Hoe kun je wetenschappers verleiden tot datamanagement? De Eramusuniversiteit Rotterdam ging met ze in gesprek. Dat bleek uiterst nuttig te zijn. 'Onderzoekers zitten niet te wachten op beleid dat zich richt op controle en dat dus op wantrouwen lijkt te zijn gebaseerd.'

Graag niet al te generiek

Gewetensvol datamanagement

Marlon Domingus

Erasmus Universiteit Rotterdam

Financiers en subsidieverstrekkers stellen in toenemende mate eisen aan de toegankelijkheid en het 'managen' van onderzoeksdata. Zogenaamde datamanagementplannen zijn inmiddels verplicht. Ook Europese en nationale gedragscodes en disciplinespecifieke gedragscodes stellen om redenen van betrouwbaarheid en controleerbaarheid eisen aan de wijze waarop onderzoekers data verzamelen, gebruiken, documenteren en archiveren.
En ten slotte stellen tijdschriften eisen aan de wijze waarop de aan het artikel onderliggende data beschikbaar moeten zijn voor onderzoek.

Vanuit het perspectief van de onderzoekers is bovenstaande benadering een verplichting van buitenaf, waaraan ze zullen moeten voldoen. Het is kenmerkend voor wat ik de 'Cambridge-aanpak' van researchdatamanagement (RDM) noem, met Marta Teperek als belangrijkste vertegenwoordigster.

> **Het werd duidelijk dat onderzoekers veel verschillende vragen tegelijkertijd krijgen**

Wij leerden Teperek in 2016 kennen en raakten onder de indruk van haar gedreven en effectieve aanpak in Cambridge. In haar recente artikel[1]: '*Is Democracy the Right System? Collaborative Approaches to Building an Engaged RDM Community*', beschrijft ze de legitieme redenen voor een *policy driven*, top-downbenadering van RDM. Tegelijkertijd geeft ze mee dat dit geen succesvolle strategie is. Bij onderzoekers leidt ze tot '*another checkbox activity*', in plaats van tot de beoogde inhoudelijke afwegingen en maatregelen rond de verantwoorde omgang met onderzoeksdata.

Integriteit

Toen we bij de Erasmus Universiteit Rotterdam (EUR) bezig waren met het opstellen van een zogenaamd *RDM-baseline-protocol,* namen we de Cambridge-aanpak zeer ter harte. De roep om zo'n protocol kwam voort uit het debat over wetenschappelijke integriteit. De richtlijn zou bindend moeten zijn voor alle EUR-onderzoekers en voor elk onderzoek – dus niet alleen voor extern gefinancierd onderzoek.

Allereerst maakten we een overzicht van wat verschillende partijen op dit moment al concreet aan de onderzoeker vragen. We ontdubbelden, categoriseerden en postioneerden de lijst naar de verschillende fasen van het generieke onderzoeksproces. Het werd duidelijk dat onderzoekers veel verschillende vragen tegelijkertijd krijgen. We maakten een prioritering van onderwerpen, met een baseline in gedachten. Dit werd onze basislijst (*mandatory*-elementen). Daarnaast was er een langere bijlage van onderwerpen die je zou kunnen bestempelen als '*mandatory if applicable*'.

