenschappen is een standaard voor metadata ontwikkeld die onderzoekers helpt om te beschrijven hoe hun data in elkaar zitten. Dat soort dingen kun je niet van bovenaf bedenken: het veld moet dat zelf regelen. En dat gebeurt ook, want voor toppublicaties moet je samenwerken."

Kwaliteitszegel

NWO laat het echter niet bij goede hoop, aldus Dekker. "Op het moment dat wij de subsidieaanvraag krijgen voor een groot onderzoeksproject, moet men aangeven dat die nieuwe data nodig zijn, en niet al elders bestaan. En om de beschikbaarheid te garanderen, hebben we samen met de KNAW DANS opgezet: Data Archiving and Networked Services."

Dat instituut doet twee dingen. Om te beginnen data archiveren en dus toegankelijk maken. Daarbij wordt de betrokken onderzoekers meteen geleerd om vanaf het begin alles nauwkeurig te beschrijven. "Ze moeten beseffen, dat ze eerst moeten vastleggen hoe die data eruit gaan zien. Want als ze dat achteraf doen, dán kost het heel veel tijd. Dus daar zijn voorlichting en kennisoverdracht nodig."

DANS kan ook een organisatie leren om gegevens zelf duurzaam en toegankelijk op te slaan. "Dan is het goed mogelijk dat een groot onderzoeksinstituut alle data zelf bewaart, maar dat DANS ervoor zorgt dat de metadata bepaalde kwaliteitsnormen halen. Daarvoor is het DANS-kwaliteitszegel ontwikkeld: dat biedt gebruikers van data zekerheid."

Laatste tranche

Dan is er nog één valkuil. Artikelen worden steeds meer verrijkt met onderliggende databestanden. Dreigt er geen risico dat uitgevers daar ook exclusiviteit op gaan claimen, legaal of niet?

'Ik wil het liefst uit de gebodssfeer blijven'

Het antwoord is nee, zo benadrukt Dekker. "Ik wil het liefst uit de gebodssfeer blijven. Maar bij grote databestanden kunnen we opneemen dat de makers verplicht worden hun gegevens te deponeren volgens de standaard van DANS. Als iemand dan niet aan zijn verplichtingen voldoet, dan is het aan de subsidiegever om te zeggen: we doen het zelf, en wel met het geld van de laatste tranche van jouw subsidie. Dat kan. Anders zijn we kapitaal aan het verspillen. Maar we zullen per geval kijken hoe te handelen."

Ten slotte wordt er naast de stokken ook gewerkt aan een extra wortel voor onderzoekers. "DANS is bezig om samen met organisaties als SURF en de Koninklijke Bibliotheek een 'persistent identifier' te ontwikkelen. Dat is een unieke code voor digitale objecten. Die maakt het ook mogelijk om datasets, én hun makers, bij naam te citeren. Dat kan heel goed werken om meer credits te geven aan degene die de data verzameld heeft."

PersID

Permanente toegang tot digitale publicaties en andere elektronische bronnen is een steeds belangrijker aspect voor de gemeenschappen voor wetenschappelijk en cultureel erfgoed. Gegarandeerde toegang tot onderzoeksinformatie vereist een ondubbelzinnige identificatie van informatieobjecten door een blijvend herkenningsteken – de 'persistent identifier'.



Tien nationale organisaties in acht Europese landen combineren hun deskundigheid in een gezamenlijk initiatief: PersID. Doel is om een internationale infrastructuur voor persistent identifiers te realiseren. Daarnaast zal een roadmap opgesteld worden – hoe zo'n infrastructuur zich zou moeten ontwikkelen. De partners hebben gekozen voor 'Uniform Resource Names for National Bibliography Numbers' ofwel de URN-NBN. Dit identifier systeem is al naar tevredenheid in gebruik bij Europese nationale bibliotheken, baseert zich op IETF (*) internet standaarden, en wordt niet gedomineerd door commerciële partijen.

De ambitie van PersID is het opleveren van een werkende infrastructuur die zowel lokaal, nationaal als internationaal toegekende URN-NBNs kan 'resolven' (dat wil zeggen: de persistent identifier herkennen en verwijzen naar het object, de bron of informatie over het object). Op zo'n infrastructuur kunnen allerlei referentie- en andere diensten worden gebouwd. Dit vereist technische oplossingen, maar belangrijker: een alliantie van elkaar vertrouwende partijen die de kwaliteit van de resolver-infrastructuur stuurt en bewaakt.

(*) De internationaal samengestelde 'Internet Engineering Task Force' voor internetarchitectuur

www.persid.org

12

INTERNATIONALE WORKSHOP: ONDERZOEKERS STIMULEREN TOT HET DELEN VAN ONDERZOEKSDATA

Door: Christoph Kümmel
Deutsche Forschungs Gemeinschaft

In september 2009 organiseerde Knowledge Exchange een internationale workshop over de voorwaarden om onderzoeksdata succesvol te kunnen delen en hergebruiken¹. Deze werd bijgewoond door 65 deelnemers uit Nederland, het VK, Duitsland, Denemarken, Frankrijk en de VS. Tijdens deze bijeenkomst in Berlijn bleek het belangrijkste obstakel om dit doel te bereiken het feit dat onderzoekers onvoldoende worden beloond voor het beschikbaar maken van hun data. Dit vormt een veel groter probleem dan technische hindernissen. Voor een bewustzijnsverandering op dit terrein is samenwerking nodig van onderzoekers, onderzoeksinstellingen, uitgevers, subsidieverleners en overheden.

Momenteel groeit het inzicht dat organisaties en individuen die onderzoek uitvoeren met publiek geld, een verantwoordelijkheid hebben om hun resultaten en bronnen beschikbaar te stellen, ook ten behoeve van vervolgonderzoek. Het is daarom de hoogste tijd om na te gaan hoe onderzoekers in de praktijk kunnen profiteren van de mogelijkheden van hergebruik van data en wat nog méér nodig is om dat hergebruik tot een succes te maken.

