Research Data Management en de Vlaamse Universiteiten: White Paper

January 1, 2018

VLIR Werkgroep Research Data Management & Open Science

Managementsamenvatting:
Open Science aan universiteiten en de dringende nood tot actie

" Duidelijke keuzes voor de economie van de toekomst, met eenvoudige structuren en instrumenten en de juiste competenties."

(toelichting beleidsbrief Economie, Wetenschap & Innovatie door minister Muyters, 9/11/217, minuut 20.58-21.12)

De economie van de toekomst is data-driven. Diegene die over data beschikt en de skills heeft om ermee aan de slag te gaan, kan als eerste trends distilleren en hierop nieuwe ontwikkelingen baseren. Dat geldt bij uitstek voor data uit onderzoek.

De Europese Commissie speelt hier volop op in, getuige hun visietekst *Open Innovation, Open Science & Open to the World*. Open Science is een paradigmashift, mogelijk gemaakt door technologische ontwikkelingen en gedreven door grote, vaak globale maatschappelijke uitdagingen. De EU ondernam daartoe concrete initiatieven. Voor EU gefinancierde projecten worden onderzoekers bij default verwacht de onderzoeksdata open ter beschikking te stellen. De European Open Science Cloud (EOSC) wordt een belangrijk project in deze missie en heeft als objectief alle data-silo's die conform zijn aan een aantal operationele en kwalitatieve principes te linken en voor onderzoek ter beschikking te stellen. De Europese Commissie schat namelijk dat tegen 2020 Open Data een marktwaarde zal hebben van 75 miljard euro voor de EU28+ en dat dit zich zal vertalen naar een nood aan 100.000 data-experten. Kwaliteitsvolle en duurzame voorzieningen rond onderzoeksdatabeheer en data delen hebben een aantoonbare efficiëntiewinst die binnen de EU28+ geraamd wordt op een geaccumuleerd bedrag van maar liefst 1.7 miljard euro tegen 2020.

Voor de academische instellingen betekent Open Science gestalte geven aan een beleid dat de ruimere schakel van onderzoekshandelingen valoriseert en erkent, research datamanagement meer centraal stelt en een passend onderwijsaanbod initieert in functie van de toekomstig benodigde skills. Vlaamse universiteiten stellen momenteel alles in het werk om daaraan het hoofd te bieden. Via de VLIR werkgroep RDM & OS (Research Data Management & Open Science) werden de krachten gebundeld. Een direct resultaat was de oprichting van een consortium voor het aanbieden van DMP-online, software voor de geleide opmaak van een datamanagement plan (DMP). Daarnaast toonden de uitkomsten van een gemeenschappelijk uitgevoerde survey onder Vlaamse onderzoekers de hoogdringendheid aan om te handelen zowel op het niveau van opleiding als investering in infrastructuur. De urgentie is ook ingegeven door diverse nieuwe regelgevingen, zoals de Algemene Verordening Gegevensbescherming (AVG/GDPR), die voor data-intensieve sectoren, zoals wetenschappelijk onderzoek, een grote en directe impact hebben. Op dit moment stoten de universiteiten en andere onderzoeksinstellingen echter op een plafond in capaciteit en voorzieningen om de grote uitdagingen waarvoor gesteld gezamenlijk aan te pakken.

Het Vlaamse wetenschappelijke onderzoek aan de kennisinstellingen heeft altijd al een sterke rol gespeeld in innovatieondersteuning en verankering van bepaalde industriële sectoren in de regio (elektronica, farma, chemie, …). Ook binnen Europa staat Vlaanderen genoteerd als een sterke innovator. Wil Vlaanderen als regio kapitaliseren op de mogelijkheden die data-gedreven onderzoek en technologie in de nabije toekomst zal bieden, dan moeten de nodige bijkomende inspanningen worden gedaan in het kader van een nieuw (Open) Onderzoeksdata beleid. Daarom worden door de academische instellingen, in gemeenschap, de volgende **korte termijn actiepunten** voorgesteld:

1. Investeer in het opzetten van **infrastructuur** aan de academische instellingen voor het ter beschikking stellen van onderzoeksgegevens aan derden, met aandacht voor specifieke licenties, auteursrecht en gebruiksindicatoren, en met garantie van interoperabiliteit met de EOSC. Deze infrastructuur dient gesteund op een duurzaam businessmodel en vergelijking met internationale modellen, waarbij de verhouding tussen centraal en decentraal aanbod wordt geoptimaliseerd.

2. Investeer in een hoogstaande gespecialiseerde **opleiding** van datamanagers zodat in de nabije toekomst de passende profielen op de arbeidsmarkt aanwezig zijn, zowel voor jobs aan de universiteiten als bij overheden en KMO's. Data skills zouden ook een intrinsiek onderdeel van de Bachelor/Masteropleiding moeten vormen en idealiter al in het secundaire onderwijs worden geïntroduceerd.

3. Indien ervoor wordt geopteerd om de Europese Algemene Verordening Gegevensbescherming (GDPR) om te zetten naar nationale wetgeving, moet erover worden gewaakt dat er geen bijkomende restricties voor onderzoek worden gecreëerd. Daarnaast moet er ook actieve steun gegeven worden aan de Europese hervorming van het auteursrecht en de bepalingen rond tekst- en data-mining.

4. Zet, net als de Europese Commissie, in op het **waarderen van alle schakels in het onderzoeksproces**, niet enkel het eindproduct, m.n. de publicaties, maar ook datasets, software en de mate van openheid van deze output. Hierbij is het evenwel belangrijk dat verplichtingen door funders of overheden rekening houden met mogelijkheden tot praktische implementatie aan de academische instellingen. Beleid en investeringen dienen hand in hand te gaan.

Sense of urgency

Onderzoeksdata vormen het kloppende hart van wetenschappelijk onderzoek en zijn de motor van de enorme vooruitgang op technologisch, sociaaleconomisch, maatschappelijk en gezondheidsvlak die generaties van innovatief onderzoek hebben mogelijk gemaakt. Hoewel onderzoekers de data met de grootste precisie behandelen voor toetsing aan de onderzoekshypothese, worden deze nadien veelal niet duurzaam bewaard. Tegelijk wordt de samenleving gekenmerkt door een toenemende dataficatie van de samenleving, waarbij alsmaar meer handelingen en processen in datastromen worden gevat. Dit maakt een specifiek type van data-gedreven onderzoek mogelijk, gesteund op de combinatie van (big) data met nieuwe analysetechnieken ('data science'). Het economische en wetenschappelijke belang van deze voorheen ongekende stroom aan data is groot - occasioneel wordt gesproken over the new paradigm for science - en vraagt om een gecoördineerde aanpak met betrokkenheid van alle stakeholders.

'Data is the new gold, just as oil was likened to black gold' Nelie Kroes

De lat voor de goede omgang met onderzoeksdata, het zogenaamde onderzoeksdatamanagement of research data management (RDM), ligt momenteel hoger dan ooit. Zo vereist een aantal nieuwe wetten en verordeningen een veel striktere omgang met persoonsgegevens, terwijl anderzijds de belangrijkste Europese en Amerikaanse financiers en uitgevers van wetenschappelijk onderzoek een steeds betere archivering van onderzoeksdata opleggen als voorwaarde voor financiering of publicatie, in vele gevallen zelfs in een vrij toegankelijke vorm om zo het hergebruik van deze data maximaal te stimuleren (*open data).* Daarnaast vragen steeds meer financiers een uitgebreide documentatie van het geplande en gevoerde RDM in de vorm van een gedetailleerd *datamanagementplan.* De wetenschappelijke publicatiecultuur krijgt daarmee een extra dimensie waar data en resultaat worden gekoppeld, wat open review initieert en de grondslag legt voor een nieuwe wijze van onderzoek waarbij de ene vinding nog dichter aanleunt bij de andere vinding. Datapublicatie wordt hierbij een na te streven objectief en het beheer van data zal voor de onderzoeker meer dan ooit een parallel traject worden dat aandacht vereist op de verschillende momenten van het volledige onderzoeksproces.

Deze ontwikkelingen stellen de academische instellingen voor onmiddellijke en significante uitdagingen die met de huidige middelen niet kunnen worden opgevangen. Wil Vlaanderen internationaal toonaangevend blijven en aansluiting vinden bij landen en regio's die de leiding nemen inzake data-gedreven onderzoek, dan zal een state-of-the-art onderzoeksdatamanagement een noodzakelijke voorwaarde zijn. Het onderzoek aan onze universiteiten steunt op de jaarlijkse productie van miljoenen gigabytes aan onderzoeksdata. Het goed omgaan met en beheren van deze enorme rijkdom vergt een nieuwe portfolio aan ondersteuning, gaande van het louter infrastructurele om de dataopslag materieel mogelijk te maken, het opmaken van een passend RDM beleid, het uitwerken van een juridisch en ethisch kader en niet in het minst het opleiden van onderzoekers daarin. Nieuwe types van onderzoek, steunend op bijzonder data-intensieve en vernieuwende technieken, zullen zowel in curricula als in de onderzoeksstrategieën moeten worden ingepast.

Deze white paper wenst vanuit een bevraging bij de universitaire instellingen de huidige stand van zaken in kaart te brengen en een beeld te krijgen van de noden en verwachtingen. Daarop steunend worden een aantal aanbevelingen geformuleerd die een noodzakelijke voorwaarde vormen om het RDM beleid aan de universiteiten verder gestalte te geven en als hefboom te fungeren in een data-intensieve onderzoekswereld.

