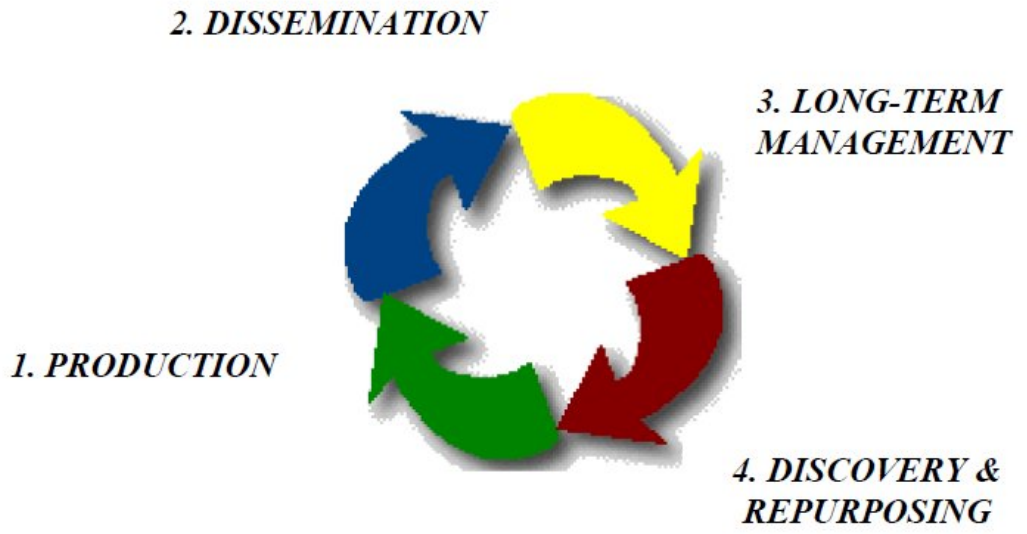
De cirkel van onderzoeksdata [1]

Eric Sieverts

Universiteitsbibliotheek Utrecht

Wetenschappelijke vooruitgang kent vaak kringlopen. Resultaten van onderzoek worden gepubliceerd of op andere wijze gecommuniceerd en geven zo weer aanleiding 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 gebiedend motto hanteert: *Stand on the shoulders of giants*.

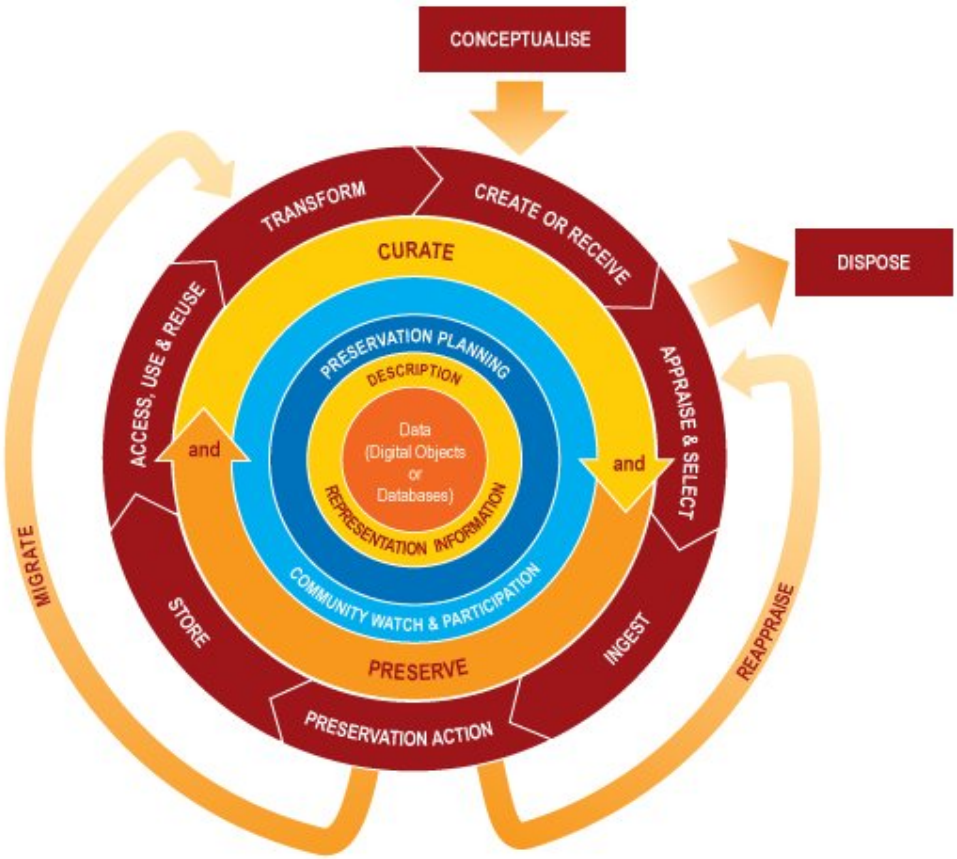
Dat die kringloopgedachte ook van toepassing is op de onderzoeksgegevens die aan wetenschappelijke publicaties ten grondslag liggen, wordt nog niet zo lang onderkend. Ook daarvoor kun je echter mooie cirkeltjes tekenen. Een Canadees onderzoek naar de Data Lifecycle doet dat simpel in vier stappen [1]:
production -> dissemination -> long-term management -> discovery & repurposing -> en dan is de cirkel weer rond naar "production".