Om hier inzicht in te krijgen werden in Berlijn 'best practices' bijeen gebracht uit een groot aantal disciplines om zo te achterhalen wat voor de onderzoeker zelf de beweegredenen van hergebruik zouden kunnen zijn.

De discussies brachten echter niet alleen de voordelen aan het licht, maar gaven ook inzicht in de obstakels die effectief hergebruik in de weg staan.

¹ Een volledig verslag en de presentaties zijn te vinden op:
<http://www.knowledge-exchange.info/Default.aspx?ID=324>

Drie invalshoeken

Het twee dagen durende programma ging uit van drie invalshoeken:

1. Het perspectief van onderzoekers afkomstig uit diverse disciplines. Wat zijn voor hen de voor- en nadelen van hergebruik van data?
2. De wetenschappelijke infrastructuur voor het delen van data. Dit werd toegelicht door instellingen die zich daarmee bezig houden.
3. Uitspraken van nationale en Europese subsidieverstrekkers en uitgevers over de mogelijkheden om hergebruik van onderzoeksdata te verbeteren.

Ulf Leser van de Humboldt Universiteit gaf een aftrap met zijn lezing over 'Social Issues in Scientific Data Exchange'. Daarin beklemtoonde hij de noodzaak om aandacht te geven aan de mentale, organisatorische en methodologische voorwaarden voor succesvol hergebruik.

De daarop volgende presentaties vanuit verschillende wetenschappelijke disciplines lieten een grote diversiteit zien aan praktijkervaringen, maar niettemin werd duidelijk dat sommige bedreigingen en hindernissen vergelijkbaar zijn.

Herhaaldelijk werd erop gewezen dat de basale technische vragen in principe reeds zijn opgelost, maar in de praktijk niet altijd goed worden aangepakt. Daarbij gaat het vooral om standaarden voor metadata en metadata-uitwisseling, maar ook om toezicht op de kwaliteit van het materiaal in de data-archieven.



Ulf Leser besteedde aandacht aan de sociale aspecten

'Mogelijkheden om data te hergebruiken kunnen alleen toenemen door een cultuuromslag'

Bovenal echter werd duidelijk dat het succesvol beheer van een data-repository, het organiseren van een data-service en het mogelijk maken van optimale uitwisseling van datasets geenszins alleen een technische kwestie is. In vrijwel elke presentatie werd gewezen op methodologische implicaties, zoals geharmoniseerde ontologieën, deugdelijke klassificatiesystemen of andere, meer discipline-specifieke, methodologische problemen die door de betreffende onderzoeksgroepen opgelost zouden moeten worden.

Beloon de onderzoekers

Zoals overtuigend werd aangedragen door Ulf Leser is de bereidheid om onderzoeksdata te delen vooral een sociale aangelegenheid, en een mentaliteitskwestie. Mogelijkheden om data te hergebruiken kunnen derhalve alleen toenemen door een cultuuromslag in die



Toehoorders volgen de discussie op de tweede dag

richting te bewerkstelligen, onder meer door beloningen in het vooruitzicht te stellen. Dit geldt zowel voor individuele onderzoekers als voor onderzoeksgroepen die zich bereid tonen hun data te delen.

In dit licht kan ook worden gedacht aan maatregelen zoals:

- trainen van onderzoekers in data-hergebruik (vaardigheden);
- creëren van bewustzijn op alle niveaus: overheden, management, onderzoekers, uitgevers;
- uitgeven van 'best-practice'-gidsen, afgestemd op de behoeften van elke discipline;
- promoten van onderzoeksprojecten waarin hergebruik van data is toegepast, bijvoorbeeld via subsidieverstrekking;
- subsidies verlenen of in het vooruitzicht stellen aan datacentra of instellingen die zich met het delen van data bezig houden;
- juridische zekerheid nastreven en een juridisch raamwerk ontwikkelen ter bescherming van onderzoekers.

Beloningen voor onderzoekers die data delen en/of data hergebruiken kunnen gezocht worden in de sfeer van subsidies, maar kunnen ook worden meegenomen in beoordelingen van onderzoek. Hergebruik zou veel relevanter zijn als bijvoorbeeld uitgevers hun auteurs vragen niet alleen hun onderzoeksresultaten maar ook hun data beschikbaar te stellen. De technische randvoorwaarden dienen ook te worden ondersteund, onder meer door data-opslagruimte te creëren, door een gespecialiseerde staf ter ondersteuning aan te stellen en tenslotte door standaarden te ontwikkelen (metadata, ontologieën) waarmee datasets kunnen worden geëvalueerd, vergeleken en gecombineerd. Maar minstens zo belangrijk is het dat onderzoekers geschoold worden en er ook in de hogere bestuurlijke echolons bewustwording gaat plaatsvinden. Bottom-up en top-down benaderingen zullen moeten worden gecombineerd. Hier ligt een taak voor



alle betrokkenen: onderzoekers, onderzoeksinstituten, politici, nationale en Europese subsidieverstrekkeners en andere belangenvertegenwoordigers, zoals SURF.

Knowledge Exchange

Knowledge Exchange (KE) is een samenwerkingsverband van toonaangevende nationale organisaties uit Europa. Knowledge Exchange ontwikkelt samenhangend beleid voor het Europese hoger onderwijs en onderzoek ten aanzien van ICT-diensten, complexe ICT-infrastructuren en informatiestromen. SURF-foundation werkt in Knowledge Exchange samen met JISC (VK), DEFF (Denemarken) en de Deutsche Forschungs Gemeinschaft (Duitsland).