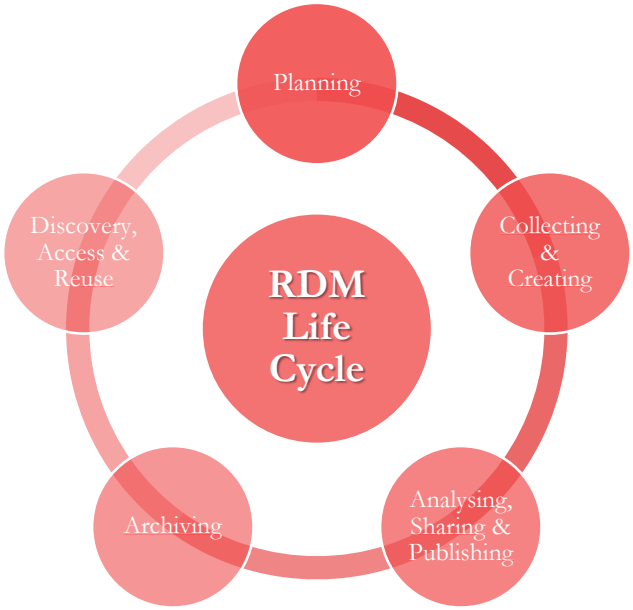
Onderzoeksdata: een brede definitie

Onderzoeksdata zijn alle digitale of fysieke data – onafhankelijk van de manier waarop deze data worden verzameld of bewaard – die gebruikt of geanalyseerd worden voor het ondersteunen van onderzoeksbevindingen, het valideren van onderzoeksresultaten of onderliggend zijn aan een redenering, discussie of berekening in het onderzoek. Onderzoeksdata omvatten het ganse spectrum van ruwe data tot de bewerkte en geanalyseerde data opgenomen of besproken in een publicatie. Onder deze data worden zowel gegenereerde data, afgeleide of samengestelde data verstaan. Dit kunnen zowel eigen gegenereerde data zijn als data ter beschikking gesteld door derden. Voorbeelden van onderzoeksdata zijn survey resultaten, statistieken, metingsresultaten, notebooks, afbeeldingen, computer gegenereerde data, simulaties, software ontwikkeld voor onderzoeksdoeleinden, computationele metadata, prints, video- en audiotapes, organismen, gensequenties, synthetische verbindingen, stalen van welke aard ook, patiëntengegevens, e.a.. Uit de enquête aan de Vlaamse universiteiten bleek dat de onderzoeksdata die Vlaamse onderzoekers verzamelen een grote variëteit bezitten en dit zowel voor wat betreft de aard (fysieke vs. digitale data), de typering -primaire (eigen) vs. secundaire (verkregen) data- alsook het desbetreffende data bestandsformaat. Deze verscheidenheid zorgt er vanzelfsprekend voor dat research data management maatwerk betekent, wil men elke onderzoeksdataset kwalitatief en duurzaam bewaren voor toekomstig onderzoek.

Wat is research data management?

Research data management gaat in essentie over het kwaliteitsvol omgaan met deze onderzoeksdata, op elk moment in de levenscyclus van data (zie figuur 1) met als doel om, in overeenstemming met de geldende en contractuele regelgeving, onderzoeksresultaten betrouwbaar te verifiëren en nieuw, innovatief onderzoek mogelijk te maken op basis van bestaande onderzoeksdata.

De levenscyclus van onderzoeksdatamanagement beschrijft de verschillende fasen die onderzoeksdata achtereenvolgens kunnen doorlopen tijdens een onderzoeksproject. Hierbij dient niet steeds elke fase noodzakelijkerwijze doorlopen te worden. Tijdens de *planningsfase*, die aanvangt vóór de start van een onderzoeksproject, maken onderzoekers een gedetailleerd plan op waarbij ze beschrijven hoe ze de diverse facetten van onderzoeksdatamanagement zullen uitvoeren in de loop van hun onderzoeksproject.



Figuur 1. Levenscyclus onderzoeksdatamanagement

Bij aanvang van het onderzoeksproject zullen vervolgens nieuwe data gegenereerd worden die a priori nog niet (samen) bestonden. De *data collecting & creating fase* gaat over hoe en in welke tools data opgeslagen wordt met inachtname van de bewaking van de toegang en veiligheid van de data. Vervolgens worden de data verwerkt en gebruikt om de vraagstelling te toetsen. De *analysing, sharing & publishing fase* verzamelt de beschrijvingen hoe de data-analyses uitgevoerd werden, welke manipulaties de datafiles ondergingen m.i.v. een versiebeheer. In deze fase worden tevens de formaten voor publicatie en sharing bepaald rekening houdend met eventuele veiligheids- en confidentialiteitsbepalingen. Vervolgens worden de data gekopieerd naar een opslaglocatie en worden ze verwijderd uit de zogenaamde actieve productie-omgeving, wat ook wel de *data archiving fase* genoemd wordt. In deze fase moet nagedacht worden over het medium waarop de data alsook de metadata het best worden opgeslagen, met inbegrip van het voorzien van de noodzakelijke back-ups. De onderzoeksdatalevenscyclus wordt eventueel beëindigd met de *data discovery, access & reuse fase*, waarbij data beschikbaar gesteld kunnen worden voor hergebruik in een nieuwe onderzoeksvraag. Idealiter voldoet deze data aan de FAIR principes.

FAIR data

Volgens de FAIR principes, geformuleerd door FORCE 11 (force11.org), moeten data:

- Findable (vindbaar): Gemakkelijk te vinden zijn door zowel mensen als computersytemen door middel van de verplichte beschrijving van metadata die de ontdekking van interessante datasets toelaat;

- Accessible (toegankelijk): Duurzaam bewaard worden op een dergelijke lange termijn zodanig dat de data gemakkelijk toegankelijk zijn en/of gedownload kunnen worden met een goed gedefinieerde licentie en toegankelijkheidscondities (Open Access waar mogelijk) voor wat betreft de metadata, of op het niveau van de eigenlijke data inhoud;

- Interoperable (interoperabel): Gemakkelijk gecombineerd kunnen worden met andere datasets door mensen en computersystemen;

- Reusable (herbruikbaar): Klaar zijn om herbruikt te worden in toekomstig onderzoek en dus om verder geprocessed te worden met computationele methoden.

Historiek en context

De omvang van de net geschetste uitdagingen vraagt om een prioritaire aanpak op diverse beleidsniveaus steunend op een gecoördineerd beleid. Het is de uitgesproken ambitie van de Vlaamse universiteiten om op het gebied van onderzoeksdatamangement zo veel mogelijk gezamenlijke beleidsvoorstellen en ondersteunende acties te initiëren, vertrekkend van overleg, aftoetsen en van elkaar leren. Met dit doel werd in 2015 door medewerkers van alle Vlaamse universiteiten een werkgroep opgericht om expertise en ervaring te delen rond concrete vooraf bepaalde probleemstellingen. Hiermee werd bewust gekozen voor een bottom-up aanpak, vertrekkende van de personen op wie de concrete uitwerking grotendeels rust. In 2017 kreeg de samenwerking een plaats in de formele overlegstructuren van de Vlaamse Interuniversitaire Raad (VLIR). Als vernieuwde werkgroep RDM en Open Science werd ze een beleidsadviserend en –voorbereidend instrument zowel naar de individuele instellingen als naar ander stakeholders.

Zorgvuldig databeheer start bij een degelijke planning, en binnen de internationale onderzoekswereld worden datamanagementplannen (DMP) steeds meer beschouwd als een onderdeel van een goede onderzoekspraktijk voor elk project dat data gebruikt of genereert. Daarom vraagt een groeiend aantal publieke onderzoeksfinanciers om een DMP te voorzien kort na de toekenning van de funding. De werkgroep vormt hier een aanspreekpunt voor de Vlaamse financiers die een datamanagentplan wensen in te sluiten. In navolging van andere financiers onderneemt ook het FWO stappen om aandacht te vragen voor een correct beheer van onderzoeksgegevens, en om inzicht te krijgen of en op welke manier data voortkomend uit door hen gefinancierde projecten en mandaten uiteindelijk ter beschikking van derden worden gesteld. De VLIR WG RDM & OS verleent hierbij advies. In concreto werd voorgesteld om incrementeel te werk te gaan en te beginnen met het invoegen van een aantal generieke vragen rond databeheer bij de aanvraagprocedure voor projecten. Het opstellen en voorleggen van een DMP volgt dan na toekenning van het project en binnen een termijn van 6 maanden. Binnen de VLIR WG RDM & OS wordt een voorstel voor een FWO-DMP template uitgewerkt. Implementatie van dit voor FWO nieuwe beleid is momenteel voorzien voor de projectronde 2018.

Onder de concrete realisaties binnen de schoot van de werkgroep behoort ook de oprichting van *DMPBelgium* in 2017. Naar analogie van buitenlandse voorbeelden beoogt dit consortium de lokale implementatie van de open-source DMPonline software, die ontwikkeld werd door het Digital Curation Centre in het Verenigd Koninkrijk. Via het gemeenschappelijke portaal DMPonline.be krijgen onderzoekers toegang tot een tool die hen op een gestructureerde wijze begeleidt bij de opmaak van een datamanagementplan via software die wordt gehost op de instelling-neutrale servers van BELNET. In hetzelfde jaar besliste KU Leuven deze software lokaal te hosten, na gedurende een jaar gebruik gemaakt te hebben van het online platform van DCC zelf. Beide toepassingen staan open voor externe partners, maar DMPonline.be is intussen uitgegroeid tot een consortium van universiteiten en onderzoeksinstellingen die zich verzameld hebben met het oog op het realiseren van schaalvoordelen. Het consortium bestaat op heden uit de vier Vlaamse universiteiten (behalve KU Leuven), de ULB, het INBO en het WIV, en staat in voor het beheren van de applicatie. Deze case is een voorbeeld van hoe in een gemeenschappelijk aanbod van software en infrastructuur schaalvoordelen te halen zijn.

Het uitbouwen van een toekomstig beleid vertrekt vanuit inzicht in de huidige praktijken en noden aan de instellingen. Daarom heeft de werkgroep beslist om in de

Een academisch RDM beleid gesteund op samenwerking, realiseren van schaalvoordelen en leren van elkaar

beginfase een bevraging uit te voeren bij onderzoeksleiders en onderzoekers over alle wetenschappelijke disciplines. Deze steunde op deelname van alle universiteiten, met uitzondering van Universiteit Gent waar de antwoorden van een eerdere bevraging rond RDM werden gerelateerd aan de huidige survey. Met het oog op vergelijkbaarheid van de resultaten is de vraagstelling geënt op belangrijke enquêtes die in het buitenland zijn uitgevoerd, zoals de Southampton Data Survey en de Austrian National Research Data Survey. De bevraging liep over verschillende dimensies:

1. *Data formats types*: deze vragen beogen een aflijnen van de data definities
2. *Data archivering, storage, back-up en loss*: doel hierbij is om inzicht te verwerven in de huidige procesmatige omgang met onderzoeksdata en de implicaties daarvan
3. *Ethical en legal issues*: hier wordt de primaire kennis getoetst in deze materie
4. *Infrastructure and services*: beoogt het in kaart brengen van infrastructurele noden voor onderzoek
5. *Collaboration and reuse*: deze vragen zijn gericht op het verwerven van inzicht in de al dan niet collectieve wijze van dataverzameling en de houding ten aanzien van hergebruik en openheid

De survey liep van maart tot juli 2016 en bevroeg zowel onderzoeksleiders (n = 224) over de praktijken binnen hun onderzoekseenheid als meer junior onderzoekers (n = 425) in detail over hun RDM. De antwoorden werden geanonimiseerd en de affiliaties per universiteit werden geaggregeerd tot op het niveau van het wetenschapsgebied (Biomedische Wetenschappen, Wetenschappen & Technologie, Sociale & Humane Wetenschappen). Opvallend aan de surveyresultaten was de grote homogeniteit over de instellingen heen zowel wat betreft de aanduiding van de pijnpunten als wat betreft de formulering van de noden. Ook tussen onderzoeksleiders en de individuele vorsers zijn geen significante verschillen merkbaar, noch tussen verschillende disciplines, hoewel sommige deelgebieden duidelijk een grotere vertrouwdheid hebben met de organisatie van het datamanagementproces.

