Data Management bij wetenschappelijk onderzoek méér dan alleen storage!!!

Werkgroep:
Alisa Westerhof (UU), Tessa Pronk (UU), Annemiek van der Kuil (3TU & TUD), Annemie Mordant (UM)

Aanleiding:
Tijdens de H2020 bijeenkomst (28-9-2015) te Den Haag met betrekking tot kosten research datamanagement (RDM), hebben Alisa Westerhof (UU), Annemiek van der Kuil (3TU) en Annemie Mordant (UM) zich aangemeld om zich te buigen over een goede definitie (gemeenschappelijke taal) voor Research Data Management. Later heeft zich Tessa Pronk (UU) hierbij aangesloten.

Doel:
Een praktisch bruikbaar overzicht maken, waarin mogelijke kosten per activiteit van het onderzoeksproces zichtbaar zijn. Het betreft een 'format' dat zowel bruikbaar moet zijn voor onderzoekers, data specialisten en subsidie aanvragers als subsidie verstrekkers. In zijn algemeenheid kan wel gesteld worden, dat hoe eerder en beter bepaalde activiteiten ondernomen worden (zoals bijv. metadateren), hoe lager de kosten zullen zijn. Een goed datamanagementplan vóór start dataverzameling helpt om extra werk en kosten te voorkomen.

Werkwijze:
We hebben ons eerst gebogen over een goede definitie van research data management en ervoor gekozen ons niet te beperken tot één definitie. Op deze manier wordt een beter beeld verschaft van de breedte van research data management en de hierbij behorende kosten.
De kosten hebben we weergegeven in tabelvorm. Als brondocument hebben we de *UK Data Service Data Management costing tool* genomen. Deze hebben we ingedeeld in zes categorieën die te onderscheiden zijn voor datamanagement in het onderzoeksproces. Deze categorieën zijn ook terug te vinden in datamanagement templates en checklists van diverse financiers en universiteiten. Als aanvulling hierop hebben we een kolom 'cost' toegevoegd. Schattingen voor het kwantificeren van bedragen in deze kolom zijn slechts ter indicatie van de orde van grootte. Bijvoorbeeld, het domein waarin het onderzoek zich afspeelt heeft invloed op de hoogte van de datamanagement kosten door specifieke richtlijnen. Per project zullen de kosten dus steeds opnieuw ingeschat moeten worden. Het document is niet compleet en heeft de status van groeidocument. Het is de bedoeling dat het wordt aangevuld cq. verbeterd.

Advies werkgroep:
- Het format betreft een eerste aanzet. Vandaar dat er behoefte is aan een kritische blik vanuit andere instellingen en aanvullingen vanuit kosten optiek van reeds opgestarte H2020 projecten.
- Belangrijk is dat het document een goed vindbare plaats krijgt.
- Opslag *ná* onderzoek, incl. datamanagement en persistentie, zou ook vanuit de funder gefinancierd moeten kunnen worden. Wie zou dit op de agenda kunnen zetten?

Definities van datamanagement.

Wat er allemaal te vangen valt onder datamanagement hangt af van de definitie die je hiervoor hanteert. Er zijn vele definities, en ze zijn in hun eigen context allen relevant. We geven hier een aantal mogelijke definities.

(Wikipedia): "Data management comprises all the disciplines related to managing data as a valuable resource"

(Casrai Data Dictionary): The activities of data policies, data planning, data element standardization, information management control, data synchronization, data sharing, and database development, including practices and projects that acquire, control, protect, deliver and enhance the value of data and information."

(businessdictionary.com): "Administrative process by which the required data is acquired, validated, stored, protected, and processed, and by which its accessibility, reliability, and timeliness is ensured to satisfy the needs of the data users".

(UM-MEMIC): "whole process of Data collection, Information management and Knowledge extraction and all the activities carried out on these data during and after the research projects".

(DTL): "In the FAIR Data approach, data should be: Findable – (vindbaar); Accessible –(toegankelijk); Interoperable –(interoperabel); Reusable – (herbruikbaar)".

