Data Management Plan

Full text version of the correspondent lesson of the German project FoDaKo[1] (Forschungsdatenmanagement in Kooperation)

Torsten Rathmann
Bergische Universität Wuppertal
Universitätsbibliothek
Gaußstr. 20
D-42119 Wuppertal
ORCID: 0000-0001-5880-1546

Abstract

This brochure describes how a data management plan can facilitate research data management, which topics should be addressed and how data management plan, milestone plan, and cost schedule are related. One section informs about software tools for setting up data management plans.

Introduction

A data management plan (DMP) is a document that describes how data is handled during a research project and after the project is completed[2]. How can a data management plan help? It determines a mandatory basis for common handling of research data. Especially if several project partners have to work together, such a common document can define responsibilities and regularize access rights. It can also facilitate coordination of the project partners in a timely sense and by defining the basis of workflows. Further, a data management plan may help to recognize problems concerning data management and to develop solutions. It may help to avoid data loss and security holes.

Another purpose of a data management plan that becomes more important is its use as an annex to a proposal for promotion. In the past few years, more and more institutions offering research funding expect a data management strategy together with a proposal. Such a data management strategy can be the first version of a data management plan. In this way, a data management plan not only facilitates handling of research data but also increases the chance of funding.

[1] https://www.fodako.de/
[2] https://en.wikipedia.org/wiki/Data_management_plan, retrieved on 02.10.2018

Main objectives of a data management plan

Data management plans are as different as research data itself. Main objectives may be for example[3]

- a description of the expected data
- strategies to protect research participants and procedures to obtain informed consent
- third party rights (e.g. privacy, intellectual property rights)
- conditions of re-use
- a specification of the corresponding metadata
- ordering and naming conventions
- a description of the workflows for generation, ingestion, sharing and publication of research data
- a planning of the long-term archiving, e.g. safety and security concepts for the bitstream preservation and a catalogue of tasks for the curation of data
- a specification of responsibilities

Little example: How you should not do it

Dear BMBF,

I want to emphasize that your additional request to enrich our proposal with a further annex, a data management strategy, has caused great joy in our working group. Here are the details:

Our primary data are created by our measuring equipment in a proprietary and rarely documented binary format. The scarce metadata are in the header of these files, i.e. data and metadata are most intimately connected. We do not complete these headers. Instead, we name the entirety of headers "Sparse Density Header Matrix" matching the other expletives in our "Action Planning".

The names of the output files, about 50000, are serially numbered by our measuring equipment, i.e. the file names are equivalently meaningful; the gaps in the numbering indicate how many failures had to be absorbed and how difficult it was to get these data. Especially important for traceability are the control and sizing files. These files are named using Arabic characters and digits matching the output file numbering with Arabic digits. As you asked about standards we obey, the characters encoding of our file names strictly follows ISO 8859-6 (Arabic). This has been a good decision since our programmer is Arab and can only speak very little English.

The control and sizing files can be found from number 1024 on. In case a re-user has been reaching this number, we add extra commas to our control files, which are in format CSV (Comma Separated Values), since re-users shall deal with our data model in detail and shall not simply import the files automatically.

[3] partly from cessda, Research data management plan, https://www.cessda.eu/content/download/239/2381/file/CESSDA%20User%20Guide%20for%20data%20management_2_Research%20data%20management%20plan.pdf, retrieved on 02.10.2018

We store our research data on tape cartridges bought about ten years ago and we still have more than enough. The reading device can no longer be built in a modern desktop computer but this is not a problem for us since we have a big pool of old computers at our disposal.

On your request, the project has been amended by a quality assurance. In the controversy with the controller, we will defend every error toughly until it cannot be denied any longer. We will name corrected datasets as the predecessor version without any notice. This is timesaving and does not alarm potential re-users.

Our scheduling is as easy as efficient. The research data management will be put into the holiday season. The colleagues can then decide on themselves if they prefer managing their data or going on holidays. Another opportunity for research data management is after the end of the working contract.