Deze white paper volgt in grote lijnen de surveyindeling en geeft voor elke dimensie de belangrijkste bevindingen weer. Toch zijn de algemene aanbevelingen een combinatie van de inzichten die volgen uit de bevraging en de ervaringen van de instellingen met de huidige implementatie van een datamanagement beleid.

De complexe realiteit van wettelijke & ethische bepalingen

In het afgelopen decennium zijn er tal van richtlijnen en wettelijke bepalingen in werking getreden die research data management zowel op Europees niveau als op nationaal niveau hoog op de agenda plaatsen en de krijtlijnen ervan bepalen. Zo ziet de Europese Commissie via het *Responsible Research & Innovation*-programma[1] en het "*Open Science*"-kader, toe op de transparantie en toegankelijkheid van wetenschappelijke publicaties, de naleving van gedragscodes en het respect voor de wetenschappelijke integriteit van Europese onderzoeksprojecten. Daarnaast verscherpen de Europese Algemene Verordening Gegevensbescherming (AVG of *General Data Protection Regulation, GDPR*), het Protocol van Nagoya en de Controle op handel in goederen voor tweeërlei gebruik (zgn. *dual use*-producten) het juridisch kader voor onderzoeksdata.

In het kader van de AVG[2] zullen organisaties die persoonsgegevens verwerken, waaronder ook de universiteiten en de onderzoeksinstellingen, vanaf 25 mei 2018 duidelijk moeten kunnen aantonen welke persoonsgegevens zij verwerken, waarom zij deze persoonsgegevens verwerken en welke maatregelen worden getroffen om de privacy van deze gegevens en de rechten van de individuen te bewaken. Dit heeft tot gevolg dat de software en andere technische tools die persoonsgegevens verzamelen en/of verwerken binnen de universiteiten en onderzoeksinstellingen up to date moeten zijn om te voldoen aan deze verordening.

Ook voor andere levende organismen is er een strenge bescherming ontstaan. Onderzoekers die planten, dieren, microben of afgeleiden daarvan in het buitenland verzamelen en/of gebruiken, zullen moeten voldoen aan de Verordening 511/2014 en het Protocol van Nagoya.[3] Onderzoekers die dergelijke genetische rijkdommen verwerven en/of gebruiken zijn verplicht om een "passende zorgvuldigheid" (zgn. *due diligence*) te betrachten opdat ze de genetische rijkdommen en de traditionele kennis met betrekking tot deze genetische rijkdommen verwerven en gebruiken in overeenstemming met de regels inzake toegang en verdeling van voordelen die het land van levering heeft vastgesteld.

Als het onderzoek kennis, materialen, methoden en technologieën genereert die mogelijks ook ingezet zouden kunnen worden voor onethische doeleinden, waar de zogenaamde "Dual use"-verordening[4] van toepassing is, moet de onderzoeker ook deze naleven en indien van toepassing de nodige vergunningen aanvragen bij de Vlaamse overheid.

Op regionaal niveau treedt het Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) in de voetsporen van de Europese Commissie voor wat research data management betreft. Zo werkt het FWO momenteel samen

[1] Zie hiervoor https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation.
[2] Verordening (EU) nr. 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (Algemene Verordening Gegevensbescherming), http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32016R0679&from=NL.
[3] Verordening (EU) nr. 511/2014 van het Europees Parlement en de Raad van 16 april 2014 betreffende voor gebruikers bestemde nalevingsmaatregelen uit het Protocol van Nagoya inzake toegang tot genetische rijkdommen en de eerlijke en billijke verdeling van voordelen voortvloeiende uit hun gebruik in de Unie, http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32014R0511&from=NL.
[4] Verordening (EG) nr. 428/2009 van de Raad van 5 mei 2009 tot instelling van een communautaire regeling voor controle op de uitvoer, de overbrenging, de tussenhandel en de doorvoer van producten voor tweeërlei gebruik, http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32009R0428&from=NL.

met de Vlaamse universiteiten (in de schoot van de VLIR) om een Vlaams model voor een datamanagementplan te ontwikkelen. Vanaf de oproepronde 2018 zal het FWO met enkele korte vragen in de projectaanvragen reeds peilen naar de belangrijkste aspecten van RDM, in een volgende fase zal een volwaardig DMP dienen aangeleverd te worden voor toegekende onderzoeksprojecten Bovendien heeft het FWO maatregelen omtrent wetenschappelijke integriteit ingeschreven in haar profielen voor onderzoekers, promotoren en instellingen. Deze maatregelen bepalen onder meer dat onderzoekers de gegevens m.b.t. het onderzoek veilig en duurzaam dienen te bewaren, dit rekening houdend met de eigenheden van het vakgebied en de aard van het onderzoek.[5]

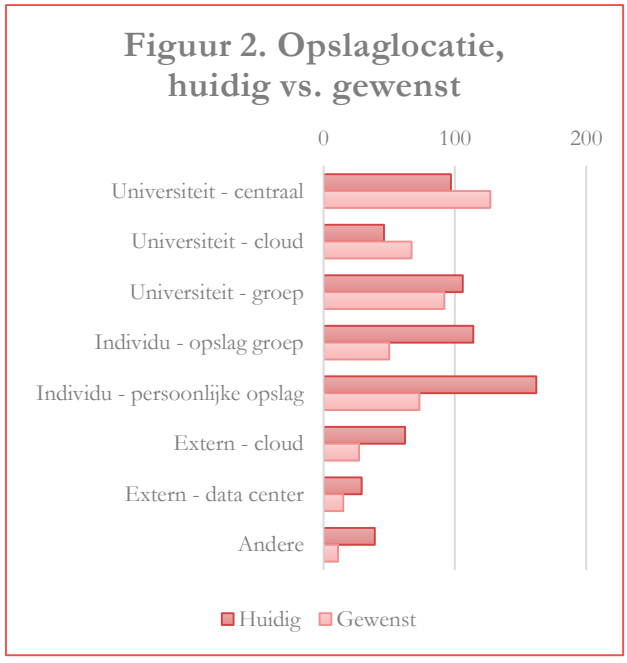
In de praktijk betekent dit dat de onderzoeker voor elk onderzoeksvoorstel moet kunnen aantonen dat alle toepasselijke juridische en ethische vereisten zullen nageleefd worden en dat hij/zij, voor zover mogelijk, dit ook opneemt in zijn/haar datamanagementplan. Voor onderzoek met betrekking tot proefdieren en/of mensen zal de onderzoeker bijvoorbeeld een gunstig advies van een daartoe bevoegde Ethische Commissie moeten kunnen voorleggen om aan te tonen dat hij/zij de Ethische Code voor het Wetenschappelijk Onderzoek en geldende wetgeving binnen zijn/haar onderzoek zal respecteren.[6] Aanvullend bepaalt deze Ethische Code dat de onderzoeker de primaire gegevens en de protocollen van het onderzoek gedurende een voldoende lange tijd en op een toegankelijke wijze bewaart. Wanneer publicaties, waaronder vooral reviews en syntheses, niet alle voor verificatie noodzakelijke details omvatten moeten deze wel beschikbaar zijn bij de onderzoeker zelf.[7] Deze ethische verplichtingen verschijnen op Europees niveau in de recent herziene 'European Code of Conduct for Research Integrity' van ALLEA, welke een specifiek hoofdstuk wijdt aan research data management.[8] Als onderzoekers bovendien persoonsgegevens verzamelen en/of verwerken binnen hun onderzoek moeten zij de privacywetgeving respecteren en de nieuwe Europese Algemene Verordening Gegevensbescherming naleven.

Het amalgaam aan wettelijke en ethische verplichtingen maakt het voor onderzoekers niet eenvoudig. Dit is niet alleen een grote uitdaging voor onderzoekers, maar ook voor onderzoekinstellingen, financiers, uitgevers en overheden. Bovendien is het voor de onderzoeker niet steeds evident om op alle wettelijke verplichtingen in te spelen, niet in het minst omdat vaak meerdere spelers de juridische spelregels bepalen. Uit de survey blijkt bijvoorbeeld dat momenteel 55% van de onderzoekers aangeeft een overzicht te hebben van de verschillende eigenaars van hun onderzoeksdata. Daarenboven lopen de antwoorden over de eigenaarsrollen zeer sterk uiteen. De regels rond intellectuele eigendomsrechten zijn op zich niet nieuw, maar toch ontbreekt er een eensgezinde visie onder onderzoekers. De bevraagde onderzoekers geven aan dat zij onzeker zijn over het spanningsveld tussen enerzijds het maken van interne afspraken omtrent intellectuele eigendom en de bescherming van gevoelige onderzoeksgegevens, en anderzijds het moeten respecteren van specifieke vereisten met betrekking tot het openstellen van onderzoeksdata.