(ZonMW): "Alle inspanningen gaan om het resultaat aan het einde van het project, namelijk dat onderzoeksresultaten en onderliggende data herbruikbaar en verifieerbaar worden. De subsidieontvanger moet dan laten zien dat de dataverzameling vindbaar, toegankelijk, interoperabel en herbruikbaar is. Data die bovendien van bron naar ruwe data repliceerbaar is."

Zie ook:

- https://en.wikipedia.org/wiki/Data_management
- http://dictionary.casrai.org/Data_management
- http://www.businessdictionary.com/definition/data-management.html
- http://memicmaastricht.nl/
- http://www.dtls.nl/fair-data/
- http://www.zonmw.nl/nl/programmas/programma-detail/toegang-tot-data-ttdata/pilot/

Kortom: Data Management bij wetenschappelijk onderzoek is méér dan alleen storage van onderzoeksdata. Omdat datamanagement nieuw is als expliciet onderdeel van een onderzoeksvoorstel en onderzoekskosten, komen onderzoekers vaak niet of in een laat stadium tot het budgetteren van de kosten van datamanagement. Dan is de tendens om alleen de kosten voor storage te begroten.

Om onderzoekers te helpen bij het begroten van datamanagement kosten is onderstaande *Guide Research Data Management and Costs* opgesteld.

Guide Research Data Management and Costs

DMP phase	ACTIVITY	COMMENTS AND SUGGESTIONS	COSTS
Preparing	**Make a Data Management Plan**	• make a DMP before you start creating data; make decisions about managing your data; consider how you can process, analyse, preserve and share your data • check if there is a department within your organization to support data management planning	2 hrs to 2 days, depending on the complexity of your project
1. Data Collection	**Acquiring External datasets** • Do you plan to use existing data, and is the data available at a commercial partner?	• your library may be able to help you acquire a license to a crucial database • in research data repositories, data can be available at no or low costs	Example: A faculty licence on a database for macro-economic analyses: €18.000/y
1. Data Collection	**Formatting and organising** • Are your data files, spreadsheets, measurements, interview transcripts, records etc. all in a uniform format or style? • Are files, records and items in the collection clearly named with unique file names and well organised?	• if planned beforehand by developing templates and data entry forms for individual data files (transcripts, spreadsheets, databases) and by constructing clear file structures – low or no additional cost • if needed afterwards – higher cost	Per project organize style, format, names can be done by a student assistant at level 1* salary *or* data manager at level 2* salary
1. Data Collection	**Transcription** • Will you transcribe qualitative data (e.g. recorded interviews or focus group sessions) as part of your research; or will you need to do this specifically so data can be more easily shared and reused? • Is full or partial transcription needed? • Is translation needed? • Will you need to develop a standard transcription template or transcription guidelines, to ensure consistent	• if part of research practice – very low or no additional cost • if not planned as part of research practice – potentially high additional cost • is additional hardware /software needed ? • consider cost of (time needed for) developing procedures, templates and guidance for transcribers	Example: Time needed for transcription - four to eight hours per hour recording, i.e. see transcribing calculator: http://www.socialsciences.manchester.ac.uk/morgancentre/methods-and-resources/toolkits/toolkit-8/

	formatting?		
1. Data Collection	**Consent for data sharing** • Do you need to ask participants for their consent for data to be shared? • Consent is essential for research in the domain of health/life sciences also for qualitative interviews	• when consent for data sharing is considered as part of standard consent procedures early in research – very low or no additional cost • when participants need to be re-contacted or re-visited to obtain -active consent– could be high cost • does this require extra preparation of information sheets and consent forms; extra time for consent discussions; or training of interviewers?	Student assistant at level 1* salary *or* data manager at level 2* salary
1. Data Collection	**Data transfer** • Are special measures needed to transfer data from mobile devices, from fieldwork sites or from home equipment to a central work server?	• is software or hardware needed for data transfer, for encryption of confidential data before transfer, or for synchronisation of data files across sites?	Free encryption or data transfer software (i.e. SurfFileSender) is available in most cases
2. Data Documentation	**Data description and Metadata** • Are data in a spreadsheet, database or data warehouse clearly marked with variable, variable labels and value labels, code descriptions, missing value descriptions, etc.? • Are validated questionnaires and standard coding used? • Are labels consistent? • Are files, records and items in the collection clearly described with well-defined metadata or a metadata standard to interpret the relations between them and to quickly select and understand the content. • Do textual data like interview transcripts need description of context, e.g. included as a heading page?	• if data description is carried out as part of data creation, data input or data transcription – low or no additional cost • if needed to be added or harmonized afterwards – higher cost • codebooks for datasets can often be easily exported from software packages	Examples: 4 hrs per single experiment (120 measurements) filling in 60 required metadata fields, with assistance of a data manager at level 2* salary Two to three weeks are costed into an average two year research grant application to prepare and collate materials for deposit http://www.data-archive.ac.uk/help/user-faq
2. Data Documentation	**Documentation** • Do you have documentation for the	• often essential contextual and methods documentation will be written up in publications and reports	Researcher at level 2* salary.

