

Grotere rol voor informatieprofessionals bij beheer van gegevensverzamelingen

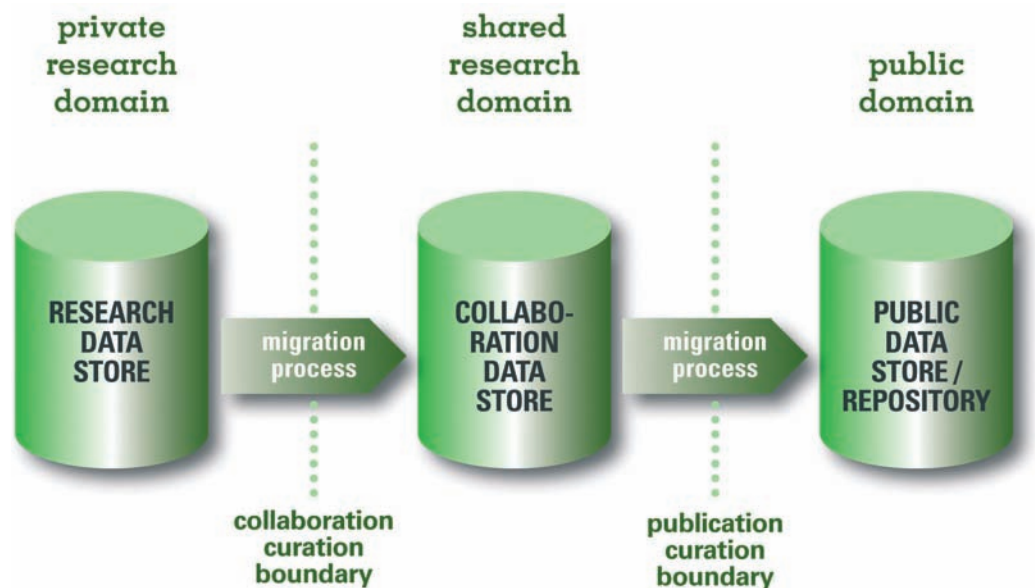
Enkele maanden geleden signaleerde Eric Sieverts in zijn column de toegenomen belangstelling voor de data die ten grondslag liggen aan veel informatie en kennis waarmee informatieprofessionals zich bezighouden. In een serie korte artikelen passeren daarom allerlei data-initiatieven en ontwikkelingen de revue. Hier de inleiding op deze serie.

Eric Sieverts

In de artikelen in deze serie laten we recente ontwikkelingen en initiatieven op het vlak van beheer en gebruik van onderzoeksdata en andere soorten gegevens aan de orde komen. Deze eerste aflevering in de serie geeft een algemene inleiding op het thema, waarbij de diverse invalshoeken kort worden belicht. De hier aangestipte facetten zullen in de loop van deze serie allemaal uitgebreider aan de orde komen. De te behandelen voorbeelden en initiatieven benadrukken dat data in toenemende mate tot het werkterrein van de informatie-professional behoren.

Soorten datacollecties

De data waar het in deze serie om gaat, kunnen van allerlei aard zijn. Het kan gaan om onderzoeksdata, de gegevens die primaire resultaten zijn van allerlei soorten onderzoek. Deze onderzoeksdata moeten op een of andere manier worden opgeslagen, beheerd en toegankelijk gehouden (of gemaakt). Daarnaast zijn er situaties waar organisaties gebruik willen maken van gegevens die door andere, vaak commercieel opererende organisaties zijn geselecteerd. Daarbij speelt de vraag hoe men deze datasets *collectieert*, hoe men toegang daartoe regelt en hoe die beheerd moeten worden. Een derde categorie data ten slotte zijn die welke (veelal vrij) op internet beschikbaar zijn. Het toekennen van betekenis aan die gegevens, opdat ze door computerprogramma's gebruikt kunnen wor-



Drie fases in de levenscyclus van onderzoeksdata (naar 'Australian National Data Service')

den, wordt steeds beter mogelijk dankzij zogenaamde *linked data*-initiatieven. Hier wat meer over die drie invalshoeken.

