

Research Data Management: more than just storage

draft translation

Working Group Members:

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

Incentive:

During the H2020 meeting on financial aspects of Research Data Management (RDM) which took place on 28-9-2015 in the Hague, Netherlands, Alisa Westerhof (UU), Annemiek van der Kuil (3TU) and Annemie Mordant (UM) volunteered to brainstorm on a good definition (in common language) for Research Data Management. At a later stage, the Working Group was broadened with Tessa Pronk (UU).

Purpose:

To create a practical and usable overview of possible Costs per Activity within each phase of the Research Process. The format has to be useful not only for Researchers, Data Specialists, but also for Funding Applicants and Funding Bodies. An early start for certain activities within the Data Life Cycle (like for instance metadata specification) will lower the costs for Data Management in the run of the Life Cycle. A good Data Management Plan before gathering the Research Data also helps to prevent extra efforts and costs later on in the Research Cycle.

Approach:

We first focused on a good definition of Research Data Management and allowed ourselves to choose more variations. This provides a better coverage of the full perspective of Research Data Management and connecting costs during the full Research Data Life Cycle.

The costs are visualised in a Table. The [UK Data Service Data Management costing tool](#) was used as a starting point. This tool was split up in the six categories connected to Data Management throughout the full Research Process. These categories were also used in Data Management Templates and in Checklists of the various Funders and Universities. In addition to this, we added a column for Costs. Estimates to quantify the sums in this column only serve as an indication of approximate height. Costs for Data Management are for example influenced by specific guidelines in the domain in which the research is carried out. For each project the costs will need to be estimated over again. The document is not complete and is meant as an organic document. Additions and improvements are very welcome.

Recommendations of the Working Group:

- The Table Format must be seen as a first attempt. A critical look, based on experiences in other institutions, and additions to the cost perspective, based on experiences in current H2020 Projects, are most welcome.
- The document must be easy accessible and easy findable.
- Storage after active Research, including Data Management and Persistency, also should be funded by the funder. Which party will be the most appropriate to put this on the agenda?

Definitions of Data Management:

The broadness of the term Data Management depends on the definition. Many definitions are available, and they all have their own relevance in their own context. To list some possible definitions:

Data management comprises all the disciplines related to managing data as a valuable resource
Wikipedia, https://en.wikipedia.org/wiki/Data_management

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.
Casrai Data Dictionary, http://dictionary.casrai.org/Data_management

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.
businessdictionary.com, <http://www.businessdictionary.com/definition/data-management.html>

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.
UM-MEMIC, <http://memicmaastricht.nl/>

In the FAIR Data approach, data should be: Findable; Accessible; Interoperable; Reusable. DTL, <http://www.dtls.nl/fair-data/>

All efforts focus on the final project result: all research results and underlying/supporting data are reusable and verifiable. Furthermore data should be replicable from source to raw data.
<http://www.zonmw.nl/nl/programmas/programma-detail/toegang-tot-data-ttdata/pilot/>

Conclusion:

Research Data Management is more than only the storage of Research Data. Data Management is a rather new explicit element of a Research Proposal and of the Research Costs. This leads to the situation that Researchers don't budget the Costs of Data Management at all, or include this in a very late stage. Most chosen tendency is in that case to only include the Costs for Data Storage in the Research Budget.

To help Researchers in calculating their Research Data Management Costs, the following Guide has been drafted:

Guide Research Data Management and Costs

DMP phase	ACTIVITY	COMMENTS AND SUGGESTIONS	COSTS
Preparing	Make a Data Management Plan	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Do you plan to use existing data, and is the data available at a commercial partner? 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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? 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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 <i>or</i> data manager at level 2* salary
1. Data Collection	Data transfer <ul style="list-style-type: none"> Are special measures needed to transfer data from mobile devices, from fieldwork sites or from home equipment to a central work server? 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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? 	<ul style="list-style-type: none"> 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 	<p>Examples: 4 hrs per single experiment (120 measurements) filling in 60 required metadata fields, with assistance of a data manager at level 2* salary</p> <p>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</p>
2. Data Documentation	Documentation <ul style="list-style-type: none"> Do you have documentation for the 	<ul style="list-style-type: none"> 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?	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Does the institution provide regular backup or not? Consider how frequently backups should be done, how many backups should be stored. 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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? 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Do external people require access to research data? 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Is there an institutional server available where you can store your data safely? Protect data from unauthorised access or use or from disclosure 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Do data need to be converted to a standard or open format with long-term validity for long-term 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> for databases, conversions may require checking for truncation, loss of metadata or annotation, loss of relationships, etc. 	
6. Data Sharing & Reuse	Anonymisation <ul style="list-style-type: none"> 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. 	<ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Do other parties hold copyright in the data? Do you need to seek copyright clearance before sharing data? 	<ul style="list-style-type: none"> is time required to seek copyright clearance? is legal advice required? 	Juridical advice at level 3* salary
6. Data Sharing & Reuse	Data sharing <ul style="list-style-type: none"> 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? 	<ul style="list-style-type: none"> 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

	<ul style="list-style-type: none"> 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? 	<ul style="list-style-type: none"> data centres will have their own metadata forms. Consider using these on beforehand 	<p>Dryad €110 once (max 20 GB)</p> <p>DataverseNL €3.60 per GB/year</p> <p>Cloud Database as a service: €160 /month (storage 5 GB, transfer 30 GB)</p>
6. Data Sharing & Reuse	<p>Data cleaning</p> <ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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 	<p>Example: Data cleaning service: €270 to well over €1800 http://datascopic.net/cost-of-data-cleansing/</p> <p>Researcher/data manager at level 2* salary</p>
6. Data Sharing & Reuse	<p>Digitisation</p> <ul style="list-style-type: none"> Do analogue or paper-based research data (maps, newspaper clippings, photographs, images, text) need to be digitised to increase their potential for sharing? 	<ul style="list-style-type: none"> 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 	<p>Example: Digitisation €0.50 per page (few pages) OR €320-390 per 1000 pages (OCR included)</p>
Overall	<p>Roles and responsibilities</p> <ul style="list-style-type: none"> Do you need to allocate roles and responsibilities for various data management activities? 	<ul style="list-style-type: none"> if multiple partner institutions, researchers or funders are involved in research – consider cost of data management planning meetings or discussions 	<p>Travel costs, lunch, time</p>
Overall	<p>Operationalising data management</p> <ul style="list-style-type: none"> What measures are needed to implement and operationalise data management throughout the research 	<ul style="list-style-type: none"> 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? 	<p>Data manager at level 2* salary</p>

	lifecycle?	<ul style="list-style-type: none">• 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.



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Working Group:

Alisa Westerhof
Universiteit Utrecht (UU); Information and Technology Services
Projectleider Onderzoek Systemen en Datamanagement
A.Westerhof@uu.nl

Tessa Pronk
Universiteit Utrecht (UU); Universiteitsbibliotheek
Data- en informatiespecialist
T.E.Pronk@uu.nl

Annemiek van der Kuil
3TU.Datacentrum & TU Delft Library
Research Data Officer
a.vanderkuil@tudelft.nl

Annemie Mordant
Universiteit Maastricht & MUMC+
hoofd MEMIC, centrum voor data- en informatiemanagement
a.mordant@maastrichtuniversity.nl