---

[1] *Deze tekst is geschreven als inleiding op het boekje "Toegang tot onderzoeksdata / SURFfoundation, Utrecht, juli 2011" http://www.surffoundation.nl/nl/publicaties/Pages/toegangtotonderzoeksdata.aspx*

Het "Curation Lifecycle Model" van het Britse DCC (Data Curation Center) 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.



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, geluids- of video-opnamen van dier- of mensgedrag, of al door menselijk toedoen bewerkte en geaggregeerde gegevens. Zulke onderzoeksdata werden door de makers altijd wel enige tijd ergens opgeslagen en

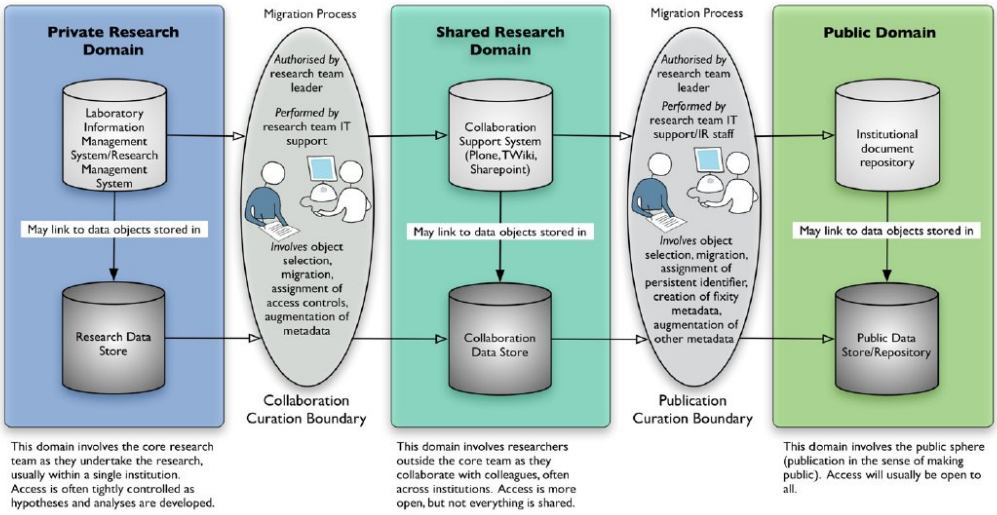
bewaard. Maar het langduriger, veiliger en duurzamer opslaan en het zorgvuldig beheren en toegankelijk houden (of maken) zijn pas recenter in de belangstelling komen te staan.
Een andere interessante categorie data is die van de *Linked Data*, datacollecties die vrij op internet beschikbaar zijn. Door op gestandaardiseerde wijze betekenis aan die gegevens toe te kennen, kunnen ze ook autonoom door computerprogramma's gebruikt worden.

