Road to a Chemistry-Specific Data Management Plan

DANIELA ADELE HAUSEN (ID)

ANN-CHRISTIN ANDRES (ID)

JOCHEN ORTMEYER (ID)

SONJA HERRES-PAWLIS (ID)

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

In order to develop a discipline-specific data management plan (DMP) template, it is important to obtain information from researchers. For a chemistry-specific template, NFDI4Chem conducted a series of interviews with 27 participants and used data from the RDA WG Discipline-specific Guidance for DMP online survey.

The interviews showed that the implementation of research data management in everyday work is a big challenge. Key findings from the interview series highlight challenges in implementing FAIR principles, with a focus on "Findability" and "Reusability." The importance of linking physical samples and data in chemistry is emphasised, with discussions on storage, archiving, and the use of tools like electronic lab notebooks and repositories. However, documentation methods, software tools, and naming conventions commonly used in chemical research are also addressed. Overall, the study underscores the need for improved resources and strategies to enhance data management practices in the field of chemistry.

All the gathered information and examples will be used to develop a DMP template in line with chemistry-specific requirements. The results provide a comprehensive outlook on the future developments of research data management (RDM) in chemistry.

CORRESPONDING AUTHOR:
Daniela Adele Hausen

RWTH Aachen University, Germany

hausen@ub.rwth-aachen.de

INTRODUCTION

In the past, it has been shown that data management plans (DMP) were an important building block for research data management (RDM). In recent years, DMPs have been used much more as generic tools by research funders, who are increasingly moving away from the original idea of a project-specific practical plan (Smale 2018).

This became clear in 2021, when the German Research Foundation (DFG) published a checklist for the appropriate handling of research data (DFG 2021). Similar to a DMP, this checklist is intended to support researchers in Germany in thinking about RDM as early as the project application stage. In contrast to a classic DMP, the DFG connects the individual questions and corresponding answers directly to the specific topics in the proposal text. This emphasises the proximity of data management to research. Similarly, to other research funders, this checklist must also be submitted at the end of the project (DFG, recommendations). Due to that, it is useful to maintain the topics over the project's funding period. Easy maintenance of the checklist is possible, for example, with the Research Data Management Organiser (RDMO) tool (RDMO Community). The DFG website also contains subject-specific statements on the handling of research data in the regarded disciplines (Deutsche Forschungsgemeinschaft 2023). As a more specific example, there is one from the Chemistry Review Board, which briefly touches on the most important areas such as physical samples, analytical measurement methods, use of lab notebooks, etc. (Fachkollegien Chemie 2021).

Chemistry has a strong affinity for documentation. Laboratory notebooks have existed for a long time, and in recent years Electronic Laboratory Notebooks (ELN) have become another important tool for documenting research and, in particular, research data. Ideally, an ELN covers the entire life cycle of the data and enables structured documentation of the individual steps (NFDI4Chem, Knowledge Base). Similarly, a DMP ideally covers the entire life cycle, but this is more at the project management level. A DMP is not used for the research data itself, but to describe how the research data is handled. It contains a detailed description of the data that go into the research project and the data that are generated in the research process. A strategy for handling the data as well as rights and role management are usually included (OpenAire). These are two different types of documentation that complement each other.

Nevertheless, the checklist on most DMP templates is not self-explanatory, and the meaning of some questions is not clear to researchers, or researchers are confused about which answer is the correct one. However, since this checklist is important for researchers in Germany and especially within the National Research Data Infrastructure (NFDI), the consortium NFDI4Chem is also dealing with the topic of DMP for chemists (NFDI4Chem Consortium).

The NFDI4Chem aims to support researchers by providing direct guidance on and in the checklist. For this purpose, the consortium has supported the working group of the Research Data Alliance (RDA WG) on *Discipline-specific Guidance for DMP* (RDA). This working group was established in 2019 and conducted a survey at international level. The objective was to reach as many disciplines as possible and to cover the various domains of the Science Europe guidance.

METHODOLOGY AND MATERIALS

To take forward the issue of a chemistry-specific template, a working group of six professors and research associates was formed within NFDI4Chem. This group first analysed the RDA WG *Discipline-specific Guidance for DMP* online survey (Wham et al. 2022) (further described as RDA survey) dataset which, after cleaning, yielded 20 results for chemistry. The working group decided to take this dataset as a base and enrich it. Additionally, to fill information gaps in the RDA survey dataset by means of a pre-questionnaire (further described as pre-questionnaire) and interview series (further described as interviews). This data collection was carried out in a structured two-step process. First, the 18-question pre-questionnaire was distributed to potential interviewees with a request to complete it before the interview. The pre-questionnaire was almost identical to the RDA survey. The only difference is the way of handling – the RDA survey was carried out with the tool SoSciSurvey and the pre-questionnaire was a text file.