[5] Zie hiervoor: http://www.fwo.be/nl/het-fwo/organisatie/wetenschappelijke-integriteit/in-het-fwo/
[6] Zie hiervoor: http://www.belspo.be/belspo/organisation/publ/Eth_code_nl.stm.
[7] Zie hiervoor http://www.fwo.be/nl/het-fwo/organisatie/wetenschappelijke-integriteit/in-het-fwo/
[8] Zie hiervoor: http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf

Lacunes in infrastructuur

Goed datamanagement vereist een degelijke infrastructuur, zowel voor het plannen van datamanagement, de productie en het beheer van dynamische onderzoeksdata (d.w.z. data in actief gebruik), als voor de duurzame bewaring en beschikbaarstelling van daarvoor geselecteerde data uit afgerond onderzoek. Bepaalde componenten van deze infrastructuur zijn reeds aanwezig binnen en buiten de Vlaamse universiteiten. Zo kan gebruik gemaakt worden van externe domein-specifieke en general-purpose data repositories en –archieven (met name in het buitenland), en bieden de universiteiten tools voor datamanagement planning alsook centrale opslag- en computing faciliteiten aan. Toch zullen de groeiende datavolumes en digitale evoluties de instellingen voor grote uitdagingen stellen, en beschikt lang niet elk onderzoeksdomein of datatype over een gepaste oplossing. In veel gevallen zijn het bijvoorbeeld de onderzoekers zelf die de (langdurige) opslag en beschikbaarstelling van de door hen gegenereerde of verzamelde onderzoeksdata organiseren, hetgeen doorgaans weinig garanties biedt op het vlak van informatieveiligheid, duurzaamheid, beschikbaarheid en bruikbaarheid. Met name het voor de toekomst bewaren, beschikbaar en bruikbaar houden van digitale onderzoeksdata is immers een complexe aangelegenheid die veel meer vergt dan het louter opslaan van gegevens (gezien o.a. de inherente kwetsbaarheid en veroudering van digitaal materiaal). Zonder de juiste aanpak en bewaaromstandigheden is het risico op dataverlies, op onvindbare of onbruikbare data, dan ook reëel, terwijl duurzame toegang tot data net uitdrukkelijk verwacht wordt door onderzoeksfinanciers. Onderzoeksdata blijven immers vaak erg waardevol ook na het afronden van het onderzoek waarbinnen ze verzameld of gegenereerd werden. Lacunes in zowel interne als externe infrastructuur dienen dus dringend te worden opgevangen en gevuld door bijkomende, al dan niet gedeelde (institutionele) infrastructuur, en daarnaast moeten reeds bestaande voorzieningen in Vlaanderen natuurlijk ook up to date en performant gehouden worden.

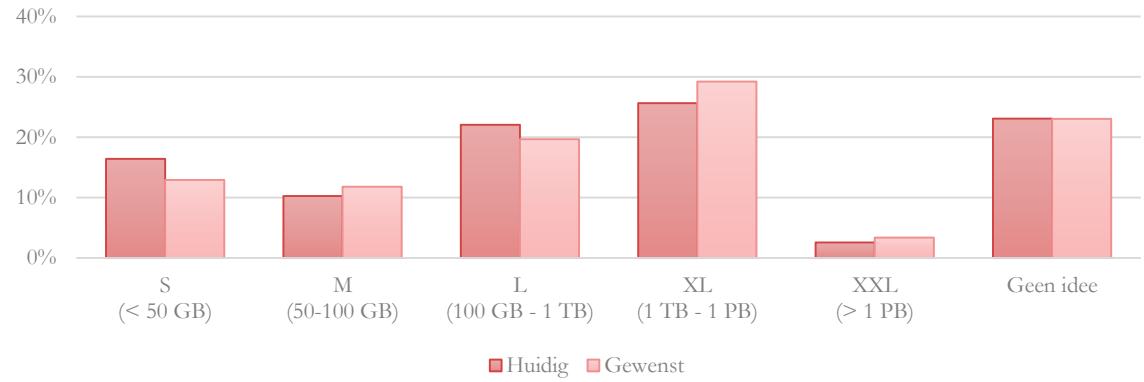


Dat deze lacunes er wel degelijk zijn en ook door onderzoekers aangevoeld worden, bewijzen de resultaten van de bevraging. Onderzoekers zowel als onderzoekseenheden gaven in eerste instantie de huidige opslaglocaties van hun onderzoeksdata weer. Bij onderzoeksleiders werd daarna ook gepeild wat hun geprefereerde opslaglocaties zouden zijn. Figuur 2 toont aan dat opslag in een kwart van de gevallen nog gebeurt op individuele opslagmedia, maar dat men deze graag inruilt voor opslagfaciliteiten die meer centraal, hetzij op niveau van de universiteit, hetzij op niveau van de eenheid worden aangeboden. Ook op het vlak van cloudopslag geeft men de voorkeur aan instellingsgebonden oplossingen in plaats van zuiver externe providers (Google Drive, Dropbox…). Een cloud storage aanbod is meestal wel al aan de instelling aanwezig, maar minder gekend dan de externe commerciële diensten die onderzoekers op individuele basis gebruiken. Om het verlies van data te vermijden moeten de onderzoeksdata op een correcte manier voorzien worden van backups. Bijna alle respondenten beweren backups te maken van

de onderzoeksdata, maar, net als bij de algemene opslag van data, gebeurt dit vaak niet via de centrale infrastructuur maar eerder aan de hand van kopieën op persoonlijke apparaten (laptop, externe harde schijf). Ook hier is het handelen ingegeven door een gebrek aan kennis van bestaande infrastructuur, waarvoor extra opleiding noodzakelijk is, in combinatie met het niet of onvoldoende ter beschikking zijn van infrastructuur – hetgeen bijkomende investeringen vergt.

Naast opslaglocatie vroegen we de onderzoekseenheden ook om een inschatting te maken van het jaarlijks volume aan data dat ze opslaan. Er gaat veel onzekerheid gepaard met het maken van deze inschatting, en zo'n 25% van de onderzoekseenheden gaf aan hier geen duidelijk zicht op te hebben, temeer omdat meestal ook op verschillende locaties wordt opgeslagen. Bij verdere bevraging van de onderzoekers bleek de beschikbare capaciteit toch in vele gevallen te beperkend. Zo ondervond een op drie onderzoekers ooit problemen bij de opslag van onderzoeksdata omwille van de bestandsgrootte. Figuur 3 toont aan dat de datanoden die men ziet voor de toekomst nog meer dan nu op grote volumes gericht zijn. Het valt op dat de ingeschatte datavolumes zich nu reeds situeren in het segment 'large' tot 'very large'.

Figuur 3. Jaarlijks opslagvolume, huidig vs. gewenst



Gezien het belang van het vrijwaren van de mogelijkheid tot verificatie en validatie van onderzoek, en tot hergebruik van bestaande onderzoeksgegevens, werden de onderzoekers ook bevraagd met betrekking tot de langdurige bewaring van onderzoeksdata. Sommige universiteiten hebben in hun RDM-beleid reeds opgenomen dat relevante onderzoeksdata voor minimum vijf jaar na afloop van het onderzoek bewaard moeten worden, maar het volgen van dit beleid hangt grotendeels af van de individuele onderzoeker. Een duidelijk kader voor datapreservatie op het niveau van de vak- of onderzoeksgroep of de faculteit ontbreekt doorgaans, en men blijkt dit eerder als de verantwoordelijkheid te zien van de onderzoeker zelf. Onderzoekers zijn echter doorgaans niet zo bewust bezig met wat er na afloop van het onderzoek met de door hen geproduceerde data moet gebeuren (data blijven in het beste geval gewoon staan waar ze stonden aan het einde van een project), en ze hebben ook zelden zelf de tijd en mogelijkheden om data echt duurzaam te bewaren en ter beschikking te stellen. Onder andere omwille van het probleem van 'obsolescence' (van fysieke dragers, hardware en software) en de nood aan een degelijke ontsluiting, vraagt het voor de langere termijn toegankelijk en bruikbaar houden van digitale data immers geschikte processen en faciliteiten.

Bij het archiveren van onderzoeksdata moet hoe dan ook nagedacht worden over wie verantwoordelijk zal zijn voor het behoud van en de toegang tot de data. Men moet de keuze maken waar deze data gedeponeerd en opgeslagen zullen worden, alsook voor wie ze uiteindelijk toegankelijk zullen zijn. Betrouwbare data repositories bieden de onderzoeker de mogelijkheid om data(sets) duurzaam te bewaren en ook online (al dan niet onder restricties) beschikbaar te maken. Uit de enquête blijkt dat een grote groep van de onderzoekseenheden kijkt naar de eigen instelling om ook op lange termijn data te bewaren en beschikbaar te maken. In eerste instantie voor collega's uit de onderzoekseenheid, maar bij uitbreiding, en in mindere mate, ook voor andere onderzoekers of geïnteresseerden zowel binnen als buiten de instelling. Toch blijft het een vrij moeilijke zaak om exact te bepalen op welk niveau diverse datafaciliteiten best worden georganiseerd, en welke eigen rol instellingen precies te spelen hebben in het ecosysteem van bestaande of op til zijnde data repositories. Bepaalde disciplines of onderzoekcentra hebben immers reeds de gewoonte dit op regionaal, nationaal of internationaal niveau te organiseren, en ook daar moet bij het uitwerken van een infrastructuurplan dus terdege rekening mee gehouden worden.

Instellingen zullen in elk geval met een stijgende nood aan centrale opslaginfrastructuur en faciliteiten voor archivering en beschikbaarstelling van onderzoeksdata geconfronteerd worden, en met aanzienlijke kosten voor het aanbieden van een adequate infrastructuur. Wil de overheid de instellingen hierin steunen of een operationele rol opnemen, dan is het essentieel dat er van een accuraat kostenmodel wordt uitgegaan. Daartoe bestaan inmiddels voldragen kostenmodellen voor 'digital curation', met als uitgangspunt dat dit een hele reeks van activiteiten omvat (inclusief, maar lang niet beperkt tot, 'archival storage') die gepaard gaan met verschillende soorten kosten (bv. kapitaallasten, kosten van arbeid) beïnvloed door uiteenlopende factoren[9]. De diversiteit van het takenpakket en de onderliggende kosten (werking, infrastructuur, personeel) maken een eenvoudige prognose van toekomstige uitgaven niet eenvoudig. Voor onderzoeksdata zijn bijvoorbeeld het Keeping Research Data Safe (KRDS)-model, dat resulteerde uit enkele studies gefinancierd door JISC in het Verenigd Koninkrijk, en het door het Nederlandse DANS ontwikkelde Cost Model for Digital Archiving (CMDA) relevant[10]. Deze studie, waarvan een overzichtstabel in de appendix is overgenomen, geeft de complexiteit van het proces weer en de veelheid aan aspecten die, naast de loutere storage, in rekening moeten worden gebracht.