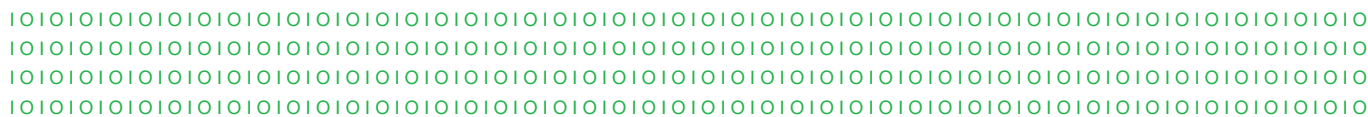
Primaire onderzoeksdata

In deze serie zal de meeste aandacht uitgaan naar onderzoeksdata, omdat er juist op dat terrein veel initiatieven zijn, waarbij bovendien bibliotheken en informatiecentra vaak een centrale rol spelen. Al geruime tijd richten universiteiten, onderzoeksinstituten en hogescholen institutionele repositories in voor het digitaal beschikbaar stellen, het blijvend opslaan en het met anderen uitwisselen van publicaties uit de eigen instelling. Die publicaties zijn meestal geba-

seerd op bij die instelling uitgevoerd onderzoek. In toenemende mate realiseert men zich dat het belangrijk is ook die achterliggende primaire onderzoeksdata te bewaren en eventueel beschikbaar te stellen. Tijdschriftredacties en subsidiërende instanties stellen soms zelfs al de eis dat die gegevens ergens beschikbaar moeten zijn. Vanuit digitale publicaties moet er dan met een permanente URL naar gelinkt kunnen worden. Dat kan aanleiding geven tot wat wel verrijkte publicaties en complexe objecten worden genoemd. Op gestructureerde wijze moeten daarin niet alleen multimediale illustraties, maar ook datasets aan teksten gekoppeld kunnen worden.

Curation continuum

Voortbordurend op bemoeienis met de institutionele repositories, probeert SURFfoundation in Nederland ook op dit terrein een voortrekkersrol te spelen, met zijn Onderzoeksdata Forum. In internationaal verband komen al langer rapporten uit die het belang van beheer van datacollecties benadrukken (1,2). Men spreekt daarbij wel van *research data curation*. Een onderzoek van de Australische Monash University laat zien dat in de loop van de levenscyclus van onderzoeksdata de rol, het gebruik en het daaruitvolgende beheer van die data geleidelijk veranderen. Men spreekt in dat verband van een *curation continuum*. Om dat wat modelmatig te



kunnen beschrijven, hebben zij echter ook *curation boundaries* aangegeven, de punten waar data duidelijk een andere fase ingaan (3).

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 gemeta-dateerd, dat die partners er ook toegang toe hebben en hun betekenis kunnen interpreteren. In een derde fase komen de onderzoeksdata helemaal in het publieke domein: in principe kan iedereen er gebruik van maken, kan er naar gelinkt worden en moeten ze ook duurzaam bewaard worden.

Deze drie fases zult u in komende afleveringen zeker tegenkomen. 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 (Data Archiving & Networked Services van de KNAW) is in het leven geroepen voor die tweede en (vooral) derde fase, waar hergebruik en digitale duurzaamheid vooropstaan.

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 experimenten met de *Large Hadron Collider*, de deeltjesversneller van CERN in Genève, gaan opleveren. Of aan de gegevens die uit gekoppelde astronomische waarnemingssystemen komen. In deze grote projecten wordt bij de opzet meteen al rekening gehouden met de massale datastromen die ze gaan opleveren. Toch is het niet alleen op het terrein van de bèta-wetenschappen dat er sprake is van grote datahoeveelhe-

den. Hogeresolutie- en hogesnelheid-multimedia en de daarbij horende grote datastromen vinden bijvoorbeeld ook ingang in de humaniora en sociale wetenschappen. In sommige kringen begint men zich dan ook te realiseren dat datamanagement van te verwachten onderzoeksdata een noodzakelijk onderdeel van elke projectaanvraag zou moeten zijn.