This was followed by an in-depth interview of 30–40 minutes covering the topics policy, FAIR principles, data, workflows/samples, software, publication, collaboration, and data

retention. The interview was minimally adapted based on the interviewee's responses of the pre-questionnaire, with specific follow-up questions tailored to each participant's situation, such as asking for examples if the FAIR principles were being implemented, or asking about challenges and barriers if they were not.

For the pre-questionnaire and the interviews, the target group was defined as academic chemists. As far as possible, the different areas of chemistry should be covered. They did not necessarily have to be associated with NFDI4Chem, but should have relevant experience in research data management (RDM), see Figure 1. In this way, 27 interviews and 27 pre-questionnaires were conducted and important information was obtained. This results in a data set of 47 for the RDA survey and pre-questionnaire (further described as survey) and 27 additional interviews.
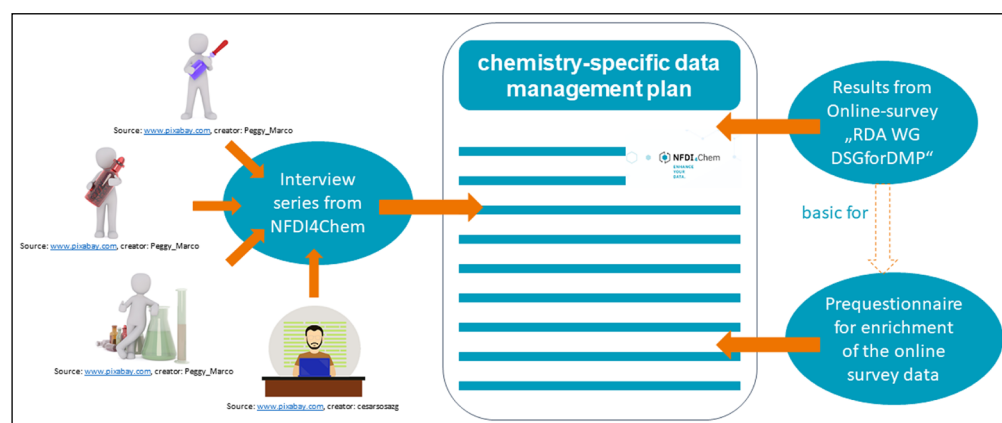

Figure 1 Data generation and reuse that will be included in a chemistry-specific DMP template.

At the current time, in the pre-questionnaire and interviews participated PhD candidates, postdocs, junior professors/ group leaders, and professors. The participants came from the various sub-disciplines of chemistry. The DFG classification was used for the online survey. This results in the distribution shown in Figure 2. This is a very rough distribution for the chemical subdisciplines. For the pre-questionnaire, a finer breakdown could be made.
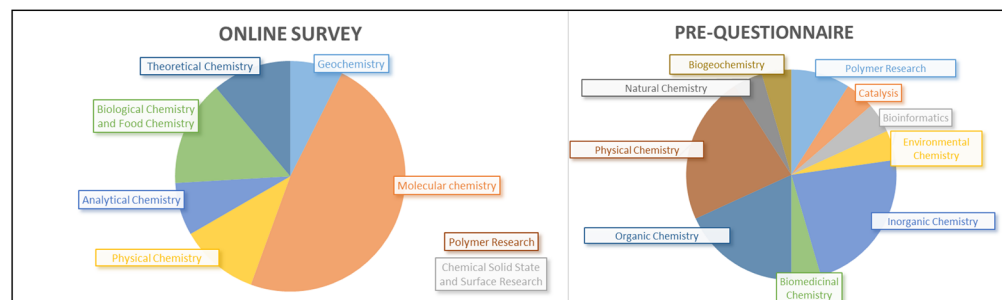

Figure 2 Chemical subdisciplines in the online survey and pre-questionnaire.

The survey, including data processing, data analysis, and data visualisation, was analysed by using Excel (2016). In the analysis of the survey data, responses are analysed by question to account for differences in response rates. The reference point is the total number of respondents. The data can be retrieved from the following link: https://doi.org/10.5281/zenodo.7443839.

RESULTS

The survey contains more questions than are shown here. Some of them have already been reproduced in other surveys (Herres-Pawlis, Liermann & Koepler 2020; Hausen 2019), e.g. on the kind of research data or on data collection. In this section the new and additional results are presented. However, the results will be included in the DMP template and the guidance.