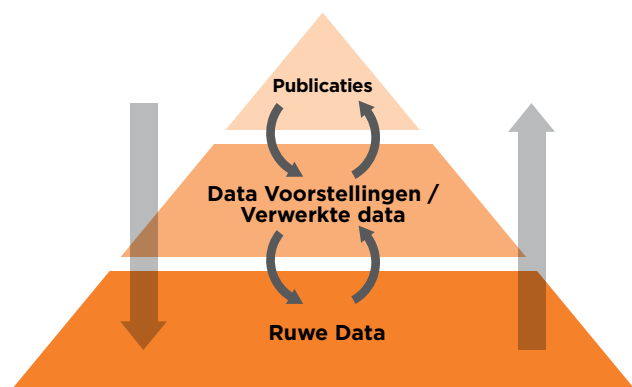
www.knowledge-exchange.info

13 VOOR ONDERZOEKSDATA IS HET ESSENTIEEL DAT ZE ALTIJD GELINKED ZIJN AAN PUBLICATIES

Interview met Eefke Smit (STM)
Door: Lilian van der Vaart

De belangstelling van Eefke Smit voor onderzoek en high-tech ontwikkelingen stamt al vroeg uit haar werkende jaren toen ze journalist was voor het NRC Handelsblad. Nu is ze 'Director of Standards and Technology' bij de International Association of STM Publishers (STM) en coördineert ze de activiteiten van het STM Future Lab Committee. Naast haar werk bij STM is ze ook actief als onafhankelijk consultant in de ontwikkeling van nieuwe initiatieven in e-publishing en andere uitgeef- innovaties. Ze brengt een uitgebreide professionele achtergrond in de uitgeefwereld mee in haar werk en was verantwoordelijk voor de ontwikkeling van bekende innovatieve toepassingen als Scopus™, Scirus™ and ScienceDirect™.

De Britse dichter T.S. Eliot schreef ooit: "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" (Waar is de wijsheid die we verloren in kennis? Waar is de kennis die we verloren in informatie?) In de context van onderzoeksdata zou Eefke Smit willen toevoegen: "Where is the information we have lost in data?" (Waar is de informatie die we verliezen in data?).



Figuur 1: Ruwe Data ligt ten grondslag aan de publicatie.

Uit: T. Hey, S. Tansley and K. Tolle (Eds). "The Fourth Paradigm: Data-Intensive Scientific Discovery", Redmond, Washington: Microsoft Research, 2009



Eefke Smit

Ze meent dat dit de vraag is die uitgevers zouden moeten pogen te beantwoorden als het aankomt op de rol en de toegevoegde waarde bij het aanpakken van de stortvloed van data. "Het is cruciaal dat onderzoeksdata en publicaties onderling gelinked blijven", zegt Smit: "Data kunnen relatief betekenisloos blijven, of in ieder geval moeilijk om te interpreteren, zonder de context erbij die de publicatie biedt. Ondertussen wordt de publicatie verrijkt door de oorspronkelijke data die ten grondslag liggen aan de beweringen en conclusies in het artikel" (zie figuur 1).

Ze wijst op het feit dat meer dan 50% van de onderzoekers hun voorkeur uitspreken om onderzoeksdata mee te sturen naar uitgevers, samen met hun manuscript¹. Ze plaatst dit in de context van een ander feit, namelijk dat veel onderzoekers nog altijd terughoudend zijn om hun data te delen met anderen en dat in veel gevallen deze terughoudendheid voortkomt uit een angst voor misinterpretatie of misbruik. Deze argwaan kan aanzienlijk worden terug gebracht door de data beschikbaar te maken in de context van het artikel. Op deze wijze krijgt de auteur ook de officiële citatie naar de data.

Verder is het ook noodzakelijk dat de metadata toegankelijk en herbruikbaar zijn en "wat voor betere metadata over data kun je je voorstellen dan die van de volledige onderzoekspublicatie?", zegt Smit. Verwijzend naar een opmerking van Andrew Treloar² tijdens

¹ Het PARSE. Insight Rapport 2010 toont aan dat hoewel 60% van de onderzoekers data van anderen zouden willen gebruiken, 40% grote bezwaren hebben met het delen van hun eigen data. 25% van de onderzoekers maken op dit moment hun data publiek beschikbaar. PARSE.Insight. "Insight into digital preservation of research output in Europe." (June 2010) and Survey report (December 2009). http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf en http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

² 'Director of Technology' van de Australian National Data Service, zie hoofdstuk 7

een recente JISC (de Britse tegenhanger van SURF) workshop over het beheren van onderzoeksdata vertelt Smit dat data nu vaak niet beschikbaar, en anders wel onvindbaar, of on-interpreteerbaar en on-herbruikbaar zijn. Om de eerste van deze drie uitdagingen aan te pakken en data niet alleen beschikbaar, maar ook nog vindbaar en interpreteerbaar te maken, is het essentieel om altijd links te hebben over en weer tussen de datasets en alle publicaties over deze data.

Links tussen publicaties en data maakt ze

1. Beschikbaar
2. Vindbaar
3. Interpreteerbaar en zelfs
4. Herbruikbaar

Wellicht kunnen uitgevers een rol spelen in het verder faciliteren van hergebruik, bijvoorbeeld door aanklikbare grafieken in een artikel op te nemen die direct verwijzen naar de data van bewuste waarnemingen of berekeningen.

Dergelijke voorbeelden passen in de missie van vele uitgevers om zich te richten op verrijking van inhoud in een omgeving die in toenemende mate open en interlinked zal zijn.

'50% onderzoekers wil dataset meesturen met manuscript'

"Dit zeggende, is er ook een substantieel dilemma waar uitgevers mee kampen", vervolgt Smit. "Terwijl het over en weer linken van data en publicaties een noodzakelijke verbetering is, is niemand gediend bij het werkelijk opslaan van

data op aparte sites van uitgevers: dat leidt slechts tot meer silos en fragmentatie op het niveau van de data en op termijn misschien zelfs tot wezenlijke problemen voor de uitgevers en de editors van de tijdschriften." Desalniettemin vraagt een groeiend aantal auteurs de uitgevers om hun data op te slaan en de meeste uitgevers hebben een natuurlijke reflex om aan de wensen van hun auteurs tegemoet te komen.

Smit: "Mijn veronderstelling is dat auteurs de uitgevers benaderen omdat er (nog) geen goede archieven beschikbaar zijn in hun eigen vakgebied". De ervaring van uitgevers die de mogelijkheid aanbieden om data mee te sturen met het artikel, of het zelfs verplichten, leert dat er groeiende problemen zijn met de verwerking van een brede diversiteit aan bestandsformaten, hosting en het preserven van een snel groeiend volume van data op de website van de uitgever. En ook om de peer-review van de data goed te regelen. Vooral

STM neemt deel aan een aantal EU projecten over data, zoals ODE en APARSEN om bij te dragen aan de ontwikkeling van standaarden en best practices in samenwerking met andere spelers in de informatie keten.