Primaire onderzoeksdata

Universiteiten, onderzoeksinstituten en hogescholen richten al lang institutionele 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 andere ten behoeve van de verifieerbaarheid van gerapporteerde resultaten, analyses en conclusies. Vanuit digitale publicaties moet er dan met een permanent URL naar gelinkt kunnen worden. Dat kan aanleiding geven tot wat wel verrijkte publicaties en complexe objecten worden genoemd. Daarin moeten niet alleen multimediale illustraties, maar ook datasets op gestructureerde wijze aan teksten gekoppeld kunnen 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* [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, enzovoort.



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 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 terechtkomen: 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 software-matige interpretatie van de bestanden te kunnen garanderen.

Bij elke overgang naar een volgende fase zullen ook beslissingen genomen moeten worden ten aanzien van onder meer de toegankelijkheid, of metadata voor die fase nog adekwaat zijn en zelfs of de gegevens überhaupt nog bewaard moeten blijven. Zo zullen zeker niet alle ooit geproduceerde data in het derde stadium terecht komen. Behalve commerciële redenen 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 drie genoemde fases komen we in de Nederlandse praktijk ook al tegen. Zo biedt het sinds kort gestarte Utrecht Dataverse Network een infrastructuur voor die eerste twee fases, waarbij de eigenaar van de gegevens nog helemaal de baas blijft over (on)toegankelijkheid van de gegevens en het gebruik ervan door anderen. DANS (de Data Archiving & Networked Services van de KNAW) is in het leven geroepen voor die tweede en (vooral) derde fase, waar hergebruik en digitale duurzaamheid voorop staan.

Data deluge

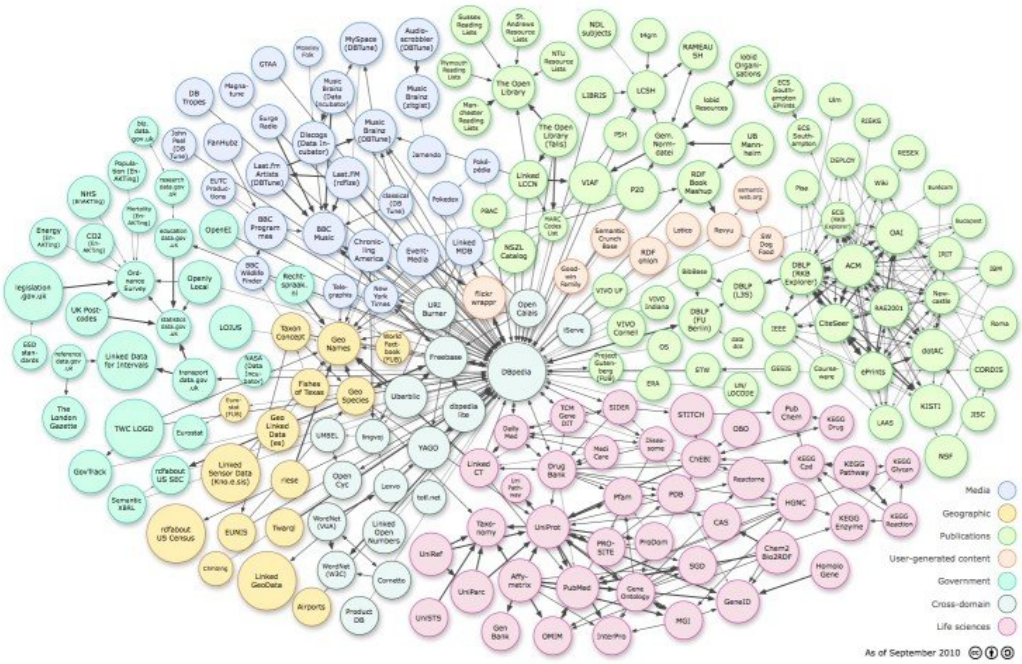
In allerlei verband wordt op dit moment gesproken over een op handen zijnde gegevensstortvloed, een *data deluge*. Bij onderzoeksdata denk je daarbij al gauw aan de enorme hoeveelheden gegevens die grootschalige experimenten, zoals met de nieuwe deeltjesversneller van CERN, gaan opleveren. Of aan de 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 vinden bijvoorbeeld hoge resolutie en hoge snelheid multimedia en de daarbij horende grote datastromen ingang. Men begint zich dan ook algemener te realiseren dat data management van te verwachten onderzoeksdata een noodzakelijk onderdeel van elke projectaanvraag zou moeten zijn.

