Report on research data management interviews conducted for HMC Hub Energy in 2022

Felix Ballani, Theresa Schaller, Leon Steinmeier

Helmholtz Institute Freiberg for Resource Technology
Helmholtz-Zentrum Dresden-Rossendorf

Mohamed Anis Koubaa, Jan Schweikert, Karl-Uwe Stucky, Wolfgang Süß

Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology

The Energy Hub of the Helmholtz Metadata Collaboration (HMC) conducted interviews with various stakeholders from the Helmholtz Research Field Energy on the topic of research data management (RDM) in 2022. The intentions were to build and serve a metadata community in the energy research field and to extend the Helmholtz-wide survey conducted by HMC in 2021 Arndt et al., 2022). Besides the deeper insight into the current state of RDM and metadata handling at the Helmholtz sites relevant to the Energy Hub the interviews focused on the related needs and difficulties of researchers and their satisfaction with the current state. Furthermore, we tried to discover already existing workflows and software solutions, to establish contacts and to make HMC better known.

In the following, we will report on the preparation, implementation and findings of these interviews. We hope to thereby make the effort and possible benefits of such interviews more apparent.

1. Preparation

The interviews were planned in a semi-structured format (see, e.g., Gill et al., 2008) with large narrative portions. On the one hand, we wanted to facilitate a guided conversation between the interviewee and us using an interview protocol, but on the other hand, we also wanted to force a strong conversational character in which the interviewee would be given the opportunity for detailed and in-depth answers and, if necessary, information that digressed from the protocol.

1.1. Questionnaire

A catalog of questions was developed in advance to provide a protocol which could guide the intended semi-structured format. However, the selection and arrangement of the questions was not based on any questionnaire design conceived for the purpose of scientifically testing specific hypotheses.

After conducting the first interviews, this catalog was again slightly revised in order to improve the conversational fluency and the weighting of a few individual aspects.

In terms of content, the questionnaire is divided into several sets of topics. At the beginning there are some general questions, among others about the professional background and the professional experience of the interviewee. This is followed by questions about data management in a specific project in which the interviewee was involved. Next is a series of more general questions about data management practices and external circumstances. This

is followed by questions about the interviewee's personal view and evaluation of the actual practice in their own data management or that of their work group. The interview ends with further general questions about data management, including a question on what an ideal RDM situation would look like, as well as a request for brief feedback on the content and form of the interview.

1.2. Privacy policy

A data protection declaration has been prepared in cooperation with the data protection officer of Helmholtz-Zentrum Dresden-Rossendorf (HZDR). The essential points regarding confidentiality of conducting and recording the conversation are as follows:

Contents of the conversation will be recorded in writing only. It will be ensured technically and organizationally that only the responsible employees of the HZDR who are entrusted with the project have access to the personal interview results. The project partners from the Karlsruhe Institute of Technology (KIT) (i.e. members of Hub Energy) will only receive anonymized data records for participating evaluation purposes. The interview results will be aggregated during the evaluation to no longer allow any conclusions to be drawn about any individual person interviewed. If the interviews cannot be conducted in person, the BigBlueButton (BBB) video conferencing service provided via the HZDR servers will be used.

1.3. Test interviews

To get a better feel for the interview process and to identify any major weaknesses in advance, we first conducted two test interviews within our institute and gathered feedback on them.

2. Implementation

2.1. Conversation initiation

To initiate the conversation, we contacted our potential interviewees via the prepared first e-mail (clearly express the request on the one hand, but not to sound too demanding on the other). We mainly requested individuals who had agreed to be contacted for an additional interview as part of the 2021 HMC survey, and in addition a few more individuals also from the Helmholtz Research Field Energy. We received no response at all to about 25% of the inquiries. On the other hand, almost all of the responses which we received were positive; in the case of one inquiry, we were referred to another person with whom we were ultimately able to conduct a conversation.

Finally we asked the interviewee to sign a declaration of consent, which then enabled us to store and further process the results of the interview in line with the privacy policy outlined above.

2.2. Interview conduct and transcript

On the interviewer side, we always conducted the interviews in pairs. One of us mainly took notes and occasionally asked supplementary questions, while the other took charge of the interview, but also made their own notes.

At the beginning of the interview, we first introduced ourselves, the Helmholtz Metadata Collaboration and the purpose of the interview, and - in order to avoid any misunderstandings - we explained once again how the data collected during the interview would be handled and which people would have access to what.

In the main part of the interview, we generally followed the content of the prepared list of questions. However, it was important to us that the interview had a conversational rather than an interrogation character. We always succeeded in achieving a good flow of conversation.