En auteursrechten dan?

In de zogenaamde 'Brussels declaration' hebben uitgevers aangegeven dat ruwe data open beschikbaar zouden moeten zijn, ze moedigen openbare beschikbaarstelling aan. Deze bewering betreft ruwe en onbewerkte data, de basis in de data pyramide (zie figuur 1). De top van de pyramide bevat de conclusies op basis van die data zoals deze in artikelen worden vermeld. Deze top is van oudsher de focus van de uitgevers. Deze focus verschuift nu naar het tweede niveau. Data gepresenteerd in grafieken, kaarten en plottings vormen vaak een onderdeel van de publicatie en zijn onderhevig aan auteursrecht. Ook selecties uit de bewerkte data zouden hieraan kunnen voldoen. De mate waarin de data is bewerkt bepaalt in hoeverre de voorstellingen van de data onderdeel uitmaken van het artikel en dus onderhevig zijn aan auteursrecht. Volgens het PARSE Insight rapport, antwoordden 57% van de uitgevers dat ze niet vragen om het auteursrecht over te dragen op data die bijgeleverd wordt bij een artikel, maar 39% van de uitgevers vragen dit wel...

dit laatste probleem heeft er toe geleid dat bijvoorbeeld het Journal of Neuroscience zijn beleid heeft aangepast. Auteurs kunnen links opnemen naar extra materiaal op een andere site, maar de uitgever accepteert niet meer zogeheten 'supplemental materials' meegestuurd met het artikel en plaatst deze niet op de site van het tijdschrift. Voor het tijdschrift Cell heeft het geleid tot een beleid waarbij de omvang en types van extra ingezonden materiaal drastisch zijn ingeperkt.

Een aanpak die Smit haalbaarder acht op de lange termijn kan het voorbeeld zijn dat in een aantal disciplines³ opduikt waarbij grote centrale data archieven rond een thema worden opgezet. Voorbeelden zijn Genbank en ProteinDataBank (PDB) of the Cambridge Crystallographic DataCenter (CCDC). Het beleid van veel tijdschriften in deze vakgebieden - waaronder Cell en Journal of Neuroscience - is om auteurs te vragen hun data daar op te slaan en de uitgever de bijbehorende toegangsnummers en links aan te leveren naar de databases voor een stabiele weblink van hun artikel naar de dataset. In een toenemend aantal gevallen levert de database ook een link van de data naar de publicatie. Deze archieven worden meestal door de community ondersteund, met een eigen kwaliteitscontrole systeem en beleid ten aanzien van zogeheten persistent identifiers en preservatie. Zo zijn vraagstukken rondom betrouwbaarheid en duurzaamheid in ieder geval in hoge mate opgelost. Dergelijk data-archief beleid is een cruciale voorwaarde om te kunnen linken naar data in een gecertificeerde repository. Alleen op deze wijze kan naadloze toegang tot en persistente links tussen data en artikelen worden gewaarborgd.

³ Bijvoorbeeld in sub-disciplines van de biomedische wetenschappen, geowetenschappen, aardwetenschappen, milieuwetenschappen en scheikunde.

Een dergelijke aanpak vergt een nauwe samenwerking tussen data-archieven, onderzoekers en uitgevers. In veel vakgebieden is dit nog nauwelijks gangbaar en daar zullen zowel praktische als culturele barrières moeten worden overwonnen om deze samenwerking tot stand te brengen. Dit kan ook niet worden ondernomen door op zichzelf staande organisaties, noch aan de uitgevers kant, noch aan de kant van data-archieven of onderzoeksinstituten.

Smit: "Data zijn te complex, te gevarieerd en te dynamisch om ze over te laten aan een laag aggregatie niveau, of dat nou gebeurt op het niveau van een individuele onderzoeker of onderzoeksgroep, zijn institutionele repository, een tijdschrift of uitgever of misschien zelfs het niveau van een land".

'Het is cruciaal dat onderzoeksdata en publicaties onderling gelinked blijven'

Een geïntegreerde aanpak voor de omgang met data kan erg waardevol zijn, omdat het leidt tot schaalgrootte, volume en nieuwe mogelijkheden voor het zoeken en leggen van verbanden door de hele collectie heen. Dit vraagt om een inspanning voor normalisatie en standaardisatie die alleen kan worden bereikt op voldoende schaal niveau: internationaal en discipline georiënteerd, aldus Smit. Compartmentalisering zou een belangrijk struikelblok vormen voor de 'openheid van data', zoals ook onderschreven door de uitgevers in de 'Brussels Declaration on STM publishing'⁴.

In een omgeving die in toenemende mate dynamisch en open is, kan certificering, betrouwbaarheid en vindbaarheid nog belangrijker worden. De archieven kunnen een rol spelen in het vastleggen van de herkomst van data (provenance). Uitgevers kunnen bijdragen aan de vindbaarheid, bijvoorbeeld door middel van applicaties op het terrein van personificering en filtering. Misschien zal Peer review van data zich meer richten op de methoden en technieken gebruikt bij data verzameling en analyse, op de integriteit van de data. Met andere woorden zich sterker richten op het proces van data management. Het kan een dienst worden, die, al dan niet voor een meerprijs, kan worden afgenomen.