Policies and FAIR

Policy can play an important role for DMPs as the two elements are often complementary. Policies often provide a framework which DMPs may wish to reflect in part. Therefore, the introduction to the DMP refers to a RDM policy and best practice examples.

When asked about their policies, 11 of the interviewees reported the implementation of written policies within their respective working groups. These policies include a range of information like experiments including standard operating procedures (SOPs), documentation wikis, and best practice examples. They also specify details such as file naming conventions and data organisation and storage. Four of the respondents shared information with their colleagues partially in writing and partially verbally, while a quarter passed on information exclusively through verbal communication. Training for new staff is typically facilitated by experienced team members, and RDM is covered in group seminars. Of the interviewees, five are actively engaged in the process of developing guidelines, utilising existing guidelines, or seeking recommendations.

An exemplary case of policy implementation can be found in the Daumann group, which has publicly shared a comprehensive policy on its working group's website (Daumann Group 2022). This policy outlines data generated during projects, provides detailed instructions for ensuring data's FAIR (Findable, Accessible, Interoperable, Reusable) characteristics, recommends tools, and prescribes documentation in electronic lab notebooks (ELNs). The rules span the entire data lifecycle, covering everything from experiment planning and execution to data storage, archiving, and publication. In addition to the example from the Daumann group, guidelines from journals such as Angewandte Chemie were also mentioned. Six interviewees stated that they find these guidelines very helpful and have incorporated aspects of them into their research group.

Furthermore, both guidelines and ELNs are regarded as crucial elements for fully implementing FAIR principles. In the survey, 12 interviewees indicated partial implementation of FAIR principles, while four interviewees did not implement them, and 11 participants abstained from responding to the question. Interviews further revealed that implementing FAIR principles is challenging under current framework conditions. Interviewees primarily focused on the 'Findability' and 'Reusability' aspects of FAIR. Findability, especially within the organisation, was seen as time-dependent, with ELNs and repositories highlighted as key tools for ensuring reusability. 'Accessibility' and 'Interoperability' aspects were viewed more critically overall.

In summary, it is evident that resources for fully implementing FAIR principles are lacking, and weaknesses in 'Accessibility' and 'Interoperability' should be addressed in the DMP template. It is imperative to gather more information and engage in discussions within NFDI4Chem and the chemical community to tackle these challenges effectively.

Storage and decommission strategy of physical samples and data

Another important part of the data management plan is storage and archiving. As mentioned previously, various sub-disciplines within chemistry produce physical samples. Currently, these physical samples are not addressed in any DMP template. However, this gap holds significant importance in the field of chemistry, given the link between data and physical samples. Almost 50% of the interviewees maintain a sample database, documenting details like the substances, their storage locations, and analysis data. In addition to its availability, the continuous maintenance of this database as well as a permanent person in charge are really important. Normally, the physical samples are stored on site in a chemical storage facility or in the laboratory itself under the necessary conditions such as darkness, low temperatures, etc. Three researchers have opted for the Compound Platform - a molecular archive - as an alternative. The Compound Platform serves as a repository for samples which makes the accessibility of substances and the collaborative work easy.

While chemical substances are stored under appropriate conditions, they often have a shorter lifespan than data. Nearly all, 25 from 27 interviewees, indicated that they store both physical samples and data indefinitely (Figure 3).