In case data were published despite this pioneering data management, we would provide a license to the data, which allows re-users to utilize this data for the reproduction of our previously published results only.

Of course, this is satire. Nonetheless, except the first paragraph and the salutation the mentioned topics should also occur in a real data management plan:

- Which types of data will be generated or evaluated ◊
- Data format
- Naming convention
- Used standards ◊
- Where and how the data will be stored ◊
- Quality assurance ◊
- Version control
- Scheduling ◊
- Access, data publishing and license ◊

The red lozenges mark those topics that are recommended[4] by the DFG. For proposals to the DFG the list is nearly complete, only *third-party rights* and *how much of the data could be relevant to other research contexts* are missing in the satire. Of course, a data management plan may contain much more but also less. If a topic does not concern your research, if this detail is still unclear or if you simply do not want to describe the details now, you may leave them out. An incomplete data management plan is better than nothing since, even if a number of details are still unclear, those points that are clear should be communicated at least inside the research project.

A data management plan is usually not a fixed document. It can be extended and altered. For example, a first version may be used as an annex to the proposal, a second version with more details during the project and a third version for long-term archiving after the end of the project. Best time to start work on the data management plan is before proposal since then data

[4] http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf, retrieved on 01.10.2018

management plan, milestone plan and cost schedule can be elaborated together and easily be coordinated.

Time schedule

Research itself usually consumes more time but the effort for research data management should not be underestimated. Not only the data is to be handled, also metadata have to be created. Data that shall be shared usually needs to be augmented with more information, for example data documentation, quality control report and a license for re-use.

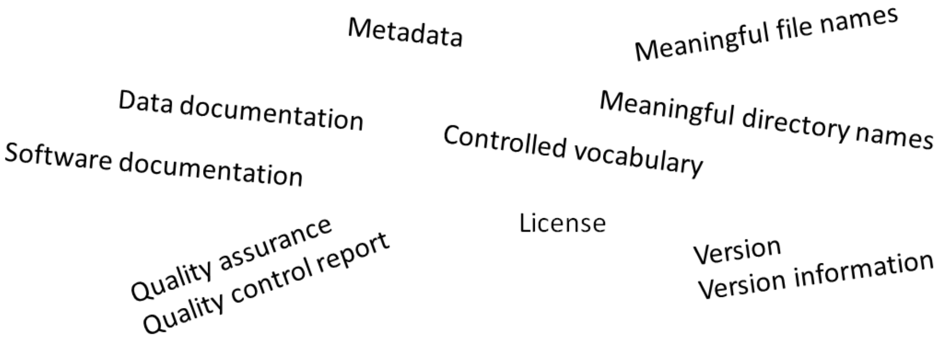

Figure 1: necessary or desirable extra information belonging to research data

In case of errors in the data a new data version may be necessary. Arrangements should be made for enabling several different versions already at the beginning of the project, e.g. an additional directory level in the file tree. Version information should sum up what has been changed. Errors in the data should briefly be described in errata.

Also utilization of a controlled vocabulary, i.e. a list of allowed entries for a part of the metadata, file headers, unit names and so on should be declared early. This concerns naming conventions for files and directories too. All these topics should be mentioned in the data management plan.

A later data management with all the components in Figure 1 is in principle possible but much more difficult and time consuming since the remembrance of details declines quickly and some of the data producers may have leaved taking with them their knowledge. Therefore, a late data management would cause the time schedule to totter. Best practice is a common development of data management plan and milestone plan before the project has been started. At that time data management steps could easily be integrated. Another positive effect: work on subsequent datasets can be organised in a way that data do not arrive all at the same time, see Figure 2. In this way, overload of staff and devices can be avoided.