⁴ Brussels Declaration on STM Publishing, door de internationale scientific, technical and medical (STM) publishing community, 1 November 2007, http://www.stm-assoc.org/public_affairs_brussels_declaration.php



SURFproject VETERAN TAPES: HERGEBRUIK VAN KWALITATIEF ONDERZOEKSMATERIAAL

Voor het SURFshare project Veteran Tapes heeft een multidisciplinaire groep wetenschappers een wetenschappelijk artikel geschreven op basis van bestaande interviewdata. In de publicatie zijn duurzame links tussen de publicatie en de onderzoeksdata, bestaande uit geluidsfragmenten en de transcripties hiervan, opgenomen. De interviewdata zijn in fragmenten opgeknipt en gekoppeld aan de inhoudelijk gerelateerde onderdelen van het artikel. Einddoel is een wetenschappelijk boek dat, voorzien van verrijkingen in een elektronische vorm, is uitgebracht. Het biedt onderzoekers het voordeel dat ze voorheen ontoegankelijke

onderzoeksdata (interviews en transcripties) nu in relevante fragmenten aangeboden krijgen in de context van het wetenschappelijk artikel. Dit leidt tot efficiënter onderzoek en meer mogelijkheden tot interdisciplinair hergebruik van de interviews en transcripties. Een mooie toepassing van het concept Verrijkte Publicaties.



14 HET WERELDWIJDE LABJOURNAAL

Door: Jan Top
Vrije Universiteit

Onlangs hebben historici een oude mythe rond de ontdekking van supergeleiding doorgeprikt¹. Het verhaal ging dat het wegvallen van de elektrische weerstand in eerste instantie werd aangezien voor kortsluiting. Toen een assistent per ongeluk wegdommelde en de koeling vergat, bleek dat de weerstand weer terugkeerde. Blijkbaar was de temperatuur de bepalende factor. Door de labjournaals van Kamerlingh Onnes nog eens goed na te pluizen is nu gebleken dat de datering van de labjournaals niet helemaal klopt, en dat er nauwkeurige aantekeningen zijn van het moment van de ontdekking. Die blijkt heel bewust gedaan te zijn en van een slapende assistent was geen sprake. Deze reconstructie was mogelijk doordat destijds het nauwgezet maken van aantekeningen gezien werd als een essentieel onderdeel van goed wetenschappelijk onderzoek. Volgens de regels van het vak beschreven wetenschappers alle waarnemingen in handgeschreven labjournaals. Voordat de computer zijn intrede deed bevatte het labjournaal dus alle data, de manier van werken en andere informatie die nodig was om het onderzoek te kunnen reproduceren. Voor het schrijven van een publicatie waren de enige bronnen de labjournaals en publicaties van gerelateerd onderzoek.

De context van data gaat verloren

Door de komst van de computer is de hoeveelheid data uit onderzoek explosief gestegen. Veel gegevens komen uit geautomatiseerde meetapparatuur, digitale registraties en alle mogelijke sensoren. Deze data staan verspreid opgeslagen in databases, spreadsheets, apparaatspecifieke bestanden, op locale schijven, op servers, enzovoorts. De aantekeningen in labjournaals – als die al gemaakt worden – zijn opgeslagen op papier of in losse tekstbestanden. Er bestaat wel specialistische software voor het ondersteunen van labwerk, maar deze biedt geen mogelijkheid om verschillende data-

bronnen te verbinden. Er is dus sprake van een ont koppeling van de beschrijving van de waarnemingen en de eigenlijke data. Als gevolg daarvan is het vrijwel onmogelijk om de betekenis (context) van de datafragmenten te achterhalen als de auteur niet beschikbaar is. Deze ont koppeling van beschrijving en data is alleen nog maar versterkt doordat de computer niet alleen sterk is in het opslaan van data, maar er ook mee kan rekenen en nog mooier, uitkomsten kan visualiseren. Onderzoekers gaan relatief snel hun data analyseren, om daarmee tot publiceerbare resultaten te komen. De enorme populariteit van het spreadsheet bij onderzoekers bevestigt dit nog eens.

‘De analyse heeft het gewonnen van de beschrijving’

Met andere woorden, *de analyse heeft het gewonnen van de beschrijving*. Het gevolg is dat experimenten vaak niet meer controleerbaar en reproduceerbaar zijn; publicaties alleen bieden onvolledige houvast. Het is moeilijk om relevante datasets te identificeren in de zee van potentieel interessante bronnen. Vooral bij spreadsheets is dit een probleem geworden; er worden nauwelijks voorwaarden gesteld aan de beschrijving. De indeling en de inhoud zijn in tweede instantie vaak niet meer te begrijpen. Wereldwijd zijn er grote hoeveelheden gegevens uit onderzoek niet beschikbaar voor controle of aanvulling op andere gegevens doordat de context onvoldoende beschreven is. In feite gaat dit ten koste van de wetenschappelijke methode.

Op wereldschaal terugwinnen van samenhang tussen gegevens

Het Semantic Web, of beter het Web of Data, biedt de mogelijkheid om het verlies aan samenhang tussen gegevens zoals die in het oude labjournaal nog bestond, terug te winnen, en dat bovendien op wereldschaal. Door de context van data formeel uit te drukken in de nieuwe standaarden RDFS/OWL wordt het mogelijk wereldwijd relevante gegevens te vinden via zoekmachines en bovendien deze gegevens automatisch met elkaar te verbinden. Het elektronisch publiceren van afgerond onderzoek op het web is al gemeengoed; de volgende stap is het delen van de onderliggende data, methodes, materiaalbeschrijvingen, enzovoorts. Tim Berners-Lee, de grondlegger van het Semantic Web, zegt hierover² dat experimentele resultaten steeds meer op het web gepubliceerd gaan worden, binnen of buiten de context van een onderzoekspublicatie. Een wetenschapper kan een experiment ontwerpen en uitvoeren, en via een webpagina de data geleidelijk delen met vertrouwde collega's. Lopende experimenten en studies ontdekken wordt eenvoudig en het werk kan aangepast worden als gevolg van interactie met *peers*, waarbij men niet hoeft te wachten tot de uiteindelijke publicatie.

¹ Ontdekking supergeleiding was geen geluk bij een ongeluk, *Volkscrant* 28 januari 2011.

² Tim Berners-Lee and James Hendler "Publishing on the Semantic Web", *Nature*, April 26 2001 p. 1023-1025

Wat is de magie achter RDFS/OWL die dit alles mogelijk maakt? Die is er niet. Wel is er een gestandaardiseerde manier om relaties tussen begrippen te leggen (triples), en een manier om naar deze begrippen en relaties te verwijzen via unieke URI's. Deze formele notatie maakt automatische verwerking mogelijk. Op dit moment is er bijvoorbeeld veel activiteit rond het Linked (Open) Data initiatief³ dat onder andere DBpedia omvat, een verzameling van feiten uitgedrukt in RDFS/OWL, onttrokken aan Wikipedia. Daarnaast groeit het aantal tools voor toepassingen op het Semantic Web snel.