This perspective contrasts with their personal beliefs about how data storage should be handled. Roughly 50% retain the data for more than 10 years, while seven interviewees consider 10 years of storage appropriate. Five interviewees suggest that storing data for less than 10 years should suffice. Moreover, 10 interviewees mentioned that data is required to compile databases, either for long-term use within their own research groups or for the broader scientific community, aiding in experiment replication and machine learning applications. The majority of the data is stored on institutional servers.
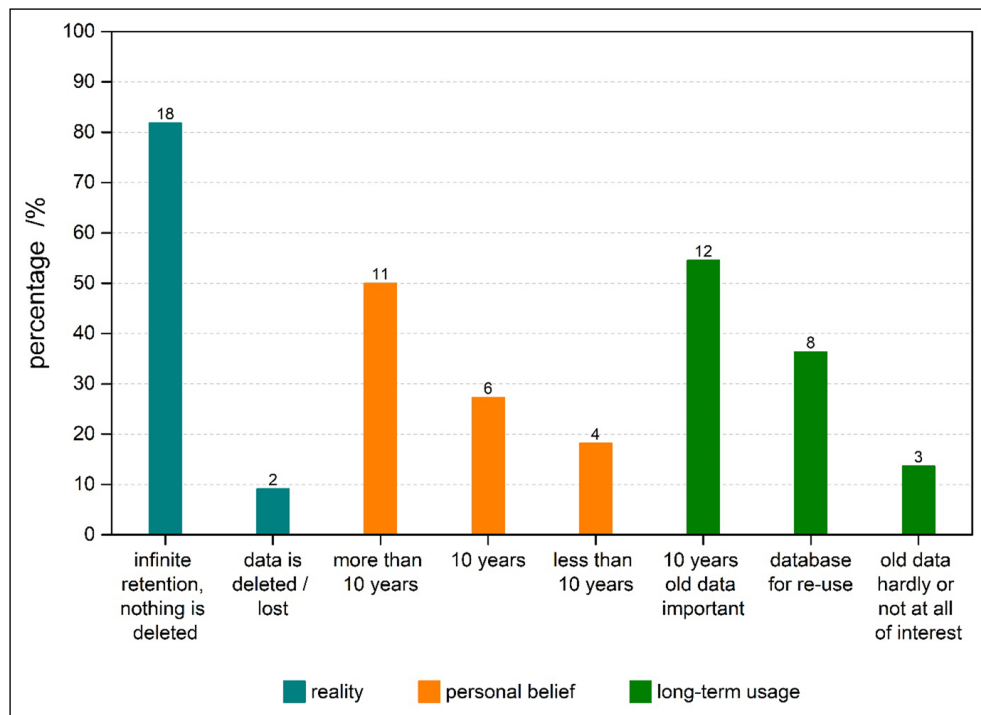
Nevertheless, in general, data at all stages of the examination process are stored, from raw data to processed spectra. This is highlighted by the responses from the survey to the question about saving different types and multiple versions of data (Figure 4).
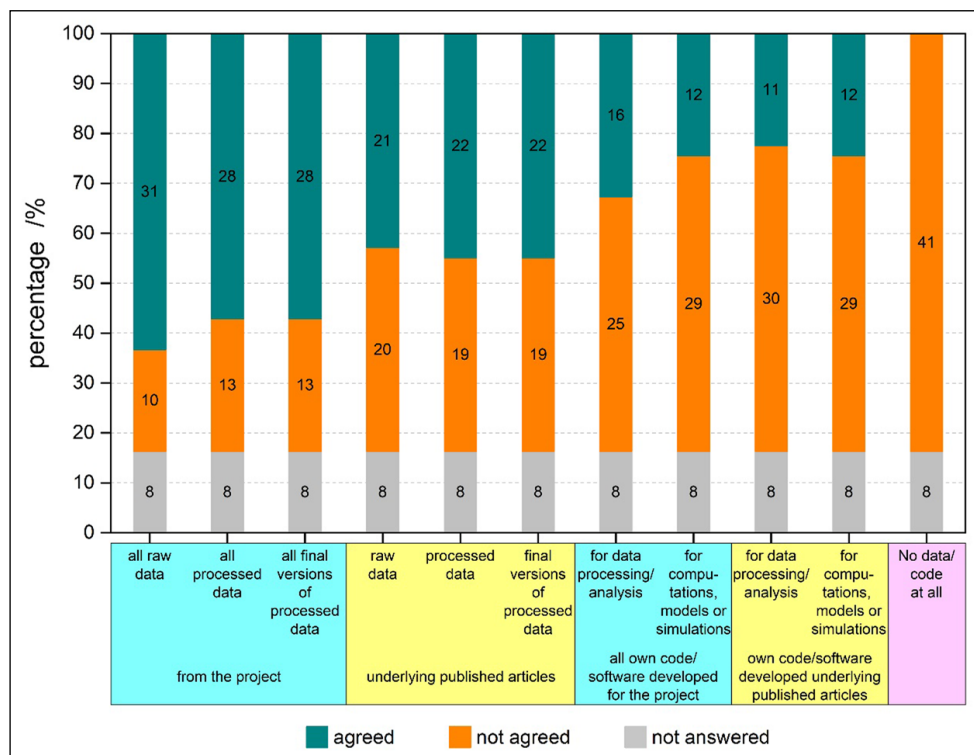
The response indicates an average of four different types of data being stored. The most common types are raw data (17 interviewees), all processed data (16 interviewees) and all final versions of processed data (14 interviewees) from the project as well as processed data underlying published articles and their final versions (12 interviewees each).

Typically, there is no specific strategy for data deletion or the disposal of physical samples. While some participants indicated that the data would not be deleted, some reasons such as the energy crisis or the actuality of the data argued for deletion or a deletion strategy. Therefore, careful consideration should be given to which data and samples warrant long-term

preservation. A chemistry-specific DMP should therefore cover the management of physical samples and include strategies for data deletion and sample disposal.