Zwemmen of verzuipen

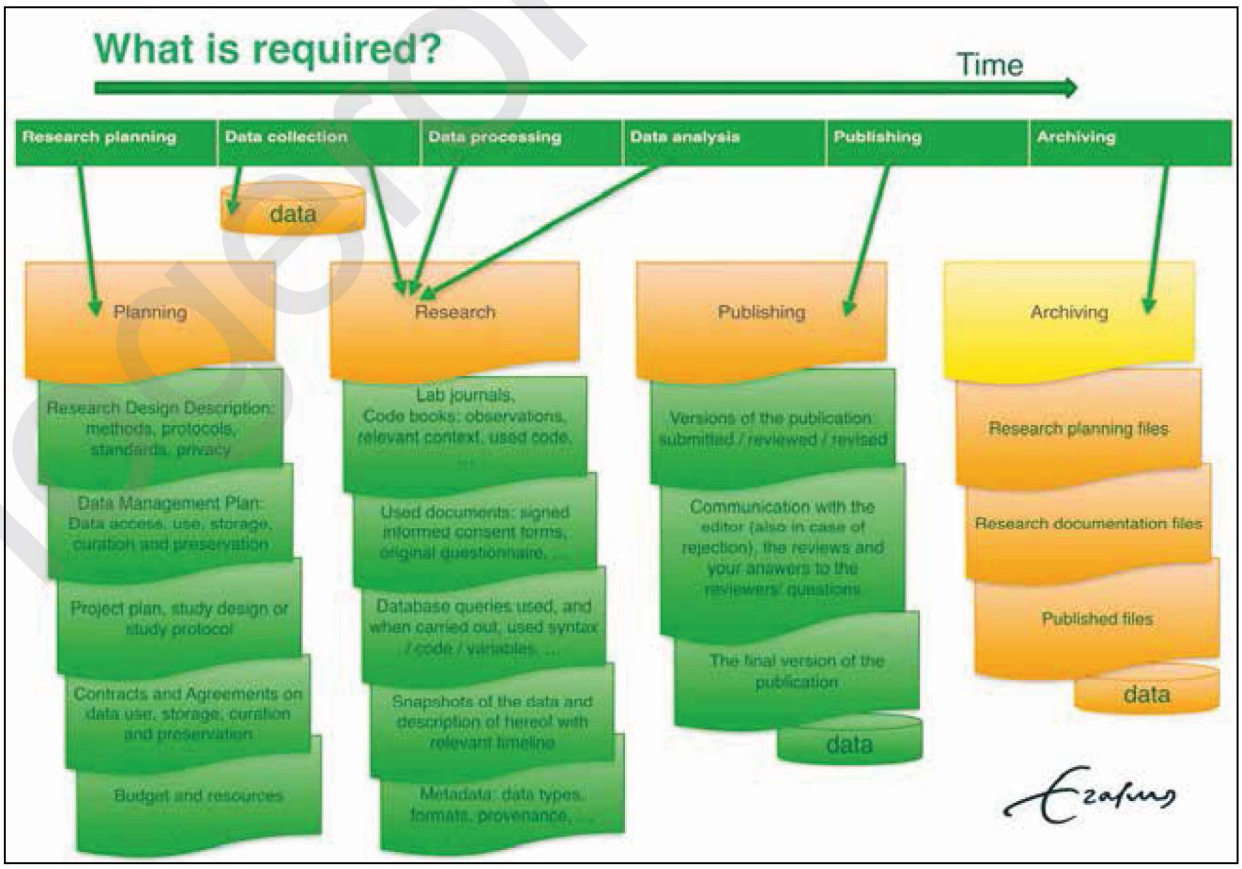
The EUR data management minimum protocol

What is required by whom?

#	type	mandatory /	requirement	VSNU	H2020	NWO	KNAW	DPR
1	research design	mandatory	Project plan, study design or study protocol (including a unique project	03_VSNU,				
2	research design	mandatory	Research question(s) / hypothesis	03_VSNU,				
3	research design	mandatory	Data Management Plan: information about data & data format, metadata content and format, policies for access, sharing, and re-use, long-term storage, curation, preservation and budget		01_H2020, 03_H2020	02_NWO	01_KNAW	
4	research design	mandatory	"Data section" ("Dataparagraaf") describing how the researcher will anticipate reuse (open access) of research data as well as how the privacy of the people involved in the project is guarded.		03_H2020, 04_H2020, 05_H2020,	01_NWO, 04_NWO	02_KNAW	01_DPR
5	agreements / legal	mandatory	Ownership of the research data / IPR agreement	01_VSNU	03_H2020		01_KNAW	
6	agreements / legal	mandatory	Agreements about the management, access, sharing, use and re-use storage and archiving of the data for all involved (also commercial partners) and all audiences		03_H2020		01_KNAW	
7	agreements / legal	mandatory if applicable	non-disclosure agreement (NDA), confidentiality agreement (CA), confidential disclosure agreement (CDA), proprietary information agreement (PIA), secrecy			04_NWO,		
8	agreements / legal	mandatory if applicable	Description of how privacy by design and privacy by default is implemented during the research project and afterwards during the long term preservation of the research data.					01_DPR, 02_DPR,
9	research design	mandatory	Description of the data items / format(s) / metadata scheme(s) / standards / identifiers / cod / labels / variable finitions from publ r commercially available data	3_VSNU	01_H2020, 02_H2020,	02_NWO	01_KNAW	
10	research design	mandatory	The data defini s and criteria use the data provider	3_VSNU	01_H2020	03_NWO		
11	research design	mandatory	Description of researchers, readers, informants, respondents, interviewees,	11_VSNU				