For example, we went back over individual answers, used questions more flexibly where necessary or even omitted them, and made an effort to extract answers to some questions from a longer flow of speech in a meaningful way. In retrospect, this was certainly not always entirely satisfactory in terms of answering all questions completely. On the other hand, many interviewees positively emphasized in the brief feedback at the end of the interview that they found this form very pleasant.

The interviews were timed to last one hour, but varied from an ample half hour to nearly two hours, depending on the time and communicativeness of the interviewees.

2.3. Debriefing

After each interview, we interviewers took time for a debriefing, which could last up to an hour. In this debriefing, we first completed and refined the notes into the final version of the interview transcript. Finally, we summarized the aspects and findings from the interview and the points being essential for us.

2.4. Anonymization

The interview transcripts were finally anonymized, in accordance with our privacy policy (see section 1.2.). The interviewees were sent this anonymized version with the request to check whether it was anonymous enough for them and, if necessary, to make suggestions for changes. Any such changes were subsequently incorporated into the anonymized version of the interview transcript.

3. Findings

We would like to point out at the outset that due to the way in which the interviewees were selected, the results as a whole cannot and should not be regarded as representative. In particular, there is a bias with regard to familiarity with or openness to RDM topics.

The findings presented here are based on a total of 27 interviews. The majority of the interviewees (21) were in a senior scientific position as project leader, group leader, deputy head of department, or head of a research facility. Of the remaining six, there was one in radiation safety quality management, two PostDocs, one project staff member, one PhD student, and one technician.

We have subsumed the most frequently addressed aspects below under three larger sets of topics. Multiple responses are possible in each case. In addition, please note: If we, for example, say that X people considered Y important, this does not automatically mean that all other participants considered Y unimportant since we might not have discussed topic Y in all interviews.

An important complex of topics can best be summarized under the term *digitization*.

- Digitization in general was explicitly considered important by two persons.
- Digitization in the form of electronic laboratory notebooks (ELN), laboratory information and management systems (LIMS) or databases was an essential aid to data management for four people.
- In contrast, six individuals found the use of an ELN/LIMS to be desirable, but had not been able to use one or were just beginning to use one.
- Likewise, six people felt that (better) IT connectivity of laboratory equipment was needed.
- Finally, automated generation of metadata was desired in connection with the use of laboratory devices by four people.

Another complex includes topics on the mediation or the general framework of RDM and concrete support or lack of available solutions therein:

- Dealing both with very large data volumes (6 people) and with a large variety of data formats (6) was perceived as problematic or at least challenging. Two interviewees complained about proprietary output formats in laboratory instruments.
- A lack of (institutional) specifications for RDM was addressed by 11 people, and among these, 7 even saw it as problematic.
- Concrete personnel support for RDM is desired by 16 persons.
- For eight people, preliminary or further training in RDM is important.
- For four people, good examples are lacking or it is difficult to communicate the usefulness of good RDM.
- Seven people commented that an important issue is the sometimes high staff turnover and the associated onboarding and offboarding processes, as a result of which e.g. data loss can also occur. Relatedly, two interviewees commented that data management is not really a topic of interest for employees on short-term contracts.
- Three people had either purchased or programmed their own solutions to support their RDM due to a lack of IT support or non-existent software solutions.
- After all, more than half of the interviewees (14) explicitly stated that they simply lack the time for better RDM. One reason cited in this context was that, up to now, the performance of researchers has essentially been measured in terms of the number of publications and there is a low value placed on good data management over publication performance (2).
- Three interviewees would like to see easier-to-use or more cohesive data management tools.
- Three interviewees also spoke out in favor of a centralized/shared collection of resources within their institutions.
- After all, nine interviewees are familiar with version control such as Git and also employ it in their work.

A further complex of topics arises around the provision of one's own data and the use of third-party data.

- Nine of the interviewees used repositories to find data. Although for some of them there are still too few data repositories or they are not yet sufficiently well filled, their necessity is shown by the simple fact that three of the interviewees have read out required data from graphics in publications, partly with the help of special software, because they could not obtain the data in any other way.
- On the other hand, eight of the interviewees have already published data or program code in repositories. Two other persons were aware of data repositories, but had not yet used them in any way, but at least knew people who had already uploaded something there.
- Two of the interviewees stated that their practice to date has been to make their own data available on request.
- Three of the interviewees discussed the fact that it may not necessarily make sense to upload all data to a repository, for example because it is raw simulation data, the data is very extensive or parts of it are sensitive.

The FAIR principles (Wilkinson et al., 2016) were known by about 70% of the interviewees, and as expected higher among those on the list from the HMC survey (about 81%) than among the other interviewees (about 33%). Insofar as awareness of the FAIR principles can be seen as an indicator of conscious RDM, there is thus a corresponding general need for further training, at least in the Research Field Energy.