Software in chemistry

In addition to storing data and samples, it is important to know which software was used to create, process, and analyse the data. As expected, chemists are using a wide range of software solutions for data collection and analysis. All of the interviewees stated they rely on method-specific proprietary software (Figure 5). In terms of data analysis, almost no one sticks exclusively to software specific to their instruments. While 10 interviewees switch directly after data collection to other programs, e.g. electronic lab notebooks (ELNs), Excel, or OriginLab, 15 participants employ both device-specific and other software during their data evaluation. Due to the lack of suitable software solutions, 12 interview partners have to use self-developed software, scripts, or code. For example, one interviewee mentioned that ~50 % of their software is in fact developed by themselves. Interestingly, already 10 interviewees apply open-source solutions, such as open-source ELNs or JupyterLab. However, only five interviewees store their data centralised on institute-provided servers or in clouds. To make it easier for researchers to complete the DMP template, those software solutions already in use should be mentioned. There should also be an indication that there is still a lot of proprietary software in use and that the NFDI4Chem consortium is working on converters for different proprietary formats.
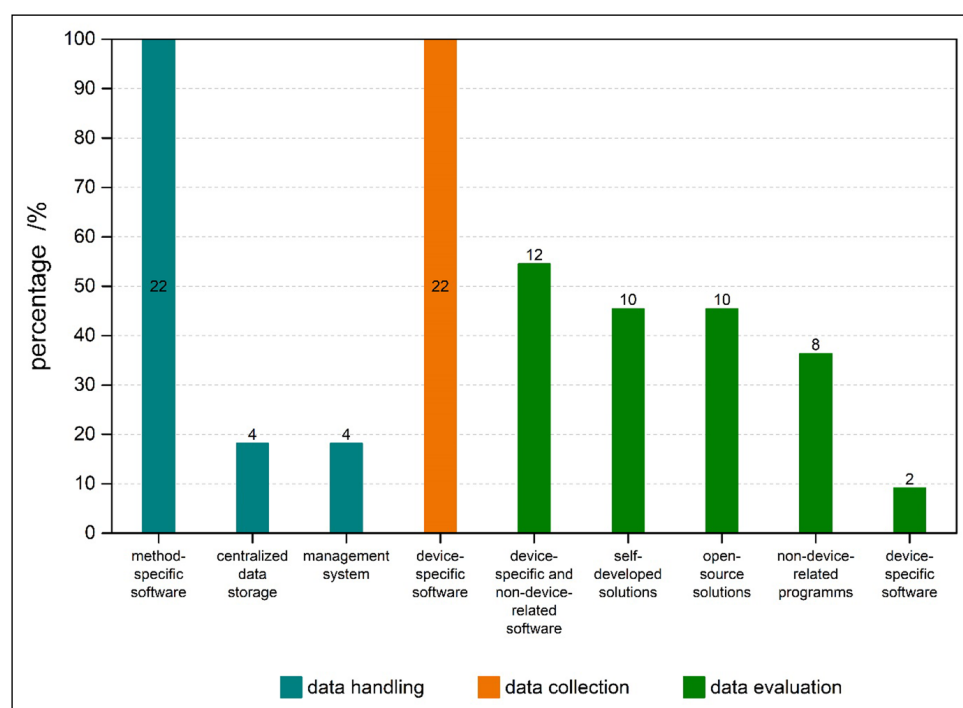

Figure 5 Distribution of categorised software solutions during management of research data. The numbers above the bars are the answers in absolute values.

Data publication, documentation and electronic lab notebooks

As mentioned earlier, chemists have physical samples as well as data that they need to document and manage. Chemists are very used to documentation because chemistry has been using laboratory books for a long time.

In the interview series two electronic lab notebooks (ELN) were explicitly mentioned – chemotion and eLabFTW, which seems to be a good open source solution for physical chemists and material scientists (Cautaerts 2021). The two interviewees emphasised its flexibility, although adapting it to existing frameworks and laboratory conditions took some time. ELNs or currently still widely used paper lab books serve as an important documentation option in chemistry. It obtains for example how molecules are synthesised, how an analysis of certain samples is performed, etc. as well as observation comments, which normally are not passed with a text or data publication. This information represents an essential resource for researchers looking to reuse the information. Therefore, it is crucial in future work to aim for a combination of these two documentation types. For now, the lab notebook topic will be included in the DMP

template to provide context regarding where data documentation occurred and where data can be readily accessed and used by following collaborators.