Er zijn bijzonder weinig goed gedocumenteerde studies rond de totale kost van een holistische databeheer. Toch kunnen enkele cijfers een indicatie geven over bepaalde deelaspecten. Een schatting uitgevoerd door Arkivum Ltd met betrekking tot de gebruikte datavolumes aan Britse universiteiten duidt op een nood van ongeveer 4TB per onderzoeker per jaar[11]. Ook al lijkt dit op eerste zicht veel, er moet terdege rekening gehouden worden met het feit dat bepaalde disciplines vrij grote storagenoden in de toekomst voorzien. Wordt dit getal vermenigvuldigd met een gemiddelde kostprijs van 400€ per jaar, een prijs die nu reeds door heel wat instellingen wordt doorgerekend aan de onderzoeker, dan loopt de kostprijs voor een instelling snel op tot

[9] U.B. Keyser et al., "D3.1 – Evaluation of Cost Models and Needs & Gap Analysis" (4C Project – Collaboration to Clarify the Costs of Curation, June 2014). De activiteiten in deze kostenmodellen zijn doorgaans gebaseerd op het OAIS (Open Archival Information System) Reference Model.
[10] N. Beagrie et al., "Keeping Research Data Safe: A Cost Model and Guidance for UK Universities" (JISC, April 2008); N. Beagrie et al., "Keeping Research Data Safe 2" (JISC, April 2010); A.S. Palaiolog et al., "An Activity-Based Costing Model for Long-Term Preservation and Dissemination of Digital Research Data: The Case of DANS", International Journal on Digital Libraries 12 (2012) 4: 195-214. doi: 10.1007/s00799-012-0092-1.
[11] M. Addis, "Estimation of research data volumes per researcher in UK HEI" (2015).

miljoenen euro's, en dit enkel voor het opslaan van gegevens. Een aantal cases die meer aspecten in beeld brengen dan louter opslag zijn gedocumenteerd in de LERU Roadmap for Research data[12]. Voor zowel de universiteiten van Oxford als de UCL wordt de kost voor de opstart van data managent services geschat op ruim 1 miljoen pond, met een recurrente jaarlijkse uitgave van zeker de helft van dit bedrag. Omdat dergelijke kosten weinig schaalafhankelijk zijn, zou dit voor dit Vlaanderen betekenen dat een directe doorstart een bedrag van 5 miljoen euro zou vergen.

Data delen & open data: de kloof tussen verwachting en realiteit

Internationaal weerklinkt een steeds luidere roep om, naast open access tot peer-reviewed publicaties, ook de achterliggende onderzoeksdata toegankelijk te maken in het kader van een meer open wetenschap. Digitale onderzoeksdata ter beschikking stellen laat immers niet enkel toe om wetenschappelijke claims te verifiëren en valideren, maar maakt ook hergebruik mogelijk voor wetenschappelijk onderwijs en onderzoek, en door de samenleving in haar geheel. De Berlin Declaration on Open Access uit 2003 vormde in dit streven een belangrijke mijlpaal, omdat het open access-begrip zich hier niet beperkte tot peer-reviewed publicaties, maar ook andere onderzoeksmaterialen omvatte, waaronder ruwe data.[13] Sindsdien werd deze boodschap verder uitgedragen op het politieke niveau door organisaties als de OESO, die in haar (ook door België ondertekende) Declaration (2004) en haar Principles and Guidelines for Access to Research Data (2007) een pleidooi hield voor open toegang tot digitale onderzoeksdata uit publiek gefinancierd onderzoek in de lidstaten.[14] Ook de Europese Commissie lanceerde in 2012 een gelijkaardige aanbeveling.[15] Datzelfde jaar verklaarden de drie ministers bevoegd voor wetenschap en onderzoek in België zich in de Brussels Declaration on Open Access to Belgian publicly funded research akkoord om de vrije toegang tot publiek gefinancierde onderzoeksresultaten te maximaliseren, onder meer door het "onderzoeken van mogelijkheden en nieuwe opportuniteiten in het brede Open Access domein (…), en daarbij Open Access tot wetenschappelijke publicaties te beschouwen als een voorbode van nieuwe initiatieven op het gebied van 'Open Data' en 'Open Science'."[16]

Inmiddels is de Open Science-beweging aan een sterke opmars bezig en groeit ook het besef dat onderzoeksdata geen louter bijproduct van onderzoek zijn, maar een waardevolle vorm van wetenschappelijke output op zich. Het hoeft dan ook niet te verbazen dat het publiek ter beschikking stellen van onderzoeksdata een steeds belangrijker aandachtspunt wordt bij internationale onderzoeksfinanciers en wetenschappelijke uitgevers. Deze steeds breder gedragen steun voor open research data als standaard in de wetenschappelijke praktijk is ingegeven door verschillende agenda's: niet enkel de bezorgdheid om wetenschappelijke integriteit, transparantie en reproduceerbaarheid speelt hier een rol, maar ook het idee dat data uit publiek gefinancierd onderzoek een publiek goed zijn, evenals de opportuniteiten voor nieuwe (vormen van) kenniscreatie en data-intensief onderzoek, voor innovatie en data-gedreven beleidsvorming, en voor een groter rendement van publieke investeringen in onderzoek dankzij datahergebruik.

---

[12] http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf
[13] *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (22 October 2003).
[14] *Declaration on Access to Research Data from Public Funding* (30 January 2004); OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007).
[15] European Commission, *Commission Recommendation on Access to and Preservation of Scientific Information* (17 July 2012).
[16] *Brussels Declaration on Open Access to Belgian Publicly Funded Research* (22 October 2012).

Tegelijk wordt door de belangrijke spelers ook wel erkend dat het beschikbaar maken van onderzoeksdata uit afgerond onderzoek een complexer verhaal is dan het realiseren van open access tot wetenschappelijke publicaties. Zo zijn er legitieme juridische, ethische en veiligheidstechnische redenen waarom de toegang tot (bepaalde) onderzoeksdata soms (al dan niet tijdelijk) beperkt moet worden: bv. de bescherming van privacy, van kwetsbare soorten of groepen, van valoriseerbare onderzoeksresultaten en intellectuele eigendom, of van de nationale veiligheid. Op Europees vlak luidt het devies dan ook steeds meer dat toegang tot onderzoeksdata "as open as possible, as closed as necessary" moet zijn (zie o.a. de RDM richtlijnen van de Europese Commissie in het kader van het Horizon 2020 programma, alsook de bepalingen in de recent herziene editie van de ALLEA Code of Conduct for Research Integrity). Dit betekent evenzeer dat men niet slechts de keuze heeft tussen hetzij volledig open, hetzij volledig gesloten data, maar dat er tussen deze uitersten nog mogelijkheden bestaan voor het delen van onderzoeksdata onder beperktere toegangs- en/of gebruikscondities. Bovenal is het van belang zich te realiseren dat het louter ontsluiten van data op zichzelf weinig waarde heeft, en dat het (al dan niet onder restricties) delen van onderzoeksdata vooral op een intelligente manier dient te gebeuren: d.w.z. in lijn met de FAIR principes (zie hoger), die bepaalde eisen stellen op het vlak van o.a. de aanwezigheid van online discovery metadata, persistent identifiers en documentatie (beschrijvende en contextuele informatie om de datasets te begrijpen), het gebruik van standaarden (bv. qua bestandsformaten, metadataschema's, ontologieën en licenties), enz.

Ondanks de toenemende aandacht voor open en FAIR onderzoeksdata bij beleidsmakers, financiers, uitgevers en andere wetenschappelijke stakeholders, blijkt de realiteit er vandaag echter nog heel anders uit te zien. Studies uit het buitenland geven immers aan dat onderzoekers lang niet unaniem zijn in het omarmen van het delen van data en open data praktijken. In sommige domeinen bestaat hierin reeds een sterke traditie (met name daar waar delen noodzakelijk is voor het onderzoek); daarbuiten erkennen veel onderzoekers weliswaar de voordelen van het delen van data, maar zijn ze weinig geneigd om dit zelf in praktijk te brengen. Bovendien, waar het wel gebeurt, verloopt het niet altijd op een manier die resulteert in FAIR data.[17] Ook de bevindingen aan de Vlaamse universiteiten liggen grotendeels in dezelfde lijn. Wanneer gekeken wordt aan wie men toegang tot onderzoeksdata verschaft, blijkt dat onderzoekers momenteel niet zo vaak delen met de wetenschappelijke gemeenschap of het bredere publiek (slechts 12% van de gegeven antwoorden in de survey). Het zeer selectief delen van data, d.w.z. met peer reviewers, met geïnteresseerde personen op aanvraag, of met bepaalde leden van de eigen instelling, komt daarentegen het vaakst voor (70%). Als er al gedeeld wordt, gebeurt het dan ook zelden via een data repository (9%), en des te meer op manieren die weinig garanties op FAIR data bieden: via e-mail, fysieke dragers, cloud applicaties en gedeelde netwerkschijven (73%).

