Workshop: Research Data Management in a Nutshell

University of Cologne, Germany
January 18 2018

Constanze Curdt
University Computing Center, University of Cologne

Jens Dierkes
University and City Library, University of Cologne

Sonja Kloppenburg
Research Management, University of Cologne

Preface

This text accompanies the presentation "RDM_AMGC_Promovierendentag_20180118_presentation_zenodo.pdf" held at the doctoral day at the University of Cologne in January 2018.
The answers are not exhaustive, rather it is intended to give you some ideas and examples that you can investigate further.
The storyline, questions, and answers are based on a similar workshop held by the Göttingen eResearch Alliance[1] in 2016 at the ELPUB Junior Scientist Day (Dierkes, Rücknagel, Gnadt 2016)[2].

[1] http://www.eresearch.uni-goettingen.de/
[2] https://www.fosteropenscience.eu/node/1662

STORYLINE WITH QUESTIONS AND ANSWERS

SETTING: A large research project funded for 3 years wants to explore artefacts from ancient cities. Two groups from two partner institutions team up to collaborate on one research topic and agree to share all captured raw and processed data. The teams organize an expedition. Both teams visit excavation places in different locations. Predominately both teams collect photographs and measurements (length, color, scales, coordinates, etc.) of objects, as well as text notes.
The expedition starts off well.

1) **DATA LOSS:** Both teams successfully collect their data. Team A returns home without critical incidents. Team B on the contrary, gets into a sandstorm. Two laptops containing data are destroyed. The destruction is discovered when they arrive at home, thus the lost data cannot be recollected in the field again. **What should have been done to prevent this incident?**

Answer:

- Have a backup strategy in place, e.g. store data and materials on several storage media, backup regularly; keep track where and how current versions are stored (versioning and naming conventions help to have an overview of existing and most recent versions.)
- Within projects roles and responsibilities must be clarified before data are collected.
- If possible institutional storage systems should be used in high frequencies. E.g. mirror central database at different locations for very important materials (dependent on frequency of change)
- Use cloud storage (e.g. Dropbox), but:
  - internet access and sufficient bandwidth are required
  - Reliability and confidentiality need to be verified

2) **DATA QUALITY:** When comparing their data, the two teams notice that photographs from Team A have lower levels of contrast and resolution than from Team B. Also the teams used different coordinate reference systems. Even with processing software, some details that Team B is interested in cannot be identified and some date were not comparable at all. **How could this have been optimized before?**

Answer:

- Discuss the research purposes respectively requirements the captured data have to fulfil BEFORE collecting the data
- Agree on compatible data acquisition methods: resolution and formats for photographs, coordinate system
- Agree on standards for documentation (e.g. describe which coordinate system has been used)
- Do quality checks on site (as early as possible): Granting early access to the captured data to all project members, so that inadequacy of pictures is discovered

3) **DATA RE-USE/ AVAILABILITY:** After several months of research and data processing, a Team B researcher wants to analyze raw data from a Team A member. However, this team member has already left the team and his colleagues do not know where to find his raw data, since they only have copies of processed data. **How could this situation have been prevented?**

Answer:

- Agree on policy to describe what happens to data, especially raw/primary data, when members leave the project
- Directory to list data and their storage location or a central database with submission agreements in place (where, when, how to submit data sets in what form?)
- Agreement on naming convention that allows to identify data sets.
- Minimum set of metadata description should be added to describe the data sets (e.g., stored within the data file or a text file in the same folder).

4) **DATA DOCUMENTATION:** The data from the Team A member was sucessfully downloaded from the shared database by the Team B member. However, substantial information on how the data were recorded, like the instruments used are missing from the database. **What measures should have been taken to better describe the data?**

Answer:

- Metadata schema to describe disciplinary data should have been used.
- Besides adding metadata, a data documentation (lab book) can be created to describe processes of data collection and data processing.
- Enable access to documentation for all (collaborating) project members

5) **SUSTAINABILITY:** After the project ends, so does the funding for the infrastructure used by the project. While the coordination has taken care of the project website being available for another 2 years, the central database which the two teams used is shut down 1 month after the project end, since no money is available for its maintenance anymore. However, several PhD students have not yet finished their thesis work and require access to the data from the database. They are now left in the final phase of their thesis, and cannot complete or verify some of their research results due to missing data availability. **What should have been taken care of to avoid this situation?**

Answer:

- Agree with service provider on database maintenance for a certain period after the project ends (Good scientific practice demands storing data in accessible way for 10 years after project end)
- Create local backups of relevant data portions and define access and reuse conditions