	data that describes the context and methodology of how data were gathered, created, processed and quality controlled?	• if all data creation steps are well documented and documentation is kept well organised during research – low or no additional cost • if documentation to be written or compiled specifically afterwards – higher cost	
3. Data Storage & Back-up	**Data backup** • Does the institution provide regular backup or not? • Consider how frequently backups should be done, how many backups should be stored.	• institutional backup – included in standard indirect cost /overheads • additional backup needed – cost according to number of copies to be kept, frequency of backup and storage media needed	Examples: University drive €0.80 per GB/y Cloud: €0.30 per GB/y 2 x Harddrive: €0.14 per GB (single purchase)
3. Data Storage & Back-up	**Data storage** • How much data storage space is needed for the entire duration of the project? • Do you need to set up a data model and accompanying database for the data?	• if storage is provided by the institution – cost is included in standard indirect costs or overheads • if additional storage needed – cost server/ disk space, as well as the cost of setting up and maintenance • Do you need a data warehouse or a database architect?	Example: Cloud Database as a service: €160/Month (storage 5GB transfer 30GB) Database architect at level 2* salary
4. Data Access & Security	**Data Access** • Do external people require access to research data?	• does remote access via VPN or secure FTP need to be arranged for external people?	Mostly researchers can make use of existing, free services
4. Data Access & Security	**Data security** • Is there an institutional server available where you can store your data safely? • Protect data from unauthorised access or use or from disclosure	• for confidential or privacy sensitive data, determining conditions for controlling access to shared data may require extra time and discussion • can security be arranged by institutional IT services or is extra software/hardware needed? • data files may need encrypting before storage or transfers	Example: TTP (trusted third party), dependent on pseudonymisation type, ca. €1.000- €30.000 Existing encryption services could be used at no costs
5. Data Preservation & Archiving	**File format** • Do data need to be converted to a standard or open format with long-term validity for long-term	• is additional software or hardware needed for conversion? • for audio-visual data, converting to open digital formats can be time-consuming or require special equipment	Researcher at level 2* salary

	preservation?	and/or software • for databases, conversions may require checking for truncation, loss of metadata or annotation, loss of relationships, etc.	
6. Data Sharing & Reuse	**Anonymisation** • Do you need to remove identifying information or conceal the identity of participants (e.g. using pseudonyms) before data can be shared? • Anonymisation needs to be consistent throughout a data collection.	• if anonymisation is planned before data collection or transcription/digitisation – cost can be lowered • for audio-visual data – anonymising/editing voices or faces can be very costly and could reduce the usefulness of data • for quantitative data (e.g. survey data) – low cost if identifiers are a priori excluded from data files, are easy to remove, or identifiable variables are coded to avoid disclosure; cost may be higher if variables need recoding afterwards to avoid disclosure • for qualitative textual data (e.g. interview transcripts) – costs can be reduced if anonymisation is carried out during transcription (or at least highlighted/coded during transcription) • cost depends on how sensitive or complex data are and how much identifying information is recorded in the data – if only removal of names is required, cost is low; pseudonymisation will require more time • for files received of participants, check file properties and edit to remove disclosive information such as editor/author name	Example: Transcribing / simultaneously anonymizing audio (speech): up until one hour per 5 minute fragment (depending on the preciseness level of transcribing) Student assistant at level 1* salary Free software is available
6. Data Sharing & Reuse	**Copyright** • Do other parties hold copyright in the data? • Do you need to seek copyright clearance before sharing data?	• is time required to seek copyright clearance? • is legal advice required?	Juridical advice at level 3* salary
6. Data Sharing & Reuse	**Data sharing** • Will your data be deposited with a data centre or institutional repository? • Which requirements exist to prepare data to particular standards e.g. regarding documentation or format?	• how long is the data required to be available, • a public repository/ data centre/ journal can provide you with the possibility to share your data for reuse. Find out what the cost are of data deposit and/or longer-term storage per year cost in time and effort needed to prepare the data for sharing and preservation	Examples: Completing a data repository upload form (i.e. 3TU Datacentrum or DANS) may take 15 min to 4 hrs