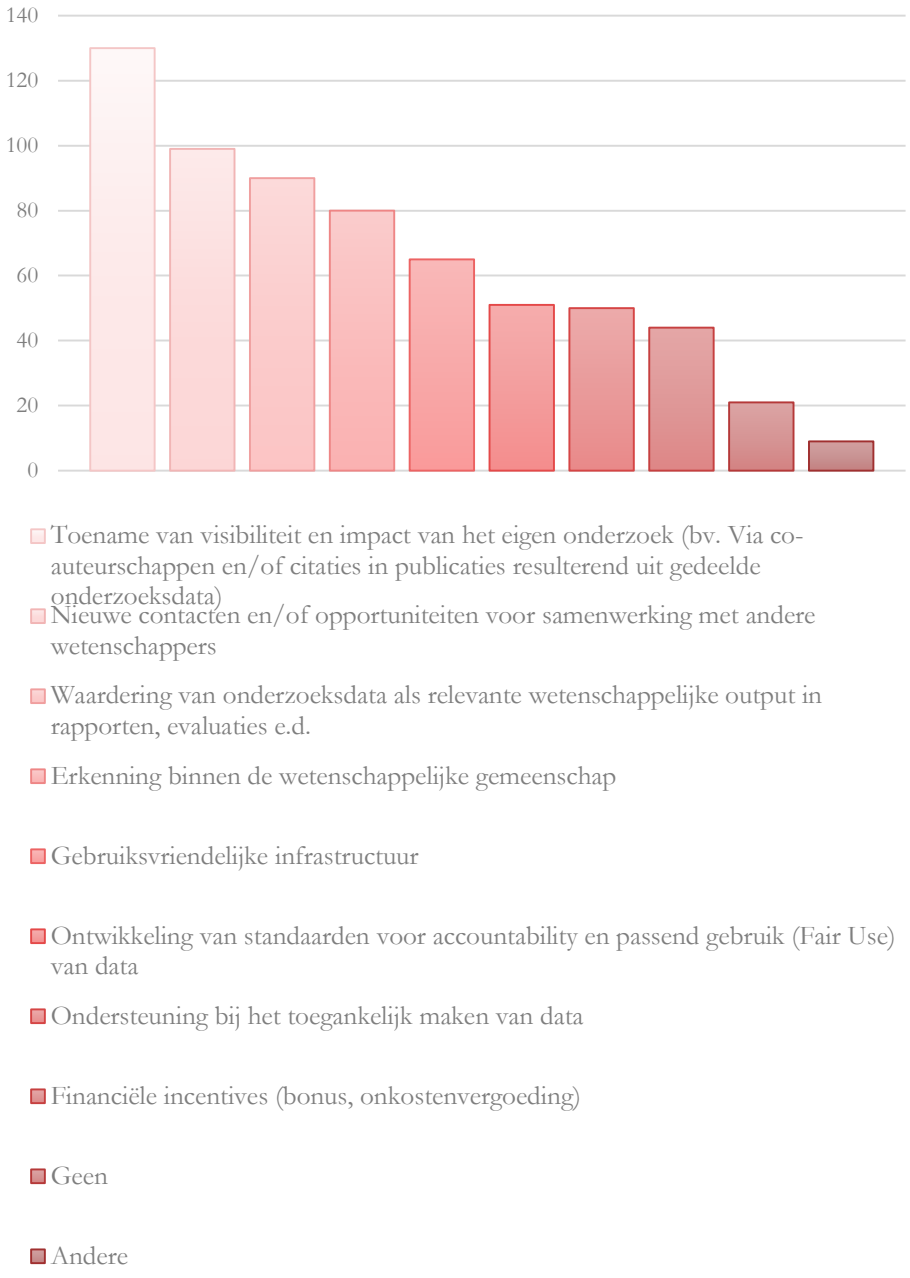
Er blijken heel wat redenen te zijn waarom onderzoekers vandaag niet overgaan tot het delen van onderzoeksdata, en deze kunnen in belangrijke mate gegroepeerd worden als juridische, praktische en sociale barrières. De juridische hinderpalen (privacyschending, IPR) blijken niet te onderschatten (24% van de antwoorden), en praktische obstakels (vereiste inspanning qua tijd/kost, gebrek aan infrastructuur, gebrek aan data standaarden en processen, gebruik van obscure formaten) worden nog vaker aangehaald (28%). Sociale barrières zoals het risico op verkeerde interpretatie/falsificatie van data, potentieel ongewenst commercieel gebruik, risico op misbruik, toegenomen wetenschappelijke competitie en gebrek aan motivatie zijn echter veruit het belangrijkst (46 %). Toch kunnen bepaalde incentives voor onderzoekers een stimulans zijn om data (openlijk) te delen. Figuur 4 toont een oplijsting van de belangrijkste incentives. Wetenschappelijke rewards

[17] Zie bijvoorbeeld: S. Berghmans, H. Cousijn, G. Deakin et al., *Open Data: The Researcher Perspective* (12 April 2017).

(erkenning in de onderzoeks-gemeenschap, waardering van onderzoeksdata als relevante wetenschappelijke output, grotere zichtbaarheid en impact van onderzoek, nieuwe contacten en opportuniteiten voor samenwerking) blijken daarbij veruit het belangrijkst te zijn (62% van de antwoorden). Daarna volgen praktische ondersteuning (gebruiksvriendelijke infrastructuur en menselijke support) (18%), de ontwikkeling van standaarden voor fair use van data (8%) en financiële incentives (7%). Voor slechts een zeer kleine minderheid is er geen enkele incentive die zou motiveren tot het delen van data of open research data (3%).

Globaal staan onderzoekers niet a priori negatief ten aanzien van het delen of openstellen van data, zij het onder voorbehoud van juridische zekerheid en infrastructurele omkadering. Er lijkt zeker nog ruimte om onderzoekers in Vlaanderen meer te stimuleren hun data zo openlijk mogelijk beschikbaar te maken. Daarbij hebben het aanbieden van de juiste incentives (vnl. wetenschappelijke erkenning en ondersteuning) alsook het wegwerken, waar mogelijk, van de bestaande (vnl. sociale en praktische) obstakels een cruciale rol te spelen.

Figuur 4. Welke incentives zouden je kunnen motiveren om onderzoeksdata te delen en ze (openlijk) beschikbaar te maken?



Opleiding op maat

Het voorzien van infrastructuur zal niet volstaan om kwalitatief en duurzaam databeheer te bekomen, hiervoor dienen onderzoekers, datamanagers en beleidsmedewerkers ook de nodige training te krijgen. De enquête afgenomen bij de Vlaamse instellingen toonde dit aan. Zo had het merendeel van de bevraagden noties van een aantal verschillende aspecten van onderzoeksdatamanagement, maar ontbreekt de kennis van het totale plaatje. Meer in detail bevraagd naar hun kennis over specifieke aspecten, bleken onderzoekers doorgaans deels tot volledig onbekend met de juiste terminologie en de bijhorende vereisten. De nood aan een gedegen opleidingskader waarbij zowel beginnende als meer ervaren onderzoekers kennis kunnen opdoen over onderzoeksdatamanagement is groot, wat de onderzoekers duidelijk onderschrijven. Het is dan ook geen toeval dat vele landen reeds specifieke RDM en andere opleidingsinitiatieven organiseren voor zowel beginnende als meer ervaren onderzoekers. Voorbeelden dicht bij huis zijn Nederland, waar (DANS-KNAW samen met Research Data Netherlands een aantal specifieke cursussen aanbiedt[18], en het Verenigd Konikrijk, waar het Digital Curation Centre (DCC) een ruim trainingsaanbod verzorgt.[19] Daarnaast organiseren steeds meer instellingen online cursussen.[20]

Dit aanbod is, ondanks de voor een leek mogelijk overweldigende complexiteit en technische veelzijdigheid, typisch eerder generalistisch: men schetst de grote lijnen en essentiële aspecten van onderzoeksdatamanagement zodat inzicht ontstaat in de samenhang van het geheel. Dit is belangrijk, omdat goed RDM vertrekt vanuit een holistische benadering. Met name voor onderzoekers is dit type opleiding cruciaal, zowel voor beginnende onderzoekers (bij voorkeur te integreren als verplicht onderdeel in de doctoraatsopleiding) als voor meer ervaren onderzoekers om bij te blijven met evoluerende technische mogelijkheden en wettelijke verplichtingen.

De theorie rond RDM omzetten in een praktijk van vindbare, toegankelijke, deelbare en herbruikbare data, vraagt naast goed geïnformeerde en betrokken onderzoekers de inzet van tal van experten met een diepgaande kennis van de verschillende deelaspecten. Denk hierbij aan samenwerking met juridische en ethische specialisten, informatieveiligheidsconsulenten (data protection officers), valorisatie-experten, bibliothecarissen en (data) archivarissen (data repository managers). Het FAIR en open bewaren van een dataset vraagt immers het toetsen van contractuele, wettelijke en ethische vereisten, onder meer met betrekking tot de mogelijkheden en beperkingen tot vrije beschikbaarheid voor derden ('open data'), het cureren en valideren van datasets, en gericht advies over de gepaste methodiek om de data te beheren (met de juiste metadata en in het juiste formaat) en te beveiligen.

Er is nood aan drie niveaus van kennis en vaardigheden: in de eerste plaats dienen onderzoekers minimaal het 'ABC' van het onderzoeksdatamanagement te begrijpen om een goed inzicht te verwerven in de onderliggende principes en in wat zij zelf (nog) niet weten, zodat zij zich gericht kunnen laten bijstaan door, ten tweede, een netwerk van diepte-experts over uiteenlopende aspecten, typisch te situeren binnen de centrale administratie van de universiteiten. Het derde niveau brengt onderzoeker en specialist samen vanuit de combinatie van een meer generalistische kennis van datamanagement en een goed begrip van het onderzoek: de 'research data manager'. Typisch bevinden deze kernspelers zich op het centrale niveau, als eerste aanspreekpunt voor

[18] https://dans.knaw.nl/nl/over/diensten/training-consultancy.
[19] http://www.dcc.ac.uk/training.
[20] Zie bv. http://datalib.edina.ac.uk/mantra; http://cradle.web.unc.edu; of http://datasupport.researchdata.nl.

onderzoekers en 'makelaar' in RDM dienstverlening. De voorlopers in onderzoeksdatamanagement investeren in datamanagers op het niveau van een faculteit of zelfs binnen een onderzoeksgroep – hoe dichter bij de onderzoekspraktijk, hoe meer discipline-specifieke vaardigheden op gebied van data curatie en archivering de onderzoeksdatamanagers zich eigen kan maken. Het is aan de universiteiten zelf om de juiste mix tussen centraal aangehechte of discipline-specifieke onderzoeksdatamanagers en diepte-specialisten te bepalen.

Men kan zich de vraag stellen op welke manieren de Vlaamse universiteiten deze verschillende vormen van expertise kunnen opleiden dan wel inhuren. Het antwoord is enigszins ontnuchterend, maar daarom niet verontrustend. De brede, laagdrempelige opleiding van onderzoekers wordt binnen de instellingen reeds op korte termijn voorzien in de vorm van eigen info- en trainingssessies. De training van de diepte-experts op het centrale niveau gebeurt eerder natuurlijk, waarbij zij (internationale) bijscholingen volgen over de eigen expertise, en zo hun kennis actueel en relevant houden. Voor de 'research data managers' bestaan in Vlaanderen (of België) echter geen specifieke opleidingen. Nochtans spreken schattingen over enorme toekomstige noden aan dit profiel, die zeker gevoed worden door de verplichtingen opgelegd door de AVG, maar ook los daarvan voortvloeien uit de steeds groeiende 'big data' afhankelijkheid van vele (bedrijfs)processen. In Vlaanderen worden, ondanks de stijgende vraag, momenteel geen volwaardige opleidingen met deze specifieke invalshoek georganiseerd, in tegenstelling tot in onze buurlanden. Indien we deze trein aan Vlaanderen laten voorbijgaan, lijkt het onwaarschijnlijk dat we voldoende kwaliteitsvolle kandidaten zullen kunnen blijven aantrekken.

Dringende nood aan het in kaart brengen van data profielen en de daarbij horende opleiding

Wat als Vlaanderen geen actie onderneemt?

Er zijn verschillende risico's identificeerbaar indien Vlaanderen geen stappen onderneemt voor een correcter en duurzamer beheer van onderzoeksgegevens.