The majority of interviewees would like to further their education in the field of RDM and FAIR data, many of them more on specific rather than general topics.

We received almost exclusively positive feedback for the form of the interview and the selection of topics addressed. Due to the nature of the topics addressed, some people even perceived the interview as one in the sense of further training or counseling.

4. Conclusion

With this report, we provide an insight into the preparation, implementation and results of the qualitative interviews on RDM conducted for the Energy Hub in 2022. Overall, this was a very valuable experience for us, as it provided us with a perhaps not comprehensive but at least wide-ranging view of various aspects of individual and structurally anchored RDM in some areas of the energy research field. As it stands today, the range in the perception and implementation of good RDM in the energy research field is wide, ranging from ignorance of the topic to very committed implementation of specific solutions.

For HMC in the energy research field, this means that on the one hand, in terms of a cultural change towards good RDM and FAIR data, more basic training should continue to be offered, but on the other hand, specific offers and training are also of high interest. In any case, this cultural change and related ideas such as a Helmholtz-wide data space must be motivated by convincing examples. However, the deficits in digitization show that sometimes the basic prerequisites on which something like a Helmholtz-wide data space is actually based are not sufficiently met.

Acknowledgments

We would like to thank all those individuals from the Helmholtz Association who made themselves available for a qualitative interview and gave us insight into their practices, challenges and achievements around data management and metadata handling.

We are grateful to all those colleagues from the Helmholtz Institute Freiberg for Resource Technology and from the Helmholtz Metadata Collaboration who supported us in any way in preparing and conducting the interviews.

References

Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., Süß, W., and Helmholtz Metadata Collaboration Working Group “Taskforce Survey“ (2022). A survey on research data management practices among researchers in the Helmholtz Association. Edited by Lorenz, S., Finke, A., Langenbach, C., Maier-Hein, K., Sandfeld, S., and Stotzka, R. HMC Report 1. HMC Office, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany, 39 pages. https://doi.org/10.3289/HMC_publ_05

Gill, P., Stewart, K., Treasure, E., and Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *British Dental Journal* **204,** 291–295. https://doi.org/10.1038/bdj.2008.192

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018. https://doi.org/10.1038/sdata.2016.18

Appendix A: Questionnaire

A.1. Introduction

Q1.A „Housekeeping" metadata:

Q1.A.1 Name (and title)
Q1.A.2 Affiliation
Q1.A.3 Research field
Q1.A.4 Position

Q1.B Scientific / Work experience & Current work:

Q1.B.1 How long have you been doing what you are doing?
Q1.B.2 What have you been doing in the past?
Q1.B.3 What exactly do you do in your current job?

A.2. Practical aspects of data management

Q2.A Specific project

Q2.A.1 Please introduce this project.
Q2.A.2 What was your role in this project?
Q2.A.3 How much do you know about everybody else's work (especially the data management)?
Q2.A.4 Was there a data management plan in the beginning?

Q2.B Data management in general

Q2.B.1 Please talk about how you handle your research data in general. (researcher)
Q2.B.1 How do you support the data management of the scientists in general? (data steward)
Q2.B.2 Which parts of your data management do you document with text or more structured metadata?
Q2.B.3 General data management requirements of institute?
Q2.B.4 What kind of plan on how to document and safe your work (data, processing steps, scripts, figures, publications, etc.) do you / the scientists you work with have when starting a new project?

Q2.B.5 What about data management with regards to ...

- Presentations at conferences/workshops?
- Publications - Figures, Scripts, Tools?
- Jointly creating publications?

Q2.B.6 Do you use data repositories for finding/getting or publishing data? Which?
Q2.B.7 What kind of long term data archiving do you apply?
Q2.B.8 Do you know or even consult a data steward / research data management expert?

A.3. Evaluation of data management

Q3.1 What do you think about your data management / the data management of the scientists?

Q3.2 Do others / your "future self" understand the documentation? (folder structure / file naming convention / documentation of workflow and strategy)

Q3.3 Evaluate the resources / general infrastructure / software used at the moment.

Q3.4 What would you need in that regard to make your data management (even) better?

Q3.5 Do you get enough support / input from your supervisor / co-workers / (infrastructure) / people working for you (like PhD students)?

Q3.6 How do scientists respond to your work? (only for people working in infrastructure)

A.4. General questions

Q4.1 FAIR: Have you heard of FAIR?

Q4.2 Interest: Are you interested in further reading/courses on this subject?

Q4.3 Obstacles: What is the main reason that keeps you from a satisfactory data management? (If data management works great, specify why and how.)

Q4.4 Utopia: If you had unlimited resources (employees, money) what would you decide to change concerning your data/metadata workflows?

A.5. Follow-up

Q5.1 Feedback

Q5.2 Do you want to be informed by us about future software developments and/or courses?