The question regarding documentation methods received different responses in the survey. In the online survey (OS), participants were asked to rate the use of various documentation types on a scale of 1-5, while in the pre-questionnaire (PQ) they indicated if they use them or not. Nevertheless, the results tend to be very similar. Notebooks (OS 7; PQ 16)[1] and naming conventions (OS 6, PQ 16) were common across both parts. The main difference was in the most frequently named item - hardware/equipment (OS 7) and ELNs (PQ 19).

Another question in the survey targeted naming conventions. Standardised conventions (19 interviewees) were followed by the part using no conventions (11 interviewees) or ad hoc conventions (4 interviewees). The naming conventions often follow a structured format, including a 2–3-digit personal code, a running number for the experiment, and details about the analysis being performed. Some experimental metadata is given in the naming, this facilitates the findability of the data and the sample equally.

Making the data findable is the first step towards data publication. Some ELNs, such as chemotion (Tremouilhac et al. 2020), offer the functionality to publish data directly from the ELN to a linked repository.

In the interview series, 24 interviewees mentioned that they publish the data in parallel with the text publication or at the end of the project. Only three interviewees have not published any data. Of interest for the DMP template is not only the list of chemistry-specific repositories like Chemotion, RADAR4Chem, etc;[2] it's also about clarifying how data should be published. In fact, responses from the interviews vary from data publication via a repository to supporting information (SI). Within the SI, the processed data are provided as images, tables, etc. in a pdf format. The practice of providing additional information in an SI is common practice in chemistry and has grown historically. Therefore, it is not surprising that chemists consider it as a data publication. The survey reflects similar trends (Figure 6), with over 50% of respondents indicating that they publish their data as SI, six interviewees each indicated data publication via a chemistry-specific repository, and using GitHub for sharing code.
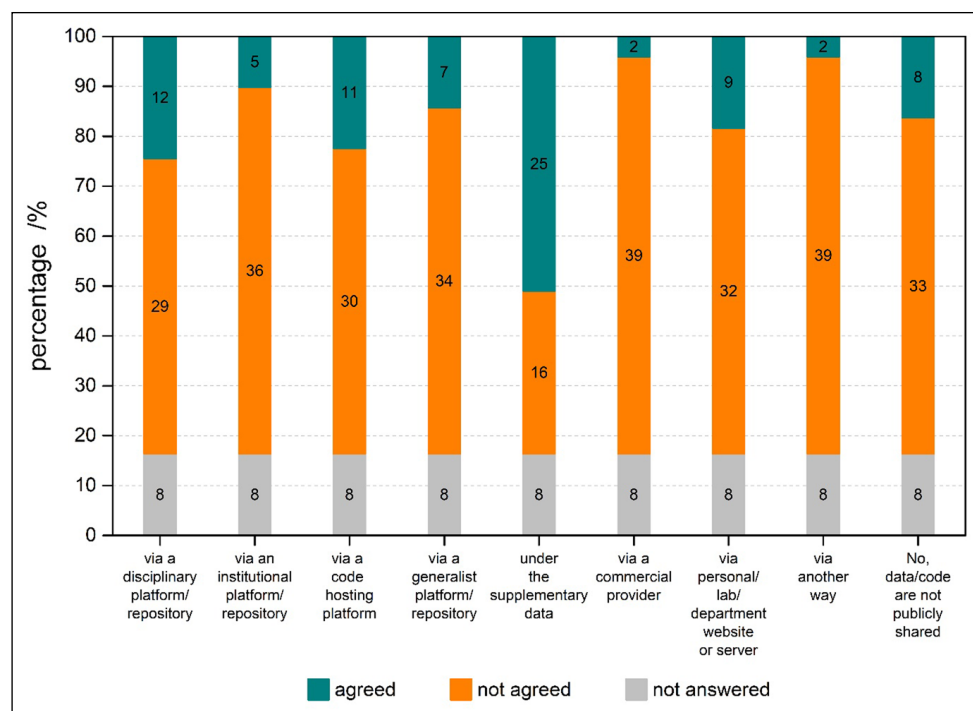

Figure 6 Survey question: Do you or the researchers you support publicly/openly share data/code for the long-term (at least 10 years) after the project? The numbers within the bars are the answers in absolute values.

---

1   Interviewees related to the single categories - online survey 20 and pre-questionnaire 27 participants.

2   A detailed list can be found in the NFDI4Chem knowledge base https://knowledgebase.nfdi4chem.de/knowledge_base/.

As already mentioned, the DMP template will ask for additional information beyond the repository to be used. The DMP should be clear about the process for publishing data and linking to related texts.