Overall | EUR generic | VSNU requirements | H2020 requirements | NWO requirements | KNAW requirements | Data Protection Regul

Figuur 1 *What is required by whom?*



Figuur 2 *What is required?*

Zwemmen of verzuipen

De volgende basislijst ontstond:

#	type	requirement:document and retain (store / archive)	what is it?
1	research design	Project plan, study design or study protocol (including a unique project identifier)	Project Plan
2	research design	Research question(s) / hypothesis	Research Description
3	research design	Data Management Plan: information about data & data format, metadata content and format, policies for access, sharing, and re-use, long-term storage, curation, preservation and budget	Data Management Plan
4	research design	"Data section" ("Dataparagraaf") describing how the researcher will anticipate reuse (open access) of research data as well as how the privacy of the people involved in the project is guarded.	Research Description
5	agreements / legal	Ownership of the research data / IPR agreement	Agreement
6	agreements / legal	Agreements about the management, access, sharing, use and re-use storage and archiving of the data for all involved (also commercial partners) and all audiences	Agreement
7	research design	Description of the data items / format(s) / metadata scheme(s) / standards / identifiers / codes / labels / variable definitions from public or commercially available data	Research Description
8	research design	The data definitions and criteria used by the data provider	Research Description
9	research design	Description of researchers, readers, informants, respondents, interviewees, participants, funders and others involved in the research project	Research Description
10	research output	The submitted version of the publication	Publication
11	research output	The reviewed and revised version of the publication	Publication
12	research logging	The response from the editor (also in case of rejection) plus the reviews	Publication Process Info
13	research logging	Your answers to the reviewers' questions	Publication Process Info
14	research output	The final version of the publication	Publication
15	research output	The publication-specific version of the dataset	Publication Supporting Data

Figuur 3 *Mandatory RDM Requirements*

Met als bijlage:

#	type	requirement: document and retain (store / archive)
1	agreements / legal	non-disclosure agreement (NDA), confidentiality agreement (CA), confidential disclosure agreement (CDA), proprietary information agreement (PIA), secrecy agreement (SA)
2	agreements / legal	Description of how privacy by design and privacy by default is implemented during the research project and afterwards during the long term preservation of the research data.
3	research design	Description of the used stimuli / input / sources
4	research design	Sampling logic; sample size et cetera (to be transparent with regards to "additional sampling")
5	research design	The boundaries of the population to which the study applies
6	research design	In case of simulated data: the seed and the used software (also: version of the used software)
7	provenance	In case of externally collected and coded data (e.g.from a market research company): maximum information about the provenance of the data.
8	provenance	Clear description of any relationship the researchers/authors may have with the external provider(s) of the data and/or funding body
9	provenance	In case of republishing own previously published work or parts thereof: a correct reference to the source as accepted within the discipline.
9	research logging	Invitation letters and communication to subjects / interviewees
10	research logging	The original questionnaire
11	research logging	Signed informed consent forms (if applicable)
12	research logging	Interview transcripts (authorised or not?)
13	research logging	Member review and / or respondent validation
14	research logging	Digital or digitised field journal - containing observations (if coded (ethogram) the decode key is provided) and relevant context (location, date, time, temperature, weather, ...)
15	research logging	Notes from the interview sessions
16	research logging	Digital scans of "paper" versions
17	research logging	First digital version of "paper" data, i.e. entered in a software application
18	research logging	Printed or saved PDFs of electronic internet documents
19	research logging	Elaborate descriptives of data that cannot be published (i.e., restricted by the terms of a Non-Disclosure Agreement) (if applicable)
20	research logging	Description of the data collection process and any differences from the project plan/study design in the collection protocol or method
21	research logging	Database queries that were used, including when the query was carried out
22	research logging	Description of data snapshots and the corresponding timeline
23	research design	Description of the data screening methods used
24	research design	The syntax, code or variables used in data processing.
25	research logging	Digital or digitised code book(s) - document containing list of code(s) used in research
26	research logging	Digital or digitised laboratory notebook, lab book or lab journal - document containing hypotheses, experiments and initial analysis or interpretation of these experiments.
27	research logging	Documentation of the data processing process: changes to the earlier plans and protocols. What was changed and why.
28	research logging	Documentation of the data processing process: any imputation (unit /item), selection, and data pruning processes.
29	research logging	Documentation of the data processing process: deleted outliers, missing data report, and how labels are linked to the survey items.
30	research design	Documentation of method, variables and underlying items (including labels) used in the final model.
31	research design	Documentation of robustness checks, sensitivity analyses, summary statistics and graphs.
32	research design	Documentation of syntax or code of your final data analysis.
33	research design	Documentation of configuration and version of the software used, including when the analysis was run.
34	research design	Documentation of syntax and output log files of the statistical software, including when the analysis was run.
35	research logging	A detailed log of all the "trials" of an experiment that you run.
36	research logging	A log of anything that was tried and didn't work or didn't yield significant results (everything that is not in the paper to be submitted).
37	research logging	The data snapshots themselves