Linked data

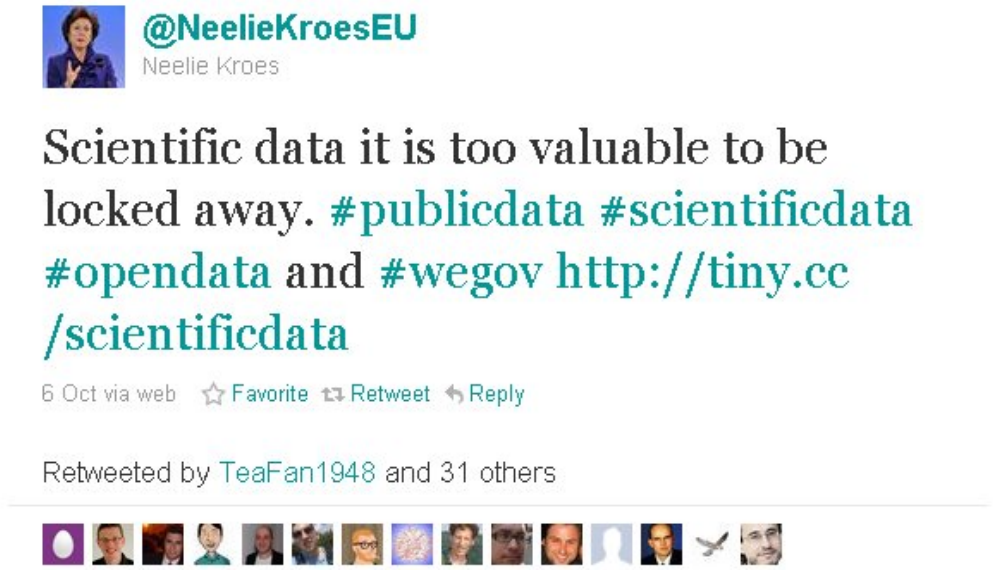
De grote dataverzamelingen die intussen op internet beschikbaar zijn, betreffen op dit moment vaak nog geen onderzoeksdata, maar allerlei algemene soorten gestructureerde gegevens. Opdat computers deze gegevens op standaardwijze kunnen lezen, interpreteren en gebruiken, en om zo allerlei systemen aan 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.

Vooral 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 DBpedia [6], 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 ontlenen. 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 datacollecties gevisualiseerd [7]. Ook daar is intussen sprake van een "deluge" van tientallen miljarden beschikbare RDF-tripels en varianten daarop.



Ook wetenschappelijke onderzoeksdata zullen uiteindelijk zo beschikbaar gesteld kunnen en moeten worden. Onlangs (6 oktober 2010) twitterde zelfs @NeelieKroesEU daarover [8]: "Scientific data it is too valuable to be locked away. #publicdata #scientificdata #opendata and #wegov http://tiny.cc/scientificdata"



[1] Stewardship of Research Data in Canada: A Gap Analysis (October 2008), http://data-donnees.gc.ca/docs/GapAnalysis.pdf

[2] DCC Curation Lifecycle Model, http://www.dcc.ac.uk/resources/curation-lifecycle-model

[3] Data curation continuum, http://ands.org.au/guides/curation.continuum.pdf

[4] Utrecht Dataverse Network, https://dataverse.library.uu.nl/dvn/

[5] Data Archiving and Networked Services, http://www.dans.knaw.nl/

[6] DBpedia, http://dbpedia.org/About

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

[8] http://twitter.com/#!/NeelieKroesEU/status/26568611661