Exemplary datasets demonstrating high-quality data publication practices, such as those recognised as the FAIRest datasets (Krausch & Giessmann 2021; Kessler et al. 2021) in chemistry by NFDI4Chem the last years will be linked from the template. These and more can also be found e.g. in the NFDI4Chem Knowledge Base under Lead-by-Example section. As mentioned earlier in the paragraph, another best practice for data publications, which should be linked in the DMP, is the ability to publish data from the chemotion ELN to the chemotion repository. Data publications with embargo periods are possible in Chemotion through a simple internal workflow. Digital Object Identifiers (DOI) are assigned to the data, so that data and text publication can be linked to each other.

Collaborations and obligations

One of the issues that makes data publication difficult is collaboration with other groups or with industry. A total of 25 researchers have past or ongoing collaborations with industry or research institutes, ranging from occasional partnerships to day-to-day collaboration. Typically, research data management, data rights and roles are contractually defined before the project starts to avoid conflicts. There are individual conflicts with regard to the publication of data. Internal discussions delay publication, even where contractual arrangements exist. The EU, which insists on publishing openly, was seen as a negative example.

Another problematic case occurs in the field of machine learning: The affiliation of data and models is not easy to define. If the training data comes from proprietary sources, the data belongs to the industry partner, even if the model used is publicly available. Further complications occur:

- The confidentiality of the data varies.
- In some collaborations, the data from the joint research project is handled openly.
- In other collaborations, access to the data is severely restricted and only the collaborators involved have access rights. The data is stored on a separate server and a partner's laptop is used to send the data in encrypted form.

Within industry collaborations, the industry partner typically presents a specific question, and the outcomes are not predetermined. However, challenges arise when the outcome is predefined prior to the project, preventing the possibility of conducting an independent analysis of the question.

Another important aspect mentioned by an interviewee should be considered from the very beginning and should already be taken into account when creating a DMP: all cooperation partners should work to follow the FAIR principles from the beginning, so that the data could be FAIR in the end. Therefore, recommendations and guidance on cooperation should be included in the DMP template. Additionally, addressing issues such as contractual regulations, data confidentiality, rights, and role management is equally critical.

SUMMARY AND CONCLUSION

In order to develop a subject-specific DMP template, a lot of information is needed to properly support the target subject. In the case of a chemistry-focused DMP template, this includes documentation practices centered around laboratory notebooks, or the linking of physical samples and data. In addition, the development is simplified by starting from an already existing or predefined template and developing it further.

In this paper and for our DMP template, we have taken the DFG checklist as a starting point, as it is relevant for our project and many projects in Germany. To gather information, we used data from the RDA online survey, interviews, and a prequestionnaire.

In summary, the data from the RDA online survey provides a basis that needs to be enriched quantitatively and qualitatively through further data collection. Through the interview series, many examples were obtained that will be used to support the chemistry-specific DMP

template. In particular, these examples can be built directly into the template as a hint and also allow for the creation of response options to choose from when writing the DMP.

Moreover, the survey and interviews revealed that linking physical samples and data is particularly important in chemistry. The DMP template needs to be expanded to include this point. In addition, the DMP template should refer to a deletion strategy as well as a disposal strategy for physical samples. In today's times, it is important to use the available resources wisely. The aspects explained in the interviews as to why data and samples are kept indefinitely make it possible to provide concrete guidance for the deletion and disposal strategy: what should be kept, for how long and under what conditions. What are the reasons for deleting data or disposing of samples? The template will ask the researcher these questions while offering concrete solutions.

During the interviews, it was possible to gather a whole range of software used in chemistry, covering different areas such as measurements, analytics, analysis, etc. It became apparent that researchers are bound to proprietary software when using a measurement technique, but as soon as it goes a step further, many turn to open source solutions. The different software solutions should be listed as answer options in the DMP template. A hint to open source solutions should also be created.

One field of application in which open source solutions are widely used is ELNs. In this context, the respondents mainly use the two open-source ELNs Chemotion and eLabFTW. ELNs are seen as an important building block for the implementation of FAIR principles in workflows.

Likewise, awareness for data publication needs to be created in the DMP template. A significant portion of chemists still 'publish data' as Supplement Information.

Restrictions on data publication arise in chemistry due to collaborations e.g. with industry and patents. Many respondents stated that they have collaborations and patents, but that this is only a small part of the work. However, it affects almost everyone, so in the DMP template attention should be drawn to the fact that the handling of data on patents or on publications must be a different one, which should also be implemented at an early stage.

