A questionnaire to estimate the needs for research data management

Frank Krüger Sascha Spors
Institute of Communications Engineering
University of Rostock
Contact:frank.krueger@uni-rostock.de

DOI:10.18453/rosdok_id00002290

August 29, 2018

The development of a data management plan requires detailed knowledge about the research data and the corresponding processes. Moreover, detailed knowledge about the entire data life-cycle in the respective research group is necessary when providing support with respect to reproducibility and data management. In order to allow curators to gain this knowledge, we developed a questionnaire, which provides a set of topics including questions to inspire discussions on data management and the data life-cycle. This questionnaire consists of a collection of questions targeting different phases of the data life-cycle. It was developed in order to gain insight about the data management practices in the subproject of the collaborative research centre 1270 ELAINE[1], but can also be used as guideline for so interviews with individual research groups.

1. Introduction

Objective of effective research data management is to enable researchers to reproduce their analyses. Moreover, curated and shared research data allow others to reuse data sets in order to validate the original research results (reproduction) or build new research upon them (data reuse). However, the curation of research data requires domain knowledge and knowledge about the data and the corresponding processes. Parts of this knowledge is typically contained in data management plans (DMPs). DMPs allow (groups of) researchers to document their current processes in research data management from data generation to archiving and publication.

This document introduces a questionnaire that provides a basis for a semi-structured interview and discussion about research data management. The purpose of this questionnaire is twofold: (1) it helps researchers to estimate the current status of their research data management in order to create a DMP and (2) it helps curators to get necessary information about the research data, the corresponding processes and the current understanding of research data management of the interviewee.

The original idea of the interview was to meet with the head of the working group and an experienced employee (e.g., post doc), since the head has typically a more general view on the research topic and the experienced researcher is more involved in the actual practices. During the first interviews it turned out that the head of the working group is necessary in the first parts of the interview (i.e., until Section 2.1) only. The other parts can typically be discussed with the researchers that are more involved in the "daily" research processes. The typical length of an interview based on this guideline is about two hours.

[1] https://www.elaine.uni-rostock.de/

2. The questionnaire

The questionnaire is designed as guideline for informal semi-structured interviews concerning to the data life-cycle. For this purpose, the document starts with a short description of the objective of the document. In brief, the objective is to gain insight into the data life-cycle of individual working groups. The design of this questionnaire is based on a literature review of data interview guidelines [2, 3, 1]. As additional source, the guideline for research data interviews of the Australian National Data Service[2] was used. In fact, parts of the questions we taken from the literature, were adjusted or newly formulated to the fit specific needs of this questionnaire.

Additionally, to provide an insightful discussion some terms that are related to research data management and reproducibility are defined. This allows to easily overcome misunderstandings for specific terms such as "research data", which is often understood as digital data in the form of tables only. Here, research data is defined as anything that is used during the scientific investigations. This includes but is not restricted to: tabular data, images, sound or video files, analysis code, such as python or R scripts, but also specimen, materials, or cultivated cells.

The first part of the interview consists of an open discussion about experiences with respect to scientific data management. Objective of this part is to gain insight of the definition of research data management of the interviewee. Typical things to be discussed during this part are the experiences with respect to the publication of data. One question that often arises during this discussion is whether the interviewee has published data before. Often an introduction to possibilities of data publication (e.g., institutional repositories, domain independent data repositories, domain specific repositories, data journals, ...) follows this question.

The remainder of the questionnaire is structured into seven categories, each of which focuses on different aspects of the data life-cycle. The following sections provide an overview of the objective of the questions of these categories. In the following these categories are introduced and questions exemplified. The entire questionnaire can be found in the appendix in section A.

2.1. Data Generation

Objective of this section is to enable the interviewer to get an understanding of the entire research process. In detail, to get a general overview of the research topic of the interviewee, the methods used for investigation and the data to be produced during research. In addition to questions about the research questions and the way data is generally generated, other questions focus on the source of the data and its documentation.

2.2. Data Collection

This part deals with a detailed identification and description of the data sources to be used and the processes to consolidate the different potentially heterogeneous data formats from these sources. Furthermore, some questions deal with the actual management of the raw data in terms of amount and organisational overview.

