Research-Data Management Plan

"Methods of non-linear algebra applied to a didactics' dataset"

Tabea Krause et al., 2022

This is a prototype of a research-data management plan (RDMP) for a collaborative math project. Tabea K. (U Leipzig) is the primary researcher, her collaborators are Julia L., Marie B., and Tobias B. (all MPI MiS Leipzig). Together they apply methods from non-linear algebra to a dataset that originated in didactics. Their mathematical approach is outlined and justified in the greater grant proposal of which this is a supporting document. This document is updated regularly to serve the investigators as guideline for their RDM.
In very broad terms, Tabea and collaborators aim to understand the relationships between the datapoints and a set of measurement variables. To do so, they use principal component analysis and graphical models from statistics as well as convex hull computations.
The researchers structure their project and this RDMP along the data life cycle into three broad stages: existing preliminary research, newly proposed research, and future/subsequent research. Together they decide on the strategy outlined below, supported by the research-data management adviser Christiane G. (U Leipzig).

Data description

The proposed research builds on existing research data in the form of a data table in .csv format. Both the metadata and the underlying raw data to that table are available as .txt and .pdf files. The preliminary research including the accompanying paper was published under a CC-BY-SA license and is accessible via the repository Zenodo (at DOI).

When analysing the existing data, we generate new research data of the following type: mathematical formulae, abstract concepts like systems of measurement variables, data collected in tables as input, code, pictures and plots of graphical models and polytopes, lists of conditional independence statements and polytope data, matrices, and vectors as output. These are saved as .pdf, .tex, .csv, .ipynb, .jl, .txt, .R, .png, .svg. The open source software R will be used. The expected total data volume will not exceed 2GB.

Data documentation

During run time of the project the generation and processing of digital objects are documented in a continually edited and updated .md file. This living document is accessible to all researchers via Tabea's group server hosted by U Leipzig. The document contains information on the files' location and technical properties and a short description of data handling and analysis. Currently there is no research-group specific metadata scheme or data documentation standard available for mathematics. We adhere to the DataCite metadata standards required by Zenodo. Christiane is in contact with the MaRDI Help Desk in order to quickly adapt to new mathematic-specific recommendations.

Data storage

Regarding version control and short-term storage during the run time of the project we use Overleaf for LaTex editing and GitHub for any computational code. Implementing the 3-2-1-rule, we always save three copies of any document in two different locations and at least one of these not locally. Every researcher keeps a current version of the files they are working on locally on their PC. Additionally, a copy of these files is placed in

the respective cloud-service (Overleaf, GitHub, U Leipzig nextcloud) also used to share files with each other. Weekly backups are made on Tabea's group server hosted by the U Leipzig Computer Center according to their storage policies.

On Tabea's group server, we implement a folder structure roughly seven broad and two to three deep. The first level contains one folder for each of the four non-linear algebra methods, one for education maths/previous work, one reports and papers, and one for general files. Deeper levels refine the first.

Regarding long-term storage, we are not aware of any archiving guidelines by our institutions or the funding agency which exeed the "good scientific practice". For the required ten-year storage we use U Leipzig's Computer Center infrastructure. Christiane is responsible for this and for the organisation of data migration if necessary. Between the researchers we archive all research data in their final version.

Access to and publishing of data

During run-time every involved researcher, the supervising parties, and RDM coordinator Christiane have access to the data. After completion we want the scientific community to have easy access to all of our final results.

In order to achieve this, we publish our research data under the creative-commons license CC-BY-SA on MathRepo and Zenodo. MathRepo is Leipzig MPI's own repository and the only math-specific repository we are aware of. Zenodo is built and operated by CERN and OpenAIRE, making it sufficiently secure and recognised in the community. It assigns a DOI to our research data and it can seamlessly include our GitHub repository. Tabea's responsibility as the lead researcher is to submit our results to an appropriate journal using, if required, that journal's infrastructure for supplementary material. She additionally puts a preprint on arXiv. The collaborators upload our project to Zenodo, including both our GitHub material and the paper publication, and look into the possibility of creating teaching material and extra explanations to be uploaded to a MathRepo page for easy reuse of our results. Christiane acts as an external adviser in this process.

Ethics and legal aspects

Usage rights have been agreed on in writing individually by the three researchers and the institutional parties involved. Arising copyright claims are checked regularly and recognised if justified. Data privacy rights do not pose a limitation to our research. At this point we are not aware of any contract law provisions regarding data publishing.

Responsibilities in RDM

The person mainly responsible for implementation, adherence, and updates to the RDMP and RDM in all stages of the project is Christiane. She also monitors the institutes' and funding parties' guidelines. The researchers cooperate in their respective capacities.

Incurring costs for RDM

We do not predict extra costs for personnel implementing RDM or hard- and software during run-time of the project. That and the consultation from internal RDM experts is encompassed in the University's budget. The same goes with the costs for long-term research-data storage given it is on the Universities servers. Zenodo and MathRepo do not generate extra costs for the researchers. If contrary to expectations any additional costs arise we will apply for funds with the project's funding agency.