1. **Economisch**: Geraamd wordt dat binnen de EU 28+ in 2020 Open Data, slechts één van de aspecten van een goed en duurzaam databeheer[21], een marktomvang van 75.7 miljard euro zal hebben[22]. Deze evolutie gaat gepaard met een stijgende nood aan personeel met expertise rond databeheer en –analyse. Tegen 2020 zouden er maar liefst 100.000 jobs in Open Data ingevuld dienen te worden[23]. Wil Vlaanderen hierin een rol opnemen en de ambities van de Beleidsnota *Werk, Economie, Wetenschap en Innovatie 2014-2019*[24] waarmaken door te investeren in kennisopbouw en innovatie (punt 1.3.1) en in een excellente kennisbasis (punt 2), dan is het aanbieden van opleiding en state-of-the-art onderzoeksinfrastructuur (punt 2.4) een absolute voorwaarde om tot een goed beleid voor onderzoeksgegevens in het algemeen en Open Data (punt 2.5) in het bijzonder te komen. **Vlaanderen zal economische opportuniteiten missen indien er niet tot actie wordt overgegaan.** Tegelijkertijd is het belangrijk om tijdig de financiële implicaties van het uitrollen van een goed databeheer in te schatten en te budgetteren. Waar de kosten voor opslag steeds lager worden, exploderen tegelijk ook de datavolumes en moet bovendien rekening gehouden worden met ontsluiting en begeleiding bij het klaarmaken van de data voor (lange termijn) opslag of ontsluiting.

2. **Maatschappelijk**: Kwaliteitsvolle en duurzame voorzieningen rond databeheer en data delen hebben een aantoonbare efficiëntiewinst die binnen de EU28+ geraamd wordt op een geaccumuleerd bedrag van maar liefst 1.7 miljard euro tegen 2020[25]. Kosten worden bespaard doordat het delen van (meta)data overbodige analyses vermijdt, maar vooral omdat diverse actoren sneller aan kwaliteitsvolle gegevens raken ter ondersteuning van of nodig voor de ontwikkeling van nieuwe producten en diensten. Een zeer fundamentele afgeleide hiervan is dat hierdoor tijdsefficiënter kan worden gewerkt, wat in zake grote maatschappelijke uitdagingen[26] pure winst betekent. Bovendien leidt inzicht geven in de beschikbaarheid van onderzoeksdata tot meer transparantie en een winst in vertrouwen bij het bredere publiek m.b.t. de relevantie van onderzoek. **Geen actie ondernemen betekent de digitale mogelijkheden niet kostefficiënt en niet tijdsefficiënt aanwenden daar waar tijdige ingrepen vaak cruciaal zijn**[27]. Bovendien geeft een passieve houding commerciële bedrijven de kans deze markt wél te ontsluiten- waar zij nu reeds volop op inzetten. De vraag moet hierbij gesteld worden **of Vlaanderen wil dat toegang tot onderzoeksdata, vaak gefinancierd door publieke kanalen, gemonopoliseerd wordt door commerciële aanbieders die doorgaans geen garantie op duurzaamheid en openheid bieden.**

[21] Onderzoeksgegevens vallen zeer vaak onder meerdere beperkingen van wettelijke of contractuele aard. Een eenduidig pleidooi voor Open Data is daarom niet realistisch.
[22] https://www.europeandataportal.eu/en/using-data/benefits-of-open-data
https://ec.europa.eu/epsc/sites/epsc/files/strategic_note_issue_21.pdf
[23] https://www.europeandataportal.eu/en/using-data/benefits-of-open-data
[24] https://www.vlaanderen.be/nl/publicaties/detail/beleidsnota-2014-2019-werk-economie-wetenschap-en-innovatie
[25] https://www.europeandataportal.eu/en/using-data/benefits-of-open-data
[26] Zoals gedefinieerd op https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges.
[27] Talrijke voorbeelden zijn raadpleegbaar op http://odimpact.org/

3. **Imago-gewijs:** Gezien het aantoonbare verband tussen het delen van onderzoeksgegevens en een stijging in het aantal citaties[28], de internationaal toenemende erkenning voor onderzoeksdata als waardevolle, citeerbare onderzoeks-output en de steeds prominentere vraag van onderzoeksfinanciers en tijdschriftuitgevers naar goed databeheer en het delen van onderzoeksdata **lopen we het risico op minder erkenning voor/impact van Vlaams(e) onderzoek(ers) en gemiste opportuniteiten voor onderzoeksfinanciering en publicaties.** Bovendien heeft Vlaanderen als onderzoeksintensieve regio met de ambitie om te excelleren in onderzoek en innovatie een reputatie hoog te houden. Op dit moment kunnen we enkel met enige na-ijver vaststellen hoe goed de ons omringende landen op het gebied van databeheer georganiseerd zijn, m.n. Nederland (met bv. DANS, 4TU.Centre for Research Data, Research Data Netherlands en het Landelijk Coördinatiepunt voor Research Data Management) en het Verenigd Koninkrijk (o.a. DCC, UK Data Service, Jisc Research data shared service). **Blijven we aan de zijlijn toekijken hoe goed onze buren infrastructuur en ondersteuning m.b.t. databeheer georganiseerd hebben en zo ware pioniers in het veld zijn geworden? We missen daarbij opportuniteiten om de hoge kwaliteit van ons onderzoek te demonstreren** en zo het vertrouwen in wetenschappelijk onderzoek in Vlaanderen te vergroten.

[28] Meest recent (2017): SPARC Europe, The Open Data Citation Advantage - http://sparceurope.org/open-data-citation-advantage/

Actiepunten

Onze onderzoeksgegevens zijn ons intellectueel kapitaal. Ze vormen de basis van nieuwe inzichten om maatschappelijke uitdagingen aan te gaan, zijn de motor van economische ontwikkelingen én de basis voor de vorming van toekomstige onderzoekers. De Vlaamse universiteiten zijn zich zeer sterk bewust van dit potentieel. De bottom-up samenwerking tussen de universiteiten die gestalte heeft gekregen in de formele overlegstructuren van de Vlaamse Interuniversitaire Raad (VLIR) als werkgroep RDM en Open Science heeft reeds geleid tot diverse geconcerteerde acties. Deze werkgroep is klaar om op basis van empirische gegevens en een toepassing die op dit moment klaar is voor gebruik (DMPonline.be), de onderzoekers verbonden aan de Vlaamse universiteiten praktisch te begeleiden bij het opstellen van een gedegen beheersplan voor onderzoeksdata. Daarbij worden reeds aan de meeste instellingen op regelmatige basis infosessies en workshops gegeven rond het belang van correct databeheer en de toepassingen en diensten waarop men een beroep kan doen. Voorlopig gebeurt dit binnen de middelen die op dit moment beschikbaar zijn.

Het is evenwel volkomen onrealistisch ervan uit te gaan dat men mits voldoende enthousiasme bij en samenwerking tussen de universiteiten alle uitdagingen in deze paper geschetst aankan. Uit onze bevraging blijkt dat er consistent drie grote noden worden geïdentificeerd waaraan niet op het niveau van één of zelfs een consortium van meerdere universiteiten kan beantwoord worden wil men efficiënt te werk gaan. Daarom roepen we gezamenlijk op tot een aantal urgente acties. Wij beschouwen deze als expliciete hefbomen voor de bestaande initiatieven waarvoor het grondwerk door de werkgroep is gelegd. Wil men in Vlaanderen het data kapitaal in onze handen écht verzilveren, als regio een rol spelen in de kenniseconomie, werk maken van open data en zich niet beperken tot het inschrijven van principes in decreten, dan zijn dringend bijkomende acties nodig, minimaal op regionaal niveau.

ACTIE 1 - INVESTEER IN INFRASTRUCTUUR

Het beschikbaar stellen van duurzame infrastructuur met een groot gebruiksgemak als hefboom voor gedegen opslag en bewaring

Uit de bevraging bij onderzoekers aan Vlaamse universiteiten bleek duidelijk een nood aan duurzame en gebruiksgemakkelijke infrastructuur om data langdurig te bewaren of te delen. Dit werd gehinderd door de afwezigheid van betaalbare oplossingen voor grote volumes of specifieke formaten. De geschatte kostprijs van langetermijnbewaring van miljoenen euros per jaar is bijzonder ontnuchterend: dit zijn bedragen die de Vlaamse instellingen onmogelijk zelf kunnen dragen.

Tegelijk worden instellingen, om op korte termijn tegemoet te komen aan de Europese regelgeving, gedwongen hun toevlucht te nemen tot diverse commerciële oplossingen om hun data per direct open te kunnen stellen, zo niet maken onderzoekers gebruik van een veelheid aan bestaande externe gratis platformen om hun gegevens

op te slaan. Daarbij worden soms de rechten op de data aan de aanbieder overgedragen. Met andere woorden – we schenken ons intellectueel kapitaal weg of gaan er toch op zijn minst slordig mee om.

Om te beletten dat kapitaal als dit onbekend (want nergens geregistreerd) en onvindbaar (want niet op een zichtbare plaats geregistreerd) op een niet-duurzame manier beheerd – en daardoor dus onbruikbaar voor de wetenschap – wordt, is een krachtdadig optreden van de regionale overheid vereist. Er moet geïnvesteerd worden in (gedeelde) infrastructuur waarop diverse data voor langere termijn bewaard kan worden, waarbij minimaal het bestaan van deze data geregistreerd wordt door kwalitatieve metadata, mét mogelijkheden om al dan niet onder restricties de data te delen én met een technologie die het mogelijk maakt aan te sluiten bij de European Open Science Cloud.[29] Deze gevraagde investering is financieel – een budget dat toelaat dit uit te bouwen moet vrijgemaakt worden – maar ook een investering in overleg. Er zal op een gestructureerde wijze moeten worden samengewerkt om technische en juridische kennis samen te brengen, en dit moet deels centraal deels institutioneel gestuurd worden in wederzijdse afstemming. Buitenlandse voorbeelden zoals DANS[30] in Nederland en DCC [31] in het Verenigd Koninkrijk zijn hiervoor goede leidraden. Wil Vlaanderen zich positioneren als een onderzoeksintensieve regio met een volwaardige rol in een actieve kenniseconomie, moet ook aan de buitenwereld getoond dat we kwaliteitsvolle data genereren en, al dan niet onder restricties, ter beschikking stellen. Zonder substantiële financiële injecties blijft een gestructureerd (open) databeleid dode letter en zullen ook de mogelijkheden van de big data science onderbenut blijven.