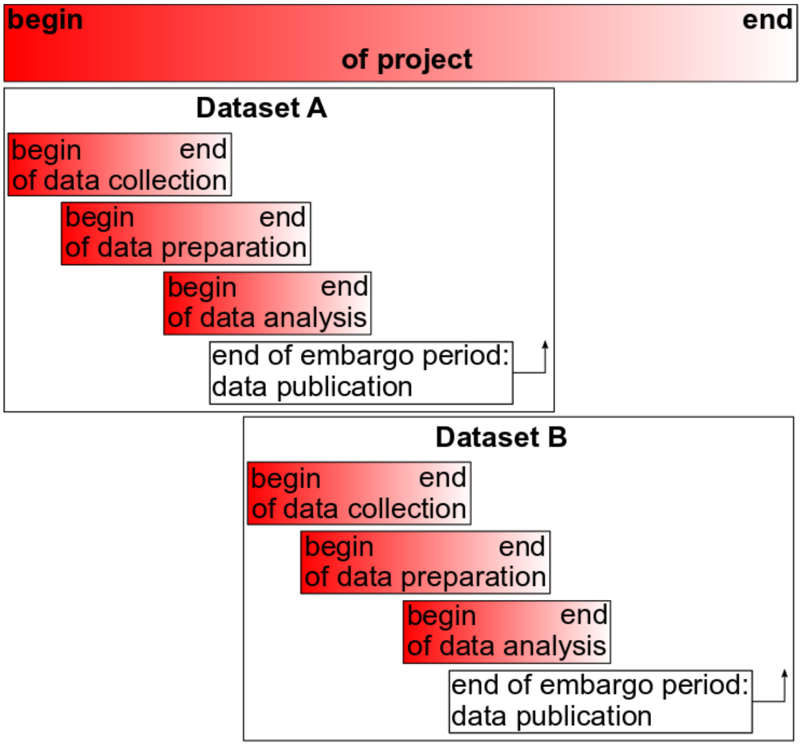

Figure 2: time schedule for research data management

Costs

Costs are seldom subject of a data management plan since institutions offering research funding usually expect a specification of costs in the form part of the proposal and the costs of research data management should be included there. Exceptions are educational science projects funded by BMBF; in this case the costs of research data management have to be included in the data management plan. Regardless of which document contains the statement of costs, it is a good idea to list the costs of research data management since these costs can often be reimbursed.

Just one paragraph with some facts about costs of research data management: labour costs are much higher than material costs. Related to steps in the data lifetime cycle, storage is the cheapest part. Selection and ingestion usually cost much more, see Figure 3. The same is true for the costs of enabling access if access is allowed, e.g. the costs for developing and running a web portal for data download and data processing. Figure 3 is the archive view[5] but, if you store the data on your own, you will likely have a similar cost partitioning.

[5] Project Radieschen (Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur), Kostenverteilung und Risiken, http://doi.org/10.2312/RADIESCHEN_008, in German, retrieved on 02.10.2018

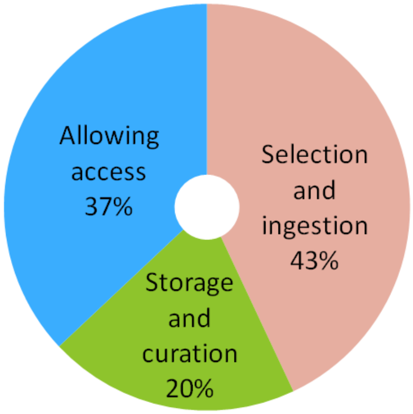

Figure 3: cost partitioning related to steps in the data management cycle

Software tools

Of course, a data management plan can be written with a word processor. Nevertheless, some special software exists for that purpose too and can facilitate the creation and expansion of a data management plan. All these tools have questionnaires, which are a good starting point and can prevent authors of data management plans from forgetting something important.

General open-source tools are DMPonline[6], DMPTool[7] and RDMO (Research Data Management Organiser)[8]. All three are hosted in the Internet, i.e only a Web browser is necessary to use them. DMPonline was developed by the British Digital Curation Center (DCC)[9], DMPTool by the University of California Curation Center (UC3). Both tools are well established and have many users. After registration one can create data management plans, which are stored on a British or American server. DCC and UC3 have decided to cooperate and to develop a common tool, DMPRoadmap[10], which is still under way.