2.3. Data Security and Privacy

Data security targets all aspects with respect to data access. It is, for instance, discussed how the access to data is managed within the working group and whether special procedures for data security exist.

2.4. Data Analysis and Visualisation

Data analysis plays an important role during the investigation. Here, questions about the typical procedures for data analysis are provided. Also a list of used applications is requested, since they typically allow to estimate the required effort to reproduce[3] or share the analyses. Often, the data analysis uses different tools or even devices (e.g. workstations, remote servers,...), which requires the data to be transferred and/or

[2] The ANDS guidelines are available at `https://www.ands.org.au/working-with-data/data-management/institutional-dm-frameworks/research-data-interviews`

[3] It is assumed that open source software allows for easier reproduction of the analysis since the software is available for each researcher.

transformed. This process might introduce additional modifications to the data. To this end, this section also discusses the use of validation procedures such as checklists and scripts for automatic validation[4].

2.5. Data Management

The questions with respect to data management focus on actual data handling of the working group after all investigations are completed. Here, backup strategies, preservation durations and the respective responsibilities are discussed. One important question is how the researchers decide when research data can be destroyed[5]. Another aspect of the data management is, whether researchers are able to restore data.

2.6. Data Formats and Metadata

The questions about the data formats and the metadata are important with respect to the effort to be spent for archival with open formats. Beside the a list of applications and data formats of general use in the working group, it is discussed whether open data formats[6] are used or at least available. In addition to the actual data, the discussion focusses on metadata, the format of the metadata and the amount of information that is recorded. Also the source of meta information is of interest (e.g., for automatic meta data extraction and curation).

2.7. Data Publication

The last part of the questionnaire addresses the openness of the researcher with respect to research data. The various possible sources (e.g., colleagues, project partners, public repositories) and the interviewee's experiences with them are discussed. Other questions focus on the publication practices of the interviewee, such as the general willingness to share data and experiences with data sharing.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) within the Collaborative Research Centre 1270 ELAINE. The questionnaire is based on a LaTeX template, which was developed by Sebastian Bader.

References

[1] Jake Carlson. Demystifying the data interview. *Reference Services Review*, 40(1):7–23, 2012.

[2] Steve Van Tuyl and Gabrielle Michalek. 2014 Faculty Data Survey Instrument. *Journal of Scholarly Communication and Librarianship*, 2014.

[3] Michael Witt and Jake R. Carlson. Conducting a data interview. Technical Report 81, Purdue University, 2007.

A. Questionnaire

In the following the actual questionnaire is included.

[4] Automatic validation scripts have advantages over manual validation since the validation itself is comprehensible and re-runnable whenever new data arrives. This is in analogy to automated tests of software.

[5] Many investigations produce massive amounts of data, e.g., numerical simulation, which will further be analysed in aggregated or filtered form.

[6] Open data formats allow the processing of the data independent from the original software that was used during the original investigation.

Data Interview Guideline for CRC 1270 ELAINE

CRC 1270 ELAINE

Frank Krüger, Sascha Spors

2018-08-29, v1.0

Version History:

v0.1 initial document
v0.2 add definitions; revise format; minor changes to questions
v0.3 data, older than 10 years
v0.4 unintentional loss of research data
v1.0 persistent URL, spelling, publication

1 Introduction

1.1 Objective

Objective of this document is to provide a semi-structured guideline for data interviews within the CRC ELAINE. The data interview is conducted in order to gain insight into the data life cycle of the different projects. This includes, but is not limited to: (1) current data management strategy, (2) storage requirements, and (3) information about metadata.

1.2 Source

This document is based on variety of documents from the literature [4, 5]. Some of them are used as informal input, others were used as direct source for questions.

1.3 Definitions

The following section provides a list of definitions that will be used within this document. For each definition, a reference is provided.

Open Science Within this document, open science is referred to as a collection of principles to make scientific research and its results accessible. According to [2, 3], among others, open science includes the following aspects: (1) open access, (2) open data, (3) open source, (4) open methodology and (5) open notebook science.

Research data This document refers to research data as "*the recorded factual material commonly accepted in the scientific community as necessary to validate research findings*" [1]. This does not necessarily include personal information of participants of research study.