Een voorbeeld van een gebruiker van *linked data* is het onderzoeksinstituut TI Food and Nutrition. Dit instituut doet experimenteel onderzoek aan voeding en voedsel voor een internationaal consortium van voedingsbedrijven, waaronder Unilever, DSM, Danone, Nestlé en FrieslandCampina. Het is een virtuele instelling, wat betekent dat onderzoekers uit verschillende universiteiten en instellingen samen onderzoek doen in scherp omliggende projecten. Het goed registreren en delen van onderzoek is een bestaansvoorwaarde van het instituut.

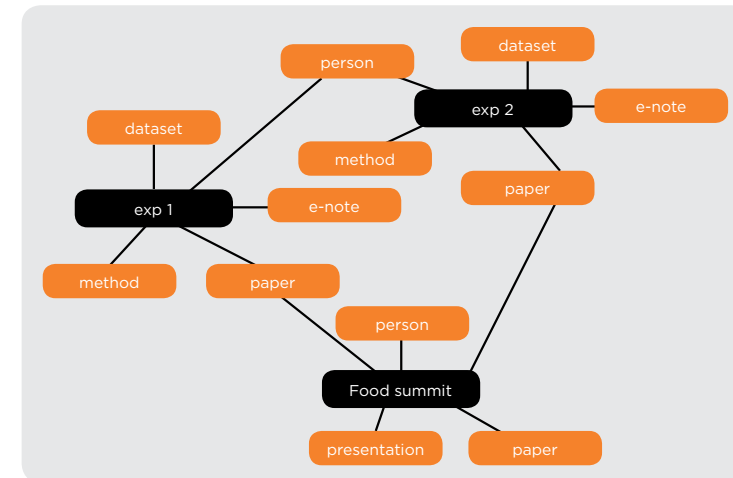
In de afgelopen jaren is een model ontwikkeld dat *top-down* de verschillende onderdelen van onderzoek beschrijft. Het omvat begrippen als onderzoeksvraag, materiaal, methode, meetapparaat en onderzoeker. Dit model koppelt publicaties aan onderliggende data, methodes, enzovoorts. Het effect is dat onderzoekers snel een dossier kunnen samenstellen voor een nieuwe publicatie of een patent. Ze kunnen data hergebruiken zonder een nieuw en kostbaar experiment uit te moeten voeren en data beschikbaar stellen voor nadere analyse. De webapplicatie Tiffany ontsluit de informatie voor de onderzoekers en partners van TI Food and Nutrition. Het model is geschikt om in de toekomst informatie te integreren van willekeurige andere bronnen die gebruik maken van een vergelijkbaar vocabulaire. De komende tijd zal de nadruk liggen op het invoeren van een elektronisch labjournaal als middel om laboratoriumobservaties op te nemen in het netwerk van onderzoeksresultaten. Op die manier ontstaat een gedeeld digitaal labjournaal dat aantekeningen, data, personen en apparaten verbindt in een netwerk van Linked Data (zie Figuur 1).

Duidelijk maken wat er gemeten is

Naast deze top-down benadering is aandacht nodig voor het annoteren van de data-elementen. Een elementair onderzoeksgegeven is een grootheid (zoals tijdstip, kracht, straatnaam, kleur, aantal) en een waarde voor die grootheid (10:00 uur, 6.5 kN, Hoogstraat, rood, 86). Een eerste stap is om het begrip grootheid en waarde uit te drukken in RDFS/OWL. De *Ontology of Units and Measures* (OM)⁴ is een voorbeeld van een vocabulaire dat grootheden, eenheden, eenhedenstelsels en dimensies uitdrukt. Door in de voorbeeldtabel in de kop te verwijzen naar de URI's van de betreffende grootheden

³ Zie <http://linkeddata.org/>

⁴ Zie <http://www.wurvoc.org/>



Figuur 1 In Tiffany vormen de verschillende onderzoeksactiviteiten een netwerk waarvan ingangen uit het elektronisch labjournaal hangen, datasets, publicaties, personen, enzovoorts.

en eenheden is voor elke gebruiker meteen duidelijk wat er gemeten is. Voor OM zijn bovendien een serie webservices gemaakt die het mogelijk maken automatisch eenheden in elkaar om te zetten, formules te controleren op consistentie, enzovoorts. Een add-in voor Excel maakt het vervolgens mogelijk om een dergelijke annotatie van data achter de betreffende cellen op te slaan. Een applicatie die deze RDFS/OWL metadata begrijpt kan dan meteen bepalen wat er gemeten is, in welke eenheden, enzovoorts. Een volgende stap is om ook de objecten waaraan gemeten is te annoteren, bijvoorbeeld met behulp van een vocabulaire over voedingsproducten. Maar ook de meetmethode, de apparatuur en de onderzoekers kunnen geannoteerd worden. We kunnen dan antwoorden krijgen op vragen als: Wie hebben nog meer aan dit materiaal gewerkt? Is die eigenschap ook door anderen gevonden? Wanneer was dat? Het annoteren van elementaire data kan men zien als een bottom-up manier om aan het netwerk van Linked Data voor onderzoek te bouwen.

Om terug te komen op Nobelprijswinnaar Kamerlingh Onnes: als hij destijds zijn ontdekking in de vorm van Linked Data de wereld over had kunnen twitteren, had dit waarschijnlijk geleid tot een golf van experimenten en een enorme versnelling van onderzoek. De keerzijde is dan natuurlijk wel dat we het verhaal over de wegdommelende assistent hadden moeten missen.