Figuur 4 *Mandatory if applicable*

Aan de hand van deze lijst gingen Pim Jansen, RDM *communicatie- en awareness officer* en ik in gesprek met onderzoekers van wie bekend was dat ze veel data gebruiken. Een aantal zaken viel ons direct op.

1. Het gebruik van termen als 'mandatory' wierp een drempel op voor een inhoudelijk gesprek. De verplichte onderdelen riepen daarnaast de vraag op of de lijst niet getuigde van wantrouwen en of niet veeleer vertrouwen de basis zou moeten zijn. Lijstjes nodigen tenslotte niet uit tot creativiteit, maar tot zogenaamd 'vinkgedrag'. Onderzoekers zien ze als administratie; de toegevoegde waarde van die inspanning wordt ze niet duidelijk.
2. De bespreking van de basislijst legde tevens disciplinespecifieke verschillen bloot, bijvoorbeeld in de timing. Zo bleek het in een bepaalde discipline gebruikelijk dat na het aanbieden van een artikel de editors van gezaghebbende tijdschriften nog aanwijzingen konden geven die van invloed waren op de onderzoeksmethode, de gebruikte bronnen en soms zelfs deels op de onderzoeksvraag. Dus zou de eis om deze aspecten voorafgaand aan het feitelijk onderzoek te documenteren tot praktische problemen leiden. Met andere woorden: een generiek onderzoeksproces als uitgangspunt nemen leidt tot algemeenheden, terwijl je juist de kern wilt raken.

De paradox

Tevens namen we waar wat ik de 'RDM-paradox' ben gaan noemen. De mate van gedetailleerdheid, zeker van de 'mandatory if applicable'-lijst, schrikt onderzoekers af. Omgekeerd, op het moment dat ze zelf RDM-ondersteuning zoeken, verwachten ze wel degelijk een duidelijk en gedetailleerd beeld van wat partijen vragen, wat de relevante antwoorden hierop zijn, en waarom.
Om die reden zijn we gaan werken met gelaagde informatie: van algemeen naar specifiek. Eerst een paar korte relevante vragen teneinde de aard van het onderzoek en de aard van de databewerking vast te stellen, om vervolgens in te zoomen op de relevante details. In veel gevallen is er namelijk geen sprake van risico's bij de databewerking, en zou je de onderzoeker niet willen lastigvallen met zaken die op hem niet van toepassing zijn.
De paradox is gelegen in het feit dat de desktopresearchbenadering weliswaar onderwerpen oplevert die relevant zijn voor de verantwoordelijke omgang met onderzoeksdata, maar dat ze dit pas in tweede

Zwemmen of verzuipen

> Primair moeten we snel kunnen inschatten of er überhaupt eisen dienen te worden gesteld

instantie, en wellicht slechts voor een deel van de onderzoekers zullen zijn.
Primair moeten we snel kunnen inschatten of er überhaupt eisen dienen te worden gesteld aan de omgang met onderzoeksdata, bijvoorbeeld aan de hand van een korte checklist, zodat het onderzoeksproces voor die eerste inschatting alleen minimaal hinder ondervindt. Alleen in die gevallen dat er wel degelijk aanvullende eisen gelden voor de omgang met onderzoeksgegevens, is het goed een uitgebreide checklist beschikbaar te hebben als referentie.
De focus op onderzoeksgegevens en de problemen die zich hierbij in het algemeen kunnen voordoen, moeten er niet toe leiden dat alle onderzoekers met een te algemeen instrument verplicht een vertaalslag moeten maken naar hun specifieke geval.

Diagnosticerende vragen

Wij stelden een korte lijst samen met diagnosticerende vragen, aan de hand waarvan de onderzoeker snel zelf kan vaststellen of er vervolgacties vereist zijn.

In general, if you answer 'yes' to some of the questions below, please contact us for support.

1. *Is the research project conducted in an international partnership? (relevance: IPR / different legal systems)*
2. *Is this partnership a public-private collaboration? (relevance: IPR / valorisation)*
3. *Is sensitive or confidential data used in the research? (relevance: data protection / privacy / valorisation)*
4. *Will the research project result in information and/or products that will become open access available, or commercially or both? (relevance: IPR)*
5. *Is an infrastructure required for the processing / analysis / storage of the research data beyond which is available at the EUR workplace? (relevance: offering of facilitating services)*
6. *Will the data processing be a manual activity, or is it automated and executed by scripts? (relevance: data protection and IPR)*

Een eerste leermoment voor ons was dat we ons niet zozeer hoefden te focussen op de probleemgevallen. We konden een grote groep onderzoekers direct al helpen met datamanagementplannen, omdat er voor hen geen bijzondere eisen golden voor de verwerking en opslag van onderzoeksgegevens. Bestuurlijk vond een vergelijkbare afweging plaats: we zijn op de goede weg, maar we dienen nog stappen te zetten om tot een bruikbare baseline te komen. We spraken af dat elke faculteit een selectie van junior en senior onderzoekers zou maken met wie wij in gesprek konden gaan om zo disciplinespecifiek inzicht te krijgen. Aldus kregen we een kijkje in de rijkgeschakeerde keuken van onderzoeksmethoden en -invalshoeken.

In de Cambridge RDM-aanpak is de dialoog met de onderzoeker het uitgangspunt. In plaats van *policy driven* eisen te stellen aan onderzoekers, gaat het erom om werkelijk te begrijpen welke rol data spelen in het onderzoek en waar risico's zijn die passende maatregelen behoeven.
Na de vele gesprekken vormde zich bij ons een realistischer beeld van de hedendaags onderzoekspraktijk. We leerden veel. Onze belangrijkste bevindingen vallen uiteen in drie aspecten.

1. Principes
De RDM-baseline van de EUR zou gebaseerd moeten zijn op deze algemeen aanvaarde principes:

1. *Build valid science.*
3. *Researchers are responsible for data management. Do not create a web of responsibilities but provide researchers with fitting tools.*
4. *For quantitative research: if you use data, the raw data, the code and the processed data used should all be available*
5. *For qualitative research: if you use data, the method used should be documented*
6. *The basic research idea and method to retrieve data / the type of data used should be basically documented (typically around the time research funding (internal or external) is applied for.*

> We kregen een kijkje in de rijkgeschakeerde keuken van onderzoeksmethoden en -invalshoeken

Zwemmen of verzuipen

> Een ander zei: 'Ik wil niet meer hoeven zeggen: waar zijn mijn data?'

Een van de onderzoekers die wij spraken formuleerde het aldus: "Ga uit van de gedragscode van de universitaire koepel VSNU en preciseer een zeer beperkt aantal minimumeisen, gekoppeld aan de Europese afspraken rond wetenschappelijke integriteit. Het ultieme doel moet zijn de bouw van valide wetenschap. Zijn wij een valide kennisbasis aan het bouwen?"