	• Do structured metadata need to be created when data are shared via a data centre or archive, e.g. completing a deposit form for the UK Data Archive? • What data will be retained and what not?	• data centres will have their own metadata forms. Consider using these on beforehand	Dryad €110 once (max 20 GB) DataverseNL €3.60 per GB/year Cloud Database as a service: €160 /month (storage 5 GB, transfer 30 GB)
6. Data Sharing & Reuse	**Data cleaning** • Do quantitative data need to be cleaned, checked or verified before sharing, e.g. check validity of codes used, check for anomalous values? • Will data match documentation, e.g. same number of variables, cases, records, files? • Does textual information in data need to be spell-checked? • Do you need to combine your data with other datasets for your research	• data cleaning takes time • if carried out as part of data entry and preparation before data analysis – low additional cost • if needed afterwards – higher cost	Example: Data cleaning service: €270 to well over €1800 http://datascopic.net/cost-of-data-cleansing/ Researcher/data manager at level 2* salary
6. Data Sharing & Reuse	**Digitisation** • Do analogue or paper-based research data (maps. newspaper clippings, photographs, images, text) need to be digitised to increase their potential for sharing?	• is additional equipment or software needed for scanning or conversion? • if simply image scanning of text – relatively low cost • if Optical Character Recognition required, with manual checking for accuracy (revising entire scanned text) – may be high cost • if manual data entry or typing needed, e.g. to digitise tabular data – may be high cost	Example: Digitisation €0.50 per page (few pages) *OR* €320-390 per 1000 pages (OCR included)
Overall	**Roles and responsibilities** • Do you need to allocate roles and responsibilities for various data management activities?	• if multiple partner institutions, researchers or funders are involved in research – consider cost of data management planning meetings or discussions	Travel costs, lunch, time
Overall	**Operationalising data management** • What measures are needed to implement and operationalise data management throughout the research	• do you need extra time and resources to implement data management throughout your research, e.g. regular team meetings, setting up a collaborative research environment?	Data manager at level 2* salary

	lifecycle?	• if staff training is required - higher cost • do you need a dedicated data manager?	

* Salary:
Level 1 (i.e. student assistant) ~ 17 euro per hour.
Level 2 (researcher, data manager) ~60 euro per hour.
Level 3 (external expert) ~160 euro per hour.

Werkgroep:

Alisa Westerhof
Organisatie: Universiteit Utrecht (UU); Information and Technology Services
Functie: Projectleider Onderzoek Systemen en Datamanagement
Email: A.Westerhof@uu.nl
Telefoon: 06 1043 2840

Tessa Pronk ;
Organisatie: Universiteit Utrecht (UU); Universiteitsbibliotheek
Functie: Data- en informatiespecialist
Email: T.E.Pronk@uu.nl
Telefoon: 030 2536754

Annemiek van der Kuil;
Organisatie: 3TU.Datacentrum & TU Delft Library
Functie: Research Data Officer
Email: a.vanderkuil@tudelft.nl
Telefoon: 015 27 85 540

Annemie Mordant;
Organisatie: Universiteit Maastricht & MUMC+
Functie: hoofd MEMIC, centrum voor data- en informatiemanagement
Email: a.mordant@maastrichtuniversity.nl
Telefoon: 043 3882341