15 CONCLUSIE EN VERDER LEZEN

Door: Wilma Mossink en Keith Russell
SURFfoundation

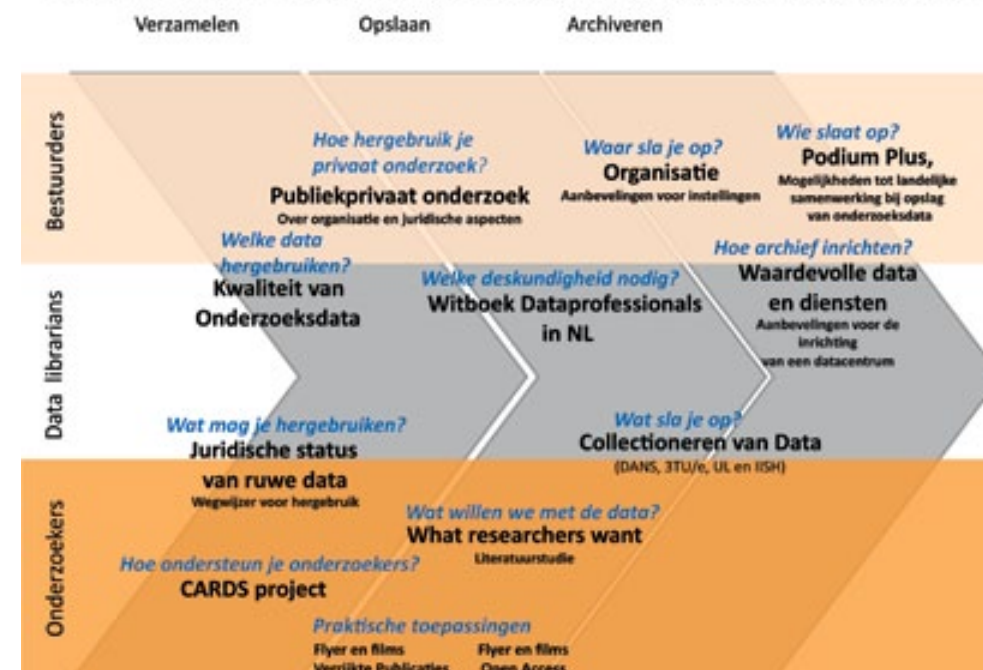
Clichés zijn er om gebezigd te worden. Daarom mogen we best constateren dat er veel in beweging is op het gebied van het opslaan, bewaren, hergebruik en koppelen van data. Dat blijkt ook uit de verschillende hoofdstukken in deze publicatie. Deze laten zien dat er in het kader van het SURFshare programma en erbuiten mooie projecten zijn uitgevoerd die ons inzicht hebben verschaft in allerlei aspecten van het omgaan met data.

Het werkpakket 'Toegang tot data' van het SURFshare programma heeft de ambitie om aan het einde van 2011 een geheel van richtlijnen, studies en aanbevelingen op te leveren die Nederlandse hoger onderwijs en onderzoek instellingen kunnen gebruiken om beleid te vormen om toegang tot en hergebruik van data te bevorderen. Het is voorbarig om te spreken van een geheel waar niets meer aan toe te voegen valt; er zijn er bouwstenen vervaardigd waarmee op een fundament kan worden gemetseld. Zo kan de toekomstvisie van wereldwijde labjournalen en een web van aan elkaar gekoppelde data gestalte krijgen.

Er is een begin gemaakt om data intensieve wetenschap op het niveau van organisatie en beleid vorm te geven. Tegelijkertijd zijn we ons ervan bewust geworden dat er nog veel (misschien wel heel veel) gedaan moet worden. Toegang geven tot data en zorgen dat deze toegang blijvend is, is geen automatisme maar vraagt een inspanning van alle partijen.

Het is voor onderzoekers nog lang niet vanzelfsprekend dat zij hun data delen met een kleine groep medeonderzoekers of studenten, laat staan met de hele wereld. Redenen van privacy of angst over misbruik van hun data weerhoudt nog velen om derden toegang te geven tot hun data. Hoe gaan we wetenschappers overtuigen dit toch te doen? Hun enthousiasme en medewerking zijn de basis voor

Proces naar duurzame opslag van onderzoeksdata voor blijvende toegankelijkheid



een nieuwe dimensie van wetenschap bedrijven. We moeten ons daarbij realiseren dat we aan onderzoekers vragen een deel van hun onderzoekstijd af te staan om ervoor te zorgen dat hun data vindbaar zijn. Verder willen wij dat zij moeite doen om bruikbare meta-data, die voor buitenstaanders verstaanbaar zijn, bij hun data te leveren en dat zij standaarden toepassen die zij wellicht niet als wenselijk ervaren.

Universiteiten, hogescholen en onderzoeksinstituten kunnen hierbij helpen maar dit gaat ook niet vanzelf. Deze instellingen kunnen bijdragen, niet alleen door onderzoekers te ondersteunen maar ook in het aanstellen of opleiden van data stewards of zelfs het formuleren van functie-eisen waaraan een data steward zal moeten voldoen.

Financiers hebben een sterke troef in handen. Zij kunnen het maken van datamanagement plannen verplicht stellen of de permanente toegang tot, in het kader van door hen betaald onderzoek, gegeneerde data als voorwaarden op nemen voor financiering. Zo kunnen zij een impuls geven aan het hergebruik van data.

Bij de bestuurders van de onderzoeks- en onderwijsinstellingen ligt de taak om afspraken te maken over liefst landelijk beleid hoe het landschap van het delen van data organisatorisch ingericht zal gaan worden. Welke onderwerpen daarbij door welke organisatie opgepakt moeten worden is beschreven in het rapport 'Organisatorische aspecten rondom de opslag van data'. De belangrijkste bevindingen zijn na te lezen in hoofdstuk 9.

Een overzicht van de studies en projecten uit het SURFshare programma over Toegang tot Data

Onverbbidlijk komt de discussie over de juridische eigendom van data en/of databanken naar boven. We willen voorkomen dat hierover net als over de juridische status van publicaties een eindeloos debat losbreekt dat ons verlamt om tot daadwerkelijke tot actie over te gaan. Daarom is het nodig dat betrokken partijen hier een standpunt over innemen en daar gezamenlijk afspraken over maken. Een charter van eigendom kan hiervoor een passende oplossing zijn.