2. Best practices in dataomgang
Een van onze gesprekspartners: *"I prefer my ideas to my data. Data are to test a theory. I care about theories, not data. I share all my data immediately online. You shouldn't keep your data for more than one paper. We look less like a fool if someone finds a mistake on data made public. You are willing to take the risk that they find the mistake. You are showing good faith. In my field I use two datasets in one paper, not many papers on one dataset."*
Een andere: "Ik wil niet meer hoeven zeggen: waar zijn mijn data?"

3. Aanbevelingen voor dataomgang
Een onderzoeker: "Persoonlijk ben ik er niet tegen om bewerkte data te delen. Een korte beschrijving aan de hand van heldere richtlijnen zal me misschien een halve dag kosten."
Een andere: "Je moet als onderzoeker een status aan je onderzoekgegevens kunnen geven – toegankelijk of niet toegankelijk. En erbij kunnen zetten tot wie je je dient te richten om toegang te krijgen."
Een derde: "Ik weet dat er collega's zijn die geen zin hebben om hun onderzoeksdata te beschrijven, maar dit moet geen kwestie van meningen zijn. De universiteit zou hierover een bindende uitspraak moeten doen. Ik wil als professional de eis van een goede omgang met onderzoeksgegevens expliciet hebben."

4. Aanbevelingen voor onderzoeksondersteuning
Een onderzoeker: "Achteraf gezien zou ik wel training gehad willen hebben aan het begin van mijn loopbaan over vragen als: van wie zijn de data? Welke afspraken moet je maken?"
Een ander: "Data worden soms in software gemaakt. Hulp bij het onderhoud van de documentatie van je software zou handig zijn – alleen al omdat je weet dat iemand anders ook naar je code gaat kijken. Ik zou daar graag structureel ondersteuning bij krijgen."
Een derde: "Standaardcontracten voor patenten et cetera zouden zeer welkom zijn. Hulp bij contractonderhandelingen idem. Ik stond er wat alleen voor."

5. Aanbeveling voor vervolgstappen
"Datamanagement is een cultuurkwestie, een mentaliteitskwestie – geen systeemkwestie. Het is een onderzoekskwaliteitsvraagstuk. Start een programma dat onderzoekskwaliteit belangrijk laat worden. Een goede onderzoeker dient aan datamanagement te doen. Mensen moeten het gevoel krijgen: ik kan niet achterblijven. *Thought leaders* moeten het goede voorbeeld geven."

Onverkort overgenomen

Met andere woorden: onderzoekers zien researchdatamanagement als een integraal aspect van een professionele attitude. Ze zitten niet te wachten op beleid dat zich richt op controle en dat dus op wantrouwen lijkt te zijn gebaseerd. Ze vragen primair om realistische *guidelines* en vooral praktische ondersteuning: advies, training, voorzieningen en infrastructuur.

Dit inzicht was ons tweede leermoment. Onze focus klapte om. We zijn het baselineprotocol gaan zien en benoemen als de professionele omgang met onderzoeksgegevens binnen de EUR, gebaseerd op de grootste gemene deler.

The following was generally accepted as a RDM baseline:
1. *The research question or hypothesis*
2. *The research method(s) chosen*
3. *The data actually used to support the research conclusions and/or recommendations*
4. *The list of known people for whom the data (3) are accessible*

> Dit inzicht was ons tweede leermoment; onze focus klapte om

5. *Documentation of the actions taken with the data and the software used*
6. *If available: the raw dataset(s)*

Als sluitstuk deden we het college van bestuur de volgende aanbevelingen:

1. *Declare that the EUR is committed to the FAIR principles[2] for data: Findable, Accessible, Interoperable and Reusable. Provided valid opt out grounds and embargo periods, the EUR intends to make research data findable and accessible.*

2. *Affirm the baseline protocol RDM for EUR research, as the largest common denominator of research professionalism with respect to research data within the EUR.*

3. *Request the RDM community – in which the EUR institutes and faculties participate – to develop specific services for implementing the RDM baseline within the research services framework.*

In februari 2017 heeft het college deze aanbevelingen onverkort overgenomen.

Marlon Domingus
is projectmanager research data management aan de Erasmus Universiteit Rotterdam en is verbonden aan het LCRDM (Landelijk Coördinatiepunt Research Data Management)

Noten

1 http://biorxiv.org/content/early/2017/01/28/103895
2 http://www.nature.com/articles/sdata201618