The results of the interviews were analysed and presented. The information collected from these interviews will be integrated into the DMP template for chemistry. In the next step, a first draft of the chemistry-specific DMP template is published under 10.5281/zenodo.10948510. The information will be included either as help, answer options or as a modified question.

ACKNOWLEDGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Daniela Adele Hausen orcid.org/0000-0001-9083-0670
RWTH Aachen University, Germany

Ann-Christin Andres orcid.org/0000-0001-9496-7906
Johannes Gutenberg-University Mainz, Germany

Jochen Ortmeyer orcid.org/0000-0003-2074-8027
RWTH Aachen University, Germany

Sonja Herres-Pawlis orcid.org/0000-0002-4354-4353
RWTH Aachen University, Germany

REFERENCES

Cautaerts, N. 2021. Using ElabFTW for materials science. *FOSDEM*, 2021. DOI: https://doi.org/10.5446/53345

Daumann Group. 2022. *Research data management plan* https://www.cup.lmu.de/ac/daumann/erc-starting-grant-lanthanophor/research-data-management-plan/.

Deutsche Forschungsgemeinschaft. 2021. *Handling of research data: Checklist for planning and description of handling of research data in research projects* https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_en.pdf.

Deutsche Forschungsgemeinschaft. 2023. *Handling of research data* https://www.dfg.de/en/research_funding/principles_dfg_funding/research_data/index.html.

eLabFTW. *Homepage.* https://www.elabftw.net/.

Fachkollegien, C. 2021. *Handlungsempfehlung zum Umgang mit Forschungsdaten* https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/handreichung_fachkollegien_chemie_forschungsdaten.pdf.

Hausen, D. 2019. *Forschungsdaten in der Chemie*, publiscologne, urn:nbn:de:hbz:79pbc-opus-15024.

Herres-Pawlis, S, Liermann, J and **Koepler, O.** 2020. Research data in chemistry – Results of the first NFDI4Chem Community Survey. *Z. Anorg. Allg. Chem.*, 646: 1748–1757. DOI: https://doi.org/10.1002/zaac.202000339

Kessler, C, Eller, J, Gross, J and **Hansen, N.** 2021. Supplementary material for 'Adsorption of Light Gases in Covalent Organic Frameworks: Comparison of Classical Density Functional Theory and Grand Canonical Monte Carlo Simulations'. *DaRUS.* DOI: https://doi.org/10.1016/j.micromeso.2021.111263

KIT. *ComPlat – the Compound Platform.* https://www.ioc.kit.edu/braese/english/1020.php.

Krausch, N and **Giessmann, RT.** 2021. Collection of UV/Vis spectra acquired while monitoring reaction progress of thymidine phosphorolysis with varying reactant concentrations. *Zenodo.* DOI: https://doi.org/10.5281/zenodo.3243352

Nationale Forschungsdateninfrastruktur. https://www.nfdi.de/.

NFDI4Chem Consortium. https://www.nfdi4chem.de.

NFDI4Chem. *FAIR4Chem Award* https://nfd4chem.t3cm.com/index.php/2022/02/03/fair4chem-award-winners/.

NFDI4Chem. *Knowledge Base* https://knowledgebase.nfdi4chem.de/knowledge_base/.

OpenAire. 2018. *Why is this a good Data Management Plan?* https://www.openaire.eu/blogs/why-is-this-a-good-data-management-plan.

RDMO Community. https://rdmorganiser.github.io/.

Research Data Alliance. *Working Group DSGforDMP* https://www.rd-alliance.org/groups/discipline-specific-guidance-data-management-plans-wg.

Smale, N, Unsworth, K, Denyer, G and **Barr, D.** 2018. The history, advocacy and efficacy of data management plans. *bioRxiv*, 443499. DOI: https://doi.org/10.1101/443499

SoSciSurvey. https://www.soscisurvey.de/.

Tremouilhac, P, Lin, C-L, Huang, P-C, Huang, Y-C, Nguyen, A, Jung, N, Bach, F, Ulrich, R, Neumair, B, Streit, A and **Bräse, S.** 2020. The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry. *Angew. Chem. Int. Ed.*, 59. DOI: https://doi.org/10.26434/chemrxiv.12195318

Wham, B, Hausen, DA, Anders, I, Sheridan, S, Ilamparuthi, S and **Turkyilmaz-van der Velden, Y.** 2022. Dataset and R code from RDA WG Discipline-Specific Guidance on DMP – Online Survey [Data set]. *Zenodo.* DOI: https://doi.org/10.5281/zenodo.8367217