Tot slot blijkt uit de bovenstaande opsomming dat we er nog lang niet zijn. Volledige open data is een mooi ideaal maar voorlopig moeten we allereerst streven naar de mogelijkheid om de data daar waar mogelijk open te maken en daar waar niet anders kan deze op te slaan, te bewaren en deze beperkt toegankelijk te maken. Maar wat we ook doen, het moge duidelijk zijn dat we de komende jaren nog hard moeten werken om onze idealen te verwezenlijken.

VERDER LEZEN

OECD report Principles and Guidelines for Access to Research Data from Public Funding

Dit rapport was het startpunt van de discussie over het openbaar maken van data en bevat nog altijd waardevolle inzichten.

<http://www.oecd.org/dataoecd/9/61/38500813.pdf>

e-IRG Report on Data Management

Dit rapport biedt een overzicht van projecten en initiatieven op het gebied van data management. Het geeft interessante aanbevelingen op de terreinen van metadata, kwaliteit en interoperabiliteit. Dit overzicht is voorzien van vele nuttige referenties en webadressen waar meer informatie kan worden gevonden.

http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

PARSE Insight: Insight into digital preservation of research output in Europe

Het Insight rapport geeft inzicht in onderwerpen die te maken hebben met permanente toegankelijkheid. Dit rapport is juist ook van belang omdat ook uitgevers betrokken waren bij de totstandkoming.

http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

The Data Curation Continuum, Managing Data Objects in Institutional Repositories

Dit artikel van Treloar, Groenewegen en Harboe-Ree uit 2007 beschrijft de verschillende silo's waarin onderzoeksdata kunnen worden opgeslagen. Dit uitgangspunt komt in vele projecten op het gebied van data management terug.

<http://www.dlib.org/dlib/september07/treloar/09treloar.html>

Riding the Wave, How Europe can gain from the rising tide of scientific data

De High Level Expert Group on Scientific Data heeft dit rapport in 2010 aangeboden aan de Europese Commissie. Het vraagt de Europese Commissie meer aandacht te besteden aan onderzoeksdata en noemt een aantal maatregelen die zouden kunnen worden ondernomen.

<http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

The Fourth Paradigm: Data-Intensive Scientific Discovery

Een toegankelijk boek over data intensive science met aansprekende voorbeelden uit verschillende wetenschappelijke disciplines. Het toont aan wat je met onderzoeksdata kunt bereiken maar geeft ook inzicht in de benodigde onderliggende infrastructuur.

<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>



COLOFON

Auteurs

Inge Angevaare (NCDD)
Baukje Beetsma (Parelsnoer)
Madeleine de Cock Buning (CIER)
Bas Cordewener (SURFfoundation)
John Doove (SURFfoundation)
Maurits van der Graaf (Pleiade Management en Consultancy, Amsterdam)
Annemiek van der Kuil (SURFfoundation)
Christoph Kümmel (DFG)
Tina van der Linden (CIER)
Wilma Mossink (SURFfoundation)
Warna Oosterbaan (NRC Handelsblad)
Allard Ringnalda (CIER)
Jeroen Rombouts (Technische Universiteit Delft)
Brenda Sorensen (DANS)
Eric Sieverts (Universiteit Utrecht)
Jan Top (Vrije Universiteit)
Lilian van der Vaart (Vaart Innovatie & Kennismanagement)
Leo Waaijers (freelance Open Access consultant)
Aad van den Wijngaart (in opdracht van SURFfoundation)

Vertalingen

Gretton & Willems Translations
Armand Heijnen
Keith Russell

Eindredactie

Armand Heijnen en Keith Russell (SURFfoundation)

Fotografie

Inge Angevaare (p.16, 18)
Annemiek van der Kuil, PhotoA (p.33, 35, 43, 45, 55)
Ivar Pel (omslag, p.4, 12, 13, 14, 25, 27, 47)
BILDSCHEIN Berlin (p.51, 52, 53)
iStockphoto (p.21, 41, 59)

Ontwerp en opmaak

Vrije Stijl, Utrecht

Druk

Drukkerij Libertas, Bunnik

© Stichting SURF, juli 2011

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



Deze publicatie verschijnt onder de Creative Commons licentie Naamsvermelding 3.0 Nederland.
<http://creativecommons.org/licenses/by/3.0/nl/>

Deze publicatie in de SURFshare reeks handelt over het delen van onderzoeksdata. Voorheen konden wetenschappers alleen tekstuele bronnen verspreiden en delen, maar dankzij het internet zijn daar immens veel mogelijkheden bijgekomen. Denk aan gedigitaliseerde afbeeldingen van door archeologen opgegraven artefacten, of een enorme database waarin alle menselijke genen beschreven staan. Het internet biedt vervolgens de mogelijkheid om die data op te slaan en te delen, om met collega's aan de andere kant van de wereld aan die data te werken.

Er is echter nog meer: met toevoeging van beeld, video of audio-materiaal zal niet alleen de communicatie tussen wetenschappers veranderen, maar zal de wetenschap zelf een ingrijpende verandering – een paradigmashift – kunnen ondergaan.

Het is voor onderzoekers nog lang niet vanzelfsprekend dat zij hun data delen met een kleine groep medeonderzoekers of studenten, laat staan met de hele wereld. Redenen van privacy of angst over misbruik van hun data weerhoudt nog velen om derden toegang te geven tot hun data, om nog maar te zwijgen over discussies over het juridische eigendom van data en/of databanken.

In deze publicatie wordt uitgelegd wat er al is ondernomen op dit terrein. Bestaande Nederlandse en internationale initiatieven worden beschreven, uitkomsten uit studies samengevat en we laten uitgevers en financiers aan het woord over de rol die zij verwachten te spelen in het delen van onderzoeksdata.



SURFfoundation
Hojel City Center gebouw D
Graadt van Roggenweg 340
3531 AH Utrecht
Postbus 2290
3500 GG Utrecht

T 030 234 66 00
F 030 233 29 60

info@surf.nl
www.surfoundation.nl

juli 2011