ACTIE 2 - INVESTEER IN OPLEIDING

Opleiding van datamanagers als hefboom voor kwaliteitsvolle gedeelde en open data: data die goed gestructureerd en gedocumenteerd zijn, in een herbruikbaar formaat opgeslagen, én bevraagbaar en opvraagbaar zijn

Infrastructuur alleen volstaat echter niet. Vooraleer datasets via een aangepaste infrastructuur aangeboden worden moet een minimum aan kwaliteit verzekerd zijn. Het heeft geen nut om data die technisch (niet compatibele formaten) of contextueel (geen documentatie of niet-gestandardiseerde metadata) niet herbruikbaar zijn op een platform aan te bieden. Gezien de reeds zware belasting van onderzoekers is het essentieel dat er geïnvesteerd wordt in het opleiden van zogenoemde datamanagers, die onderzoekers bijstaan van bij het begin van een onderzoek om efficiënt kwaliteitsvolle datasets te creëren die uiteindelijk voor delen in aanmerking komen. Deze datamanagers dienen een zeer grote affiniteit te hebben met onderzoek, een gedegen kennis te hebben van alle aspecten rond databeheer, inclusief de wettelijke bepalingen, en technisch de vertaling te maken tussen wensen van onderzoekers en de beschikbare infrastructuur. Het initiatief vanuit

[29] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
[30] https://dans.knaw.nl/nl/.
[31] http://www.dcc.ac.uk/.

de VVDAB tot het herinstellen van een opleiding 'informatiemanagement' verdient daarom alle steun vanuit de regionale overheid in acht nemend dat deze opleiding ook experts in onderzoeksdatamanagement kan aanleveren. Naast het investeren in opleiding, zal ook bijkomend budget beschikbaar moeten gemaakt worden om deze professionals ook daadwerkelijk aan onze instellingen aan te stellen.

ACTIE 3 - SCHEP EEN DUIDELIJK WETTELIJK KADER

Indien ervoor wordt geopteerd om de Europese Algemene Verordening Gegevensbescherming (GDPR) om te zetten naar nationale wetgeving, moet erover worden gewaakt dat er geen bijkomende restricties voor onderzoek worden gecreëerd.

Actieve steun aan de Europese hervorming van het auteursrecht en de bepalingen rond text and data mining als hefboom om ook als onderzoeker ten volle gebruik te kunnen maken van digitale technologieën

Uit de bevraging kwam duidelijk naar voren dat onderzoekers vaak aarzelen om data te delen of publiek ter beschikking te stellen omwille van het ontbreken van een duidelijk wettelijk kader. Hieronder vallen diverse aspecten die vaak op het niveau van de individuele instelling of contract geregeld of geëxpliciteerd dienen te worden, zoals het eigendomsrecht op data of de voorwaarden waaraan gebruikers van secondaire data moeten voldoen. Wat betreft het gebruik en delen van persoonsgegevens is er echter wél een taak voor de overheid weggelegd. Als de vernieuwde Europese Algemene Verordening Gegevensbescherming met zo min mogelijk verdere restricties in het gebruik van persoonsgegevens wordt overgenomen door Vlaanderen, zorgt dit ervoor dat alle onderzoekers en de hen ondersteunende diensten zich aan wettelijk kader dienen te houden dat in Europa zoveel mogelijk uniform en niet-restrictief is. Ook op Europees vlak dient vanuit Vlaanderen actief steun te worden gegeven aan de voorliggende hervorming van het auteursrecht en de bepalingen rond *text and data mining*, conform het gezamenlijke standpunt van een brede coalitie aan onderzoeksinstellingen[32] waardoor onderzoek ten volle gebruik kan maken van de bestaande digitale technologieën en kan concurreren met landen waarin regels veel soepeler zijn.

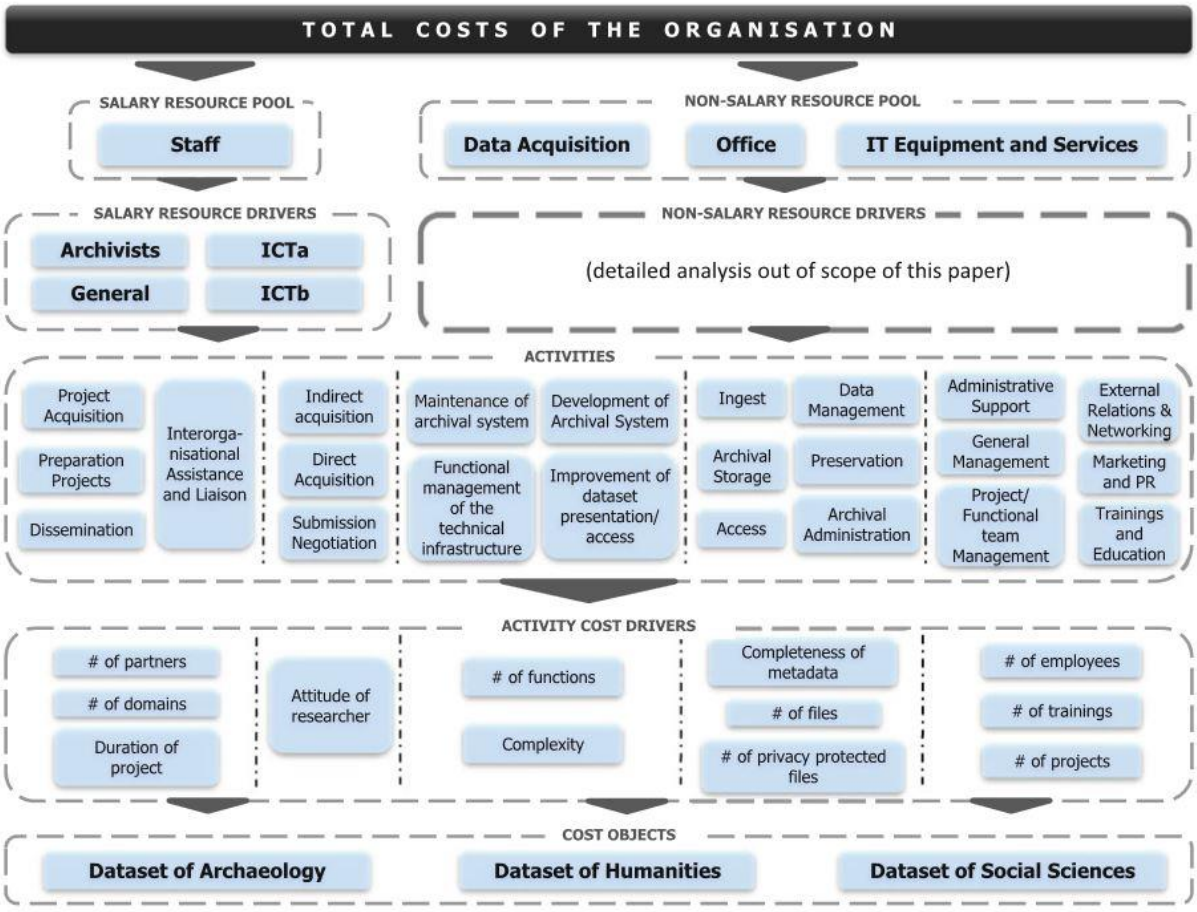
[32] http://www.eua.be/activities-services/news/newsitem/2017/01/12/eu-copyright-reform-eua-and-leading-research-groups-push-for-more-change.

ACTIE 4 - INCENTIVES VOOR OPEN SCIENCE

Erkennen van niet-traditionele publicaties als waardevolle onderzoeksoutput als hefboom voor het ter beschikking stellen van (de metadata van) onderzoeksgegevens

Deze aanbeveling is met opzet als laatste verwoord omdat deze geen zin heeft als de voorgaande hefbomen niet of onvoldoende in werking getreden zijn. Die onderzoekers die nu reeds aangeven gegevens te willen delen moeten hiervoor zo goed mogelijk technisch en inhoudelijk ondersteund worden. Voor de vele onderzoekers die aangeven dat het niet krijgen van wetenschappelijke erkenning een belangrijke belemmering is om data ter beschikking te stellen dient elke instelling na te denken over welke incentives aangeboden kunnen worden. Nu reeds zijn er instellingen die expliciet erkenning geven voor niet-traditionele publicatie output. De regionale financiers blijven evenwel in gebreke. Daar blijven publicaties dé manier waarop onderzoek gevaloriseerd wordt en dus impact kan creëren, waarbij men een zeer traditionele impliciete definitie van het begrip publicaties hanteert, zijnde tijdschriftartikels, boeken, boekhoofdstukken en congresbijdragen. Deze definitie is aan herziening toe: alle academische resultaten die publiek worden gemaakt, bij voorkeur na peer review, dienen als wetenschappelijke publicaties erkend te worden. Dat geldt dus ook voor datasets, databanken, software code e.d.m.. 'Publiek' kan daarbij betekenen dat de hele inhoud openlijk beschikbaar is, maar ook dat enkel de metadata publiek zijn waarbij nog steeds, onder voorwaarden eventueel, de data of code kan gedeeld worden. Men mag namelijk niet uit het oog verliezen dat om redenen van commerciële exploitatie of andere valorisatiemethoden, beide ook een opdracht van de universiteiten, delen vaak niet mogelijk is. In deze gevallen mag het niet-delen ook niet bestraft worden. Er dient een opening gecreëerd te worden waarbij het delen van onderzoeksresultaten op andere manieren dan in een klassieke publicatie ook erkenning oplevert bij onze financiers. Ook hier ligt een verantwoordelijkheid van de Vlaamse overheid.

Appendix 1: schema van het DANS-ABC model



Bron: A.S. Palaiolog et al., "An Activity-Based Costing Model for Long-Term Preservation and Dissemination of Digital Research Data: The Case of DANS", International Journal on Digital Libraries 12 (2012) 4: 195-214. doi: 10.1007/s00799-012-0092-1.

Dit document werd samengesteld door de leden VLIR-WG Research Data Management en Open Science.

KU Leuven: Hannelore Vanhaverbeke & Joke Claeys
Universiteit Antwerpen: Jord Hanus & Marianne De Voecht
Universiteit Gent: Myriam Mertens
Universiteit Hasselt: Hanne Elsen & Sadia Vancauwenbergh
Vrije Universiteit Brussel: Lucy Amez & Kyle Van Gaeveren

Naar deze paper kan gerefereerd worden als:

VLIR WG RDM & OS, Research Data Management en de Vlaamse universiteiten: White Paper, Oktober 2017, 26p.