Repeatability Repeatability describes that an experiment or a data analysis is repeatable by the original researcher by use of the original data sources and analysis tools. Challenges here include long term preservation of data, code, as well as applications

Reproducibility In contrast to repeatability, reproducibility here denotes the repetition of an experiment or a data analysis by others than the original researcher. As for repeatability, the original data is available. The challenge here is, to provide a comprehensive documentation of the entire process.

Replicability Replicability describes the repetition of an entire study based on the description from the literature without access to the original data or analysis tools. Similar as for reproducibility, an extensive documentation of the original study is necessary for replication.

2 Experience with Data Management

- Describe your experience in scientific research data management

3 Data Generation

1. What are typical research questions which are aimed at based on research data?
2. Describe how research data is typically generated in your research group
3. What kind of research data is generated?
4. How much research data do you typically generate (e. g. volume per study, size per month)?
5. Do you use public available data (e. g. published by other researchers)?
6. What other sources of data do you use?
7. How do you document the data generation (e. g. selection of participants, device configuration, . . .)?

4 Data Collection

1. What are the different sources you collect your data from (e. g. questionnaires and clinical devices)?
2. Describe how the data from the different sources are merged
3. Do you keep a list or registry of datasets?
4. How much research data do you have now?

5 Data Security and Privacy

1. How do you ensure data security?
2. Do you manage data access control?
3. Do any privacy issues arise during your studies (e. g. pseudonymization)?

6 Data Analysis and Visualisation

1. Describe how the collected data is used to gain insight in the data generating process (i. e. preprocessing, cleaning, transformation, statistical analysis, . . .)?
2. Which applications/environments are used for data analysis (e. g. R, SPSS, . . .)?
3. Do you use visualisation for data analysis?
4. How do you ensure validity of the research data (e. g. checklists, automatic validation scripts, . . .)

7 Data Management

1. Where do you store your research data (e. g. mobile disk, researcher's notebook, central storage server, . . .)?
2. Do you use a backup for your data?
3. Who is responsible for backing up and maintaining your research data (e. g. administrator, technical staff, . . .)?
4. How do manage different versions of the research data?

5. How long do you typically preserve the research data?
6. How do you decide whether to delete research data (e. g. data not used for analysis, dataset to large, . . .)?
7. Do you typically create any data management plans (e. g. in projects funded by DFG)?
8. Did you ever restore and use data that was archived more than ten years ago?
9. Did you ever lose any research data unintentionally?

8 Data Formats and Metadata

1. Which applications are employed during the data life cycle (e. g. simulation software, EEG software . . .)?
2. Which data formats are primarily used within you research group?
3. Are there any open formats (e. g. comma separated values (CSV), extensible markup language (XML), . . .)?
4. In how far is the data documented?
5. What metadata do you capture?
6. What are the sources of metadata that you use (e. g. documentation of clinical device)?
7. Do you use any metadata standards (e. g. Dublin Core[1], DDI[2])

9 Data Publication

1. Do you share your data with other researchers that were not involved in the original data generation?
2. Do you make your data publicly available?
3. Do you use any domain specific data repositories (e. g. PANGEA, PhysioNet, . . .)?
4. Do you publish any kind data description (e. g. data articles)?
5. Which licenses do you use for data publication (e. g. CC BY)?
6. Do you use persistent identifiers (e. g. digital object identifier (DOI), persistent URLs, . . .) to refer to the published data?

References

[1] OMB CIRCULAR A-110. Grants and agreements with institutions of higher education, hospitals, and other non-profit organizations, 1999. https://www.whitehouse.gov/omb/circulars_a110.

[2] Stefan Kasberger and Christopher Kittel. OpenScienceASAP: Was ist Open Science? last accessed:16.10.2017.

[3] Nancy Pontika, Petr Knoth, Matteo Cancellieri, and Samuel Pearce. Fostering open science to research using a taxonomy and an eLearning portal. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW 15*. ACM Press, 2015.

[4] Steve Van Tuyl and Gabrielle Michalek. 2014 Faculty Data Survey Instrument. *Journal of Scholarly Communication and Librarianship*, 2014.

[5] Michael Witt and Jake R. Carlson. Conducting a data interview. Technical Report 81, Purdue University, 2007.

[1] http://www.dublincore.org/
[2] http://www.ddialliance.org/