Datathecarissen en datavocaten

Voor onderzoek wordt niet alleen gebruik gemaakt van eigen datasets of van die van collega's waarmee men samenwerkt. Voor bijvoorbeeld economisch onderzoek heeft men vaak allerlei door andere organisaties verzamelde en gevalideerde gegevens nodig. Voor beheer van zulk gecollectieerd en (vaak) ingekocht materiaal is bij de bibliotheek van de Erasmus Universiteit een Data Centrum opgezet. Dat initiatief, evenals de eerder beschreven ontwikkelingen, maken duidelijk dat behoefte ontstaat aan een nieuw soort functionaris, een *data librarian*, zoals bij de Universiteit Tilburg is aangesteld. Misschien moeten we daar maar eens een goed Nederlands woord voor introduceren: de datathecaris (een term die in Google nog niet te vinden is). Op taken en competenties van zo'n functionaris zullen we in deze serie ook ingaan. Waar beheerders van datasets in elk geval mee te maken krijgen zijn juridische aspecten. Dat betreft niet alleen het auteursrecht op die data, maar bijvoorbeeld ook privacy aspecten in gevallen waar datasets gegevens over personen bevatten. Een van de auteurs van een juridische wegwijzer zal ons daarover bijpraten. (4)

Linked data

Intussen zijn ook op internet grote dataverzamelingen beschikbaar. Dat betreft niet specifiek (en vaak zelfs helemaal niet) onderzoeksdata, maar allerlei soorten gestructureerde gegevens in het algemeen. Opdat computers die gegevens op standaardwijze kunnen lezen, inter-

preteren 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 Data om deze dataverzamelingen te omschrijven. In RDF worden alle gegevens beschreven als zogenaamde tripels 'subject - predicaat - object', wat wil zeggen dat zo wordt aangegeven dat *iets* (een subject) een eigenschap (predicaat) heeft, waaraan een waarde (object) wordt toegekend; bijvoorbeeld 'dit_artikel - is_geschreven_door - eric_sieverts'. Beschikbare bronnen voor linked data worden daarom wel *RDF triple stores* genoemd. Op de site [Linkeddata.org](http://linkeddata.org) is een indrukwekkende wolk van dergelijke onderling gelinkte datacollecties gevisualiseerd. Ook daar is intussen sprake van een *deluge* van miljarden beschikbare RDF-tripels en varianten daarop. Vooral Tim Berners Lee is een grote stimulator van het zo beschikbaar stellen van datacollecties. Die ziet hij namelijk als de ruggengraat voor het semantisch web. Tamelijk centraal in de linked-data-wolk staat de DBpedia, 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 ontleen. De letters DB in DBpedia zijn een aanwijzing dat het hier eigenlijk om een database-achtige extensie van het web gaat. Ook de VIAF, de *virtual international authority file*, waarin een groot aantal bibliotheken via OCLC samenwerkt, is intussen als linked data beschikbaar. Linked data is niet meer alleen een speeltje voor ICT'ners, maar kan ook van nut worden voor bibliotheken.

Tot zover deze algemene inleiding. De komende afleveringen zullen hopelijk nog duidelijker mijn stelling illustreren dat ook voor informatieprofessionals een rol is weggelegd bij het steeds belangrijker beheer en gebruik van gegevensverzamelingen. <

Literatuur

- 1) e-IRG report on data management (2009). European Strategy Forum on Research Infrastructures. 95 blz. (<http://is.gd/clyMA>)
- 2) Insight into digital preservation of research output in Europe (2009). PARSE Insight. 83 blz. (<http://is.gd/clz8F>)
- 3) Data curation continuum (2010). Australian National Data Service. (<http://is.gd/clzKF>)
- 4) De juridische status van ruwe data; een wegwijzer voor de onderzoekspraktijk (2009). Stichting SURF. 54 blz. (<http://is.gd/clzJf>)

Eric Sieverts is redacteur van InformatieProfessional.



Thema's



Een voorlopig overzicht van te verschijnen bijdragen:

- > Linked data
- > Taak en competenties van de data librarian
- > Modellen en standaarden voor beheer van datasets
- > Een universitair datacentrum
- > DANS: data archiving, duurzaamheid, persistente identifiers
- > Infrastructuur voor zelf-archiveren van datasets
- > Met datasets verrijkte publicaties
- > 3TU, een datarepository voor technische wetenschappen
- > De juridische status van ruwe data.