RDMO is also still under way but may already be used. The software has been developed by a DFG project named RDMO too. The software can be tested at the AIP Potsdam[8] but a real data management plan should only be set up with an institution's instance. The developers of RDMO even want to extend their software to a real organiser as this is specified in the name. Announced and already partly realised features are

- metadata extraction, i.e. metadata can be extracted from the data management plan and can be filled-in into the data repository
- Definition of tasks for which, when they are due, a reminder is sent via email

[6] https://dmponline.dcc.ac.uk/, retrieved on 02.10.2018
[7] https://dmptool.org/, retrieved on 02.10.2018
[8] https://rdmo.aip.de/, retrieved on 02.10.2018
[9] http://www.dcc.ac.uk/, retrieved on 02.10.2018
[10] https://github.com/DMPRoadmap, retrieved on 02.10.2018

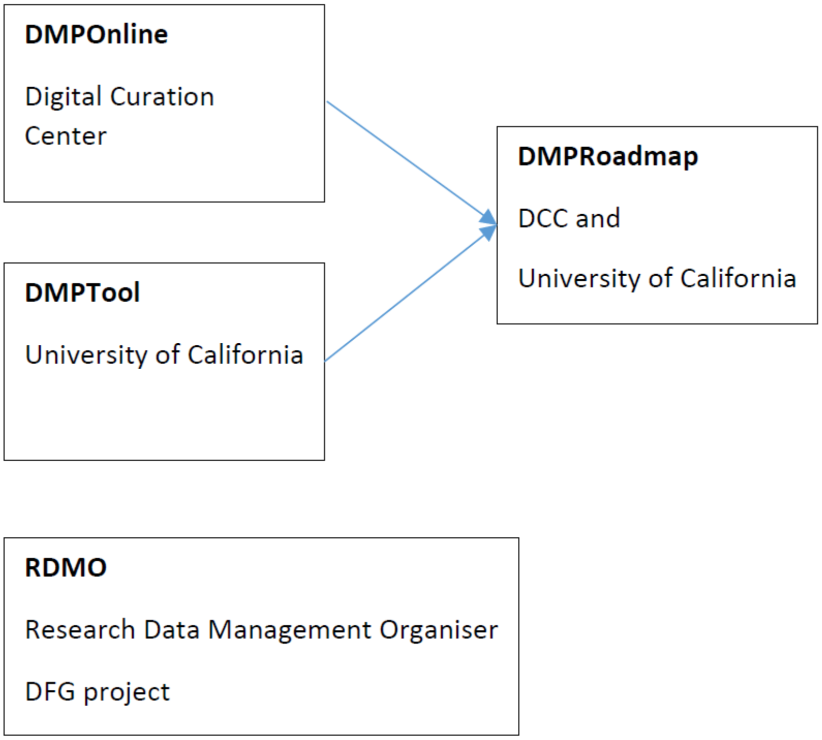

Figure 3: generic tools for the creation of data management plans

Already now, the verbose RDMO questionnaire may be a help.

Additionally, some scientific communities have been working on subject-specific tools. Examples are the DMPTY Wizard[11] of CLARIN-D[12] for the humanities and DataWiz[13] for psychology.

Further reading

Kristin Briney, Data Management for Researchers, Pelargic Publishing, 2015, ISBN 978-1-78427-011-7

Rathmann, Torsten, Requirements of the Research Funding Organisations for Research Data Management, Full text version of the correspondent lesson of the German project FoDaKo, https://doi.org/10.5281/zenodo.1464981

[11] https://www.clarin-d.net/de/aufbereiten/datenmanagementplan-entwickeln, retrieved on 02.10.2018
[12] https://www.clarin-d.net/de/, retrieved on 02.10.2018
[13] https://datawiz.leibniz-psychology.org/DataWiz/, retrieved on 02.10.2018

Examples for data management plans:

https://esrc.ukri.org/files/research/international/esrc-dfid-example-data-management-plan/
https://www.gla.ac.uk/media/media_418168_en.pdf
https://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung