RESEARCH ARTICLE

A survey on the dissemination and usage of research data management and related tools in German engineering sciences

Tobias Hamann [1]
Amelie Metzmacher [1]
Patrick Mund [1]
Marcos Alexandre Galdino [1]
Anas Abdelrazeq [1]
Robert Schmitt [1, 2]

1. Laboratory for Machine Tools and Production Engineering (WZL) | Chair of Production Metrology and Quality Management & Institute for Information Management in Mechanical Engineering (Prof. Robert Schmitt), RWTH Aachen University, Aachen.
2. Fraunhofer Institute for Production Technology (IPT), Aachen.

Date Submitted:
2023-01-01

Keywords:
Research data management, RDM, Survey, Dissemination, Usage of Research Data Management

Data availability:
Data can be found here:
https://doi.org/10.5281/zenodo.7645548

Software availability:
Software can be found here:
https://doi.org/10.5281/zenodo.7645548

Corresponding Author:
Tobias Hamann
t.hamann@wzl-mq.rwth-aachen.de

Abstract. As the amount of collected and analysed data increases, a need for data management arises to ensure its usability. This also applies in research. This challenge can be addressed by Research Data Management (RDM), which brings clear focus on the reusability of data. To understand the status quo of the application of research data management in engineering sciences in Germany, as well as possible challenges and improvement chances, a survey was conducted over the last quartal of 2020. Over 168 (n=168) researchers from the engineering sciences in Germany provided their view via a questionnaire that contains 216 question items. The results give information on the interviewees' knowledge and perceived relevance of research data management in their daily research activities. For instance, the application of research data management related tasks, data sharing with third parties, usage of different tools and the involvement of different file formats were part of the survey. The survey closed with questions regarding RDM specifications, support structures and questions on reasons that could prevent researchers from adapting sustainable RDM. This paper presents the results of the study, providing an overview over the current RDM in engineering and pointing out possible measures and strategies to foster it, namely the integration of guidance and education for research data management. Along the paper we publish the collected data set to enable further analysis and reuse (e.g. for extended statistical analysis).

1 Introduction

As the amount of data has been growing for years [1]–[3], the effort required to manage this data increases. Adding to the sheer amount of data, the requirements of data processing and data reuse further raise the effort in data management. Especially in the context of engineering and industry 4.0 data has to be managed to facilitate the application of related methods as, for example, machine learning [4], [5]. This is not only relevant for industrial applications but also related research performed in engineering sciences. The interest in data collected or generated in the context such research projects is raising as well [6]. Data can be reused to enhance the own research or validate existing results. Therefore, research data management (RDM) is becoming more and more important in many research areas, including engineering. As a result, research data management is introduced to engineering researchers. To facilitate this process, the current progress as well as requirements have to be scouted.

The question arises, what the current status of research data management among researchers in engineering sciences is. As soon as this question is answered, it will become clearer, in which contexts RDM is already applied successfully and in which areas more support is needed. After that, conclusions can be drawn, deriving reasons against the application of RDM and possibilities how RDM can be improved to fit the needs and demands of researchers better.

To get a glimpse on the status quo of research data management in engineering sciences, an explorative survey has been deployed, which asked researchers about the use of RDM in the context of their activities. The survey could sketch out the status of RDM in engineering. Key findings are the knowledge and usage of RDM tools and support structures as well as possible reasons for researchers to not integrate or apply RDM in their research.

To establish a framework delineating the terms of RDM, it is imperative to commence with a precise definition of RDM. ”Research data management encompasses the processes of transforming, selecting and storing research data with the common goal of keeping it accessible, reusable and verifiable in the long term and independent of individuals” [7] while research data is ”(digital) data generated during scientific activity (e.g. through measurements, surveys, source work)” [8].

Furthermore, the context of this survey shall be clarified. Within the framework of the NFDI4Ing consortium, the use and management of research data is to be disseminated and improved. In order to achieve the required improvement, so-called Archetypes and community clusters were used to categorise the research landscape in engineering. These Archetypes cover common fields of research methodologies (e.g. working with experimental or field data, using code or working with material samples). A researcher can relate to more than one Archetype in a fluent way. The community clusters separate the researchers thematically into the five DFG classifications of the engineering sciences that were valid when NFDI4Ing was founded [9].

This survey was prepared and conducted within the NFDI4Ing’s Archetype Frank. Frank’s methodology revolves around the concept of many participants (either as researchers or observed individuals), both human and artificial [9]. Potential users have a background that ”is mostly informed by production engineering, industrial engineering, ergonomics, business engineering, product design and mechanical design, automation engineering, process engineering, civil engineering and transportation science.” [9]. To facilitate the application of RDM, the needs of researchers should be met. To identify such needs, it is necessary to conduct interviews and surveys among a broad cross-section of researchers, who identify with Archetype Frank or work in similar environments [9]. In addition, Archetype Frank has a strong overlap with production engineering and mechanical engineering as stated above, which leads to a partial representation of the NFDI4Ing’s CC41 ”Mechanical and industrial engineering (CC41)” [9] as well.

While there are some publications on the status quo of RDM in general, there is not yet a survey on RDM in engineering sciences with a broad approach in Germany. Therefore, this survey aims to penetrate the circle of potential RDM users in engineering, specifically Archetype Frank in an explorative manner. The survey is intended to give Archetype Frank an overview of the status quo and to enable it to ask more specific questions, for example in interviews or further surveys.

Following this introduction, the next chapters focus firstly on the ”Related work”, before the ”Methodology” used as well as the ”Results” are presented. The paper closes with a ”Discussion” and a ”Summary and Outlook”.

2 Related work

To screen the papers addressing similar questions on the status quo of RDM, a literature review has been performed. This literature review aims to get an overview over similar approaches in the context of RDM. While the focus is set on engineering, other disciplines are also considered whenever they offer an adequate perspective on the topic of this paper.

2.1 Procedure of the literature review

The literature review was performed on the platforms ScienceDirect, Web of Science and IEEE Xplore. The review was last updated in November 2023. Only results newer than the original FAIR Principles [10] were considered relevant, causing results to not date back further than 2016. To perform the review, a search string was compiled based on the terms shown in table 1.

	survey	OR analysis	OR audit	OR check	OR inquiry
AND	research data management				
AND	engineering				

Table 1: Inclusion criteria for the literature review

Firstly, resulting search string was used in three search engines listed in table 2. Afterwards, the results of the search engines were filtered as far as possible (see table 2). Lastly, the resulting papers were exported in the .ris format along with their abstracts.

Search Engine	Last Searched	Filters Used	Results
ScienceDirect	08.11.2023	Year: 2016 or newer	164
Web of Science	08.11.2023	Year: 2016 or newer	53
IEEE Xplore	08.11.2023	Year: 2016 or newer	6
		Sum:	223

Table 2: Used search engines, filters and results for the literature review

The .ris files were imported to the PICO Portal to screen the collected papers for their relevance based on their abstracts. For this screening, certain exclusion criteria were formulated. These are listed in table 3. Any papers matching the exclusion criteria as well as any duplicates were removed from the review process.

Criteria Number	Exclusion Criteria
1.	Not related to research data management
2.	Not a survey or interview or similar data collection
3.	Not related to engineering sciences
4.	Not containing information on the current status of RDM usage/application

Table 3: Exclusion criteria for the literature review

The resulting 23 papers were screened a second time, but based on their full texts. It has to be mentioned that the full text of Todorova et al. about ”Comparative Findings from Data Literacy Survey in Three Bulgarian Universities” [11] was not accessible at the time this paper was written and is therefore not included. Lastly, six papers have been chosen by the full text review.

In addition to the systematic literature review, other sources of literature have been considered as well. The journals ing.grid and BausteineFDM have also been consulted to identify papers that are relevant but are not listed on the aforementioned platforms. Also, Zenodo as an catch-all repository has been consulted. BausteineFDM contained one more paper relevant in this context while in ing.grid’s preprint server, two additional papers could be found. Zenodo included three additional relevant publications. These six papers are also included in this review.

2.2 Results of the literature review

Björnmalm et al. conducted a survey on institutional level on which 21 universities of science and technology united within CESAER participated. They see the challenges of RDM in the lack of “specific instructions (or links to relevant guidelines)” [12] of RDM policies and “support at a faculty level” [12] and in the lack of “lack of trainers in RDM practices” [12]. It is concluded that there are on the one hand too many generic RDM tools, but on the other hand yet too few specific ones. Also, the missing “incentives for researchers that reward and incentivise implementation of RDM practices into everyday workflow” [12] are criticised. One of the recommendations they draw from their survey are the introduction of discipline-specific workflows, that “should provide information tailored to science and technology disciplines, e.g. data infrastructures available for the different types of data produced, different tools for documentation, implications of producing data following the FAIR principles, and when and how to publish their research data. In essence, help researchers make better sense of high-level (university-wide) requirements” [12]. Another recommendation is, to utilise “solutions with open APIs to facilitate the integration of relevant tools and software and to safeguard long-term function” [12].

A presentation of Costanzo et al. on IASSIST 2023 contained the results of two surveys from 2019 and 2022. The focus was laid on the application of the ”Tri-Agency RDM Policy” [13], that states “to support Canadian research excellence by promoting sound RDM and data stewardship practices” [13]. Main institutions representing the “Tri-Agency RDM Policy” are the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) [13]. Main barriers for the proper application of RDM are the ”lack of resources (time, budget, personnel etc.) [,] lack of institutional understanding and awareness of the Tri-Agency expectations [and] lack of availability of support materials” [13].

Austin et al. reviewed ten engineering research projects that have been conducted as Open Research Data pilots at the Horizon 2020 research programme. While the paper sets a focus on avantgarde projects that specifically aim for the application of RDM, the findings for engineering sciences still offer a value for this paper. The ”need to demonstrate to researchers the value of data management” [14] is clearly stated to point out the need for a change in research culture. More than half of the involved partners rejected data sharing. Another challenge is the effort of RDM, as ”data gathering tasks will remain a significant burden [...] until [...] data technologies (i.e. interoperability standards) required for seamless data exchange and aggregation” [14] have been developed. While possible solutions are also discussed, the presented challenges in the presented projects can be expected to occur in most research projects in engineering sciences.

Wilms et al. present ”a quantitative study of the factors affecting researcher’s intention to comply with guidelines on handling research data” [15]. A total of 111 researchers from the discipline of information systems in Germany responded to the survey. While the subject of information systems is part of the IT sciences, it is still considered technical enough for this paper. They point out that the ”overall acceptance of RDM policies is low” [15], that ”90 % of the participants indicate that they do not use institutional or national standards” [15] for research data management and that ”a large part of respondents claimed not to practise RDM” [15]. The ”requirement to comply with possible guidelines is clearly not sufficient to convince researchers to change their current inadequate data management strategies” [15]. On the one hand, uncertainty is listed as one possible explanation, as it results from the fear of losing control over the own data, on the other hand ”uncertainty can prevent people from choosing an option even if they evaluate it as more beneficial” [15]. Another reason for the lack of RDM usage is the ”perceived increased workload” [15]. A possible solution might be the provision of technologies to support RDM and “convince them that no additional technical effort is required” [15].

In 2021, Polona Vilar and Vlasta Zabukovec conducted an online survey on research data management in Slovenian science, including engineering sciences [16]. They differentiate between the perception and the behaviour of researcher to point out groups of researchers based on their discipline. They state that researchers from the engineering sciences perceive RDM as unproblematic and are rather convinced by it. In terms of behaviour, engineering researchers show a considerable spread in their answers. Some do not utilise metadata and follow no file-naming conventions/standards, while others often use file-naming conventions/standards along with version-control systems and are experienced with public-domain data.

A similar survey has been conducted in Iceland by Palsdottir in 2017. Out of the 139 respondents about 39% originated from sciences, containing engineering sciences [17]. It was found that ”the researchers had limited knowledge about the procedures of data management [, ...] it is not a normal practice in their research work” [17] and ”that there is an urgent need to increase the researcher’s knowledge and understanding of the importance of data management [...], as well as to provide them with the resources and training that enables them to make effective [...] use of data management methods” [17]. It is concluded that information specialists are needed to assist in the design of RDM services to support researchers in their data management [17].

From March to May of 2020, Israel et al. ”conducted an online survey among research physicists in Germany [...] to determine the status of their RDM and the resulting agenda for an NFDI consortium" [18]. While the focus lies on physicists, it has a very similar scope to this papers goal in performing a broad survey on the status quo of RDM. 237 complete answers from universities all over Germany could be collected via the survey. This survey was also conducted in the context of the German National Research Data Infrastructure (NFDI) initiative. Their findings point out that "documentation of research activities is not as seamlessly digitized" [18], for instance instead of electronic laboratory notebooks (ELNs), paper laboratory notebooks are still being used. The main challenges of RDM are stated as the "complexity in data structures and formats (69% approval), the large number of tools and methods (61% approval), complexity of documentation (59% approval), and confusion about underdeveloped metadata standards (50% approval)" [18]. Their most important conclusion in the context of this paper is the following: "The 2020 survey on RDM in physics has shown that making data FAIR needs to start at the foundational level of terminology, file formats and, most importantly, awareness." [18]. Physics sciences in Germany do "not live up to the standards of RDM best practices" [18].

In contrast, Ortloff et al. [19] point out that the "interviewed partners are aware of the Open Access requirements and the FAIR principles" [19] and that "most of the partners are strongly aware of the benefits provided by extended data usage and the respective demands" [19]. While they conclude that "there are concerns regarding IP protection and data security" they also state that "establishing proper templates, guidelines, and training for data collection, analysis, and sharing" can improve RDM practices. A cultural shift is seen as urgently needed in many of the interviewed organisations [19]. These conclusions are drawn from a "spotlight investigation" [19] based on expert interviews, not a wide range of researchers from engineering.

When taking a look at life sciences and engineering in the universities in Egypt, Jordan and Saudi Arabia, Elsayed and Saleh [20] found, that "42% [of researchers are] unfamiliar with data management plans" [20] and "more than half [... have] no data management plan". They state, that "despite researchers' recognition of the importance of data sharing, they lacked the capability to actually share data" [20] and that "the practice of depositing data in open data repositories was not prevalent" [20]. "56.9% indicated that they needed training in RDM" [20].

A presentation by Melissa Cheung at IASSIST May 2021 points out restrictions on data sharing in engineering. Again, the concern about "intellectual property rights (24.4%)" [21] is listed as very important, second to the "Need to publish before sharing (50.2%)" [21].

Chawinga et al. describe motivational factors as well as challenges listed in 105 papers. While the motivational factors shall not be discussed here, the challenges of RDM need to be taken into consideration although the focus of Chawinga et al. is set on funding and institutional matters, they still point out that 92.4% of papers list the data sharing skills as an issue for RDM [22].

Wuchner et al. present a case study with no broad survey. Still, there are findings specifically relevant for engineering sciences. They point out the lack of clearly defined or even standardised processes. Additionally it is stated, that "for the researcher, obtaining the project partner's consent for publication was the biggest hurdle" [23], reinforcing the statement of Ortloff et al. [19] about concerns regarding intellectual property protection. If researchers are introduced to new tasks, assistance is needed, for example, in the case study "the researcher needed assistance in the publication process, especially since it was his first" [23]. There is a "need for experts to assist researchers with data publications and overall research data management" [23], last but not least because ”data publications – especially FAIR ones – are a major challenge for researchers” [23].

While their paper is set in neuroimaging, Borghi and Van Gulick point out the current challenges of RDM in their field. They figure that the researchers ”ubiquity indicates that there is not an optimal amount of communication about the importance of RDM even within individual research groups or projects” [24]. Additionally, they point out limitations of RDM and reasons against data sharing. Limiting factors are ”the amount of time it takes [... with at least] 69.60%[, a] lack of best practices [... with at least] 43.20%[, the] lack of incentives [... with at least] 32.18% [and the] lack of knowledge/training [... with at least] 32.80%” [24]. The main reason against data sharing is the fear of use of not yet analysed/sensitive data, with 50.43% respectively 30.43%.

While the presented literature does not fully match the scope, all relevant findings are discussed in chapter 5. All publications presented either include RDM (in engineering) in a broader (e.g. nation wide) survey like [16] and [17] or refer to certain use cases or projects like [14]. The focus on RDM in Germany can only be found in related fields like IT sciences [15] or physics [18]. No literature found contains direct information on the status quo of RDM in engineering.

3 Methodology

This chapter introduces the methodology of the conducted survey. Firstly, the interviewees and the approach are discussed, followed by the surveys structure and the categories of questions contained. As a result both the interviewees and the questions are clarified before the results are discussed in chapter 4. The survey was implemented within the online tool `soscisurvey.de`.

3.1 Interviewees and Approach

The survey took place from October to December 2020. 168 researchers were interviewed, most of which are employed as research assistant seeking a doctoral degree (64.2%) (see figure 1).

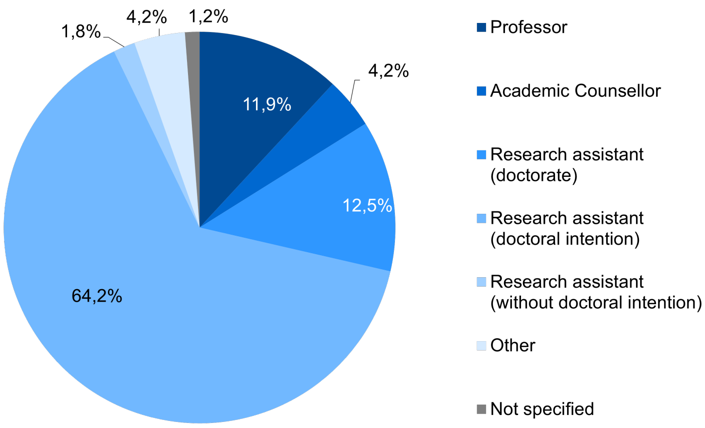

Figure 1: Occupation of the surveys participants

The surveyed researchers are composed of members of the ”Scientific Society for Production Engineering” (”Wissenschaftliche Gesellschaft für Produktionstechnik”, in short WGP), the ”Scientific Society for Product Development” (”Wissenschaftliche Gesellschaft für Produktentwicklung”, in short WiGeP) and researchers from the RWTH Aachen Cluster of Excellence

”Internet of Production” (IoP) as well as members of the ”Fraunhofer-Verbund Produktion”. These consortia stand for ”Cutting-edge research [...] in the area of basic research as well as applied and industrial research” [25] with a ”close collaboration with economy and science” [26] as well as a strong focus on ”application-oriented research” [27]. The IoP states a ”balanced composition of participating researchers from five faculties at RWTH Aachen University and six non-university research institutions” on their website [28].

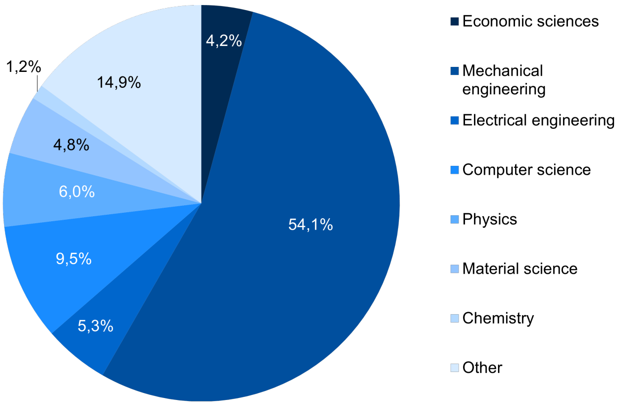

Figure 2: Subject area of the surveyed participants

All of the listed organisations are focused on engineering, particularly in mechanical engineering and production technology. However mechanical engineering often involves interdisciplinary approaches. Thus, plenty of subject areas are represented within the interviewees. As a result, the survey represents not only Archetype Frank but also gives insights into Community Cluster 41. Figure 2 depicts the subject areas of the interviewees. More than half of the surveyed researchers are from the subject area of mechanical engineering. The other half is a wide mix of different subject areas. While some more are in the scope of mechanical engineering and production technology than others, all of them are researching within the context of production technology.

3.2 Survey Structure and Questions

The survey consists of 216 question items, starting with a demographic inquiry of the respondents’ data to validate the fit of the respondents. This is followed by an exploratory self-assessment, which contains three introductory questions to the overall usage and knowledge of RDM.

Category	Number of questions
Demographic data	7
Explorative questions	15
General RDM questions (FAIR , DMP, DLC)	3
Data life cycle	27
Tools	116
File formats	39
Specifications and support structures	8
Acceptance aspects (free text)	1

Table 4: Summary of the topics and their corresponding number of question items within the survey

Interviewees were questioned if they are aware of the FAIR principles [10] for research data, if they (or a third party, if applicable) create a data management plan and if they base their research on the data life cycle. The self-assessment is followed by detailed questions of how research projects carried out along the data life cycle as proposed by forschungsdaten.info [29]. The questionnaire is rounded off by the question about the support available to the respondents. The opportunity to add further comments via free text is given to the respondents throughout the survey. The structure of the questionnaire with question categories of the survey and the corresponding numbers of questions contained can be found in table 4. Free text answers are included within the numbers of questions stated in the table.

4 Results

After validating the fit of the respondents background in terms of discipline and employment, the actual evaluation of the survey results follows. This chapter is based on the structure of the survey mentioned in chapter 3.2 and is subdivided accordingly.

4.1 RDM Knowledge and Perceived Relevance of RDM

The first set of non-demographic questions aims at providing a rough assessment of the respondents knowledge on RDM in general. Regarding research data handling, more than half of the respondents stated that their knowledge was moderate or lower. Only 42.3% stated that they had a high or very high level of knowledge regarding the handling of research data (see figure 3). At the same time, over 57% of respondents rate RDM as important or very important. Only about 15% perceive RDM as unimportant or completely unimportant (see figure 4).

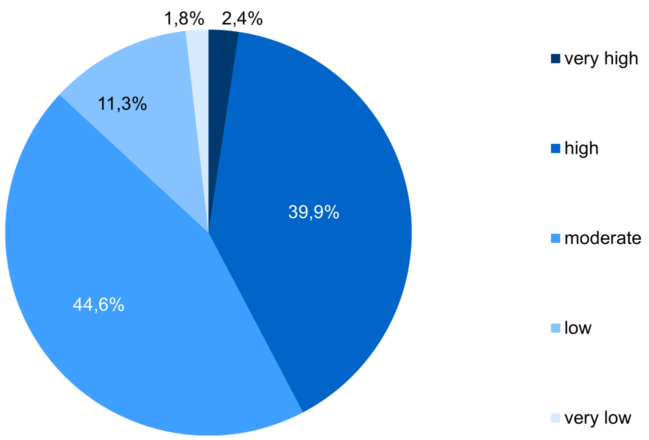

Figure 3: Self-assessed RDM knowledge of the participants

When comparing those two statements above, there seems to be a gap between the group of researchers with (very) high RDM knowledge and a (very) high perceived importance of RDM. There are 14.7% less researchers who have a RDM-knowledge specified as high or above than there are researchers who perceive RDM as at least important. This leads to the first conclusion of this paper, that there is a gap in knowledge of researchers. Additionally, missing knowledge may also lead researchers into perceiving RDM less important, potentially widening the gap.

1. *There is a need for RDM knowledge among researchers in the engineering sciences,*

specifically for researchers of the Archetype Frank respectively amongst researchers in the field of mechanical engineering and production technology (CC41).

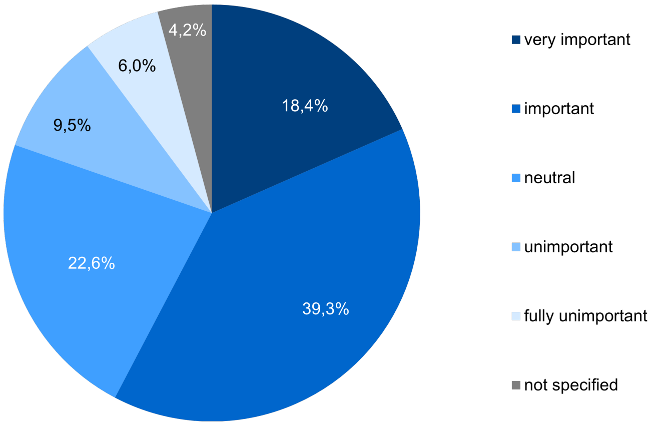

Figure 4: Perceived relevance of RDM among the participants

To better understand the relevance and reliability of the self-accessed RDM knowledge, the following question was asked: ”Have you ever heard of the FAIR principles (Findable, Accessible, Interoperable, Reusable) [10] for research data?”. The responses are shown below in figure 5.

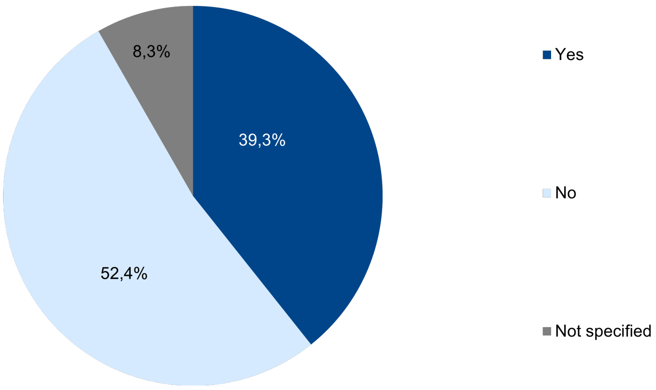

Figure 5: Percentages of interviewees who have ever heard of the FAIR principles, see [10]

The survey also asked for the usage of the Code of Conduct of the ”Guidelines for Safeguarding Good Research Practice” published by the DFG [30]. These have already been applied several times by almost three quarters of all respondents (see figure 6), however this does not lead to a consistently high level of knowledge regarding research data management. The correlation coefficient between these factors is 29.5%, which does indicate a mild correlation. Generally speaking, the correlation coefficient measures how close two values are linearly dependant [31]. As the correlation coefficient is positive, this indicates an increase in RDM-related knowledge when a person regularly uses the DFG guidelines. This effect can also be seen in figure 6.

A similar effect, can be seen between the perceived relevance of RDM in the interviewees own dissertations and the knowledge about RDM (see figure 7). Here, the correlation coefficient amounts to 33.1%, indicating a mild positive correlation, meaning that the more important RDM is perceived in context of the one’s own dissertation, the more one knows about RDM [31].

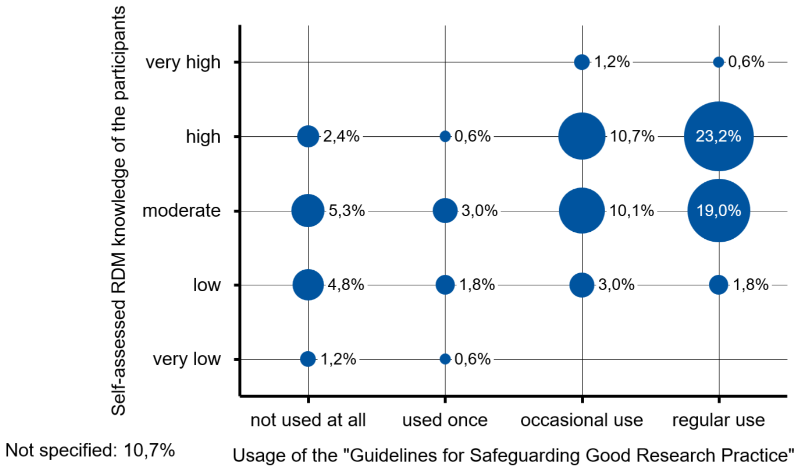

Figure 6: Perceived relevance of RDM among the participants in dependency of the usage of the Code of Conduct of the ”Guidelines for Safeguarding Good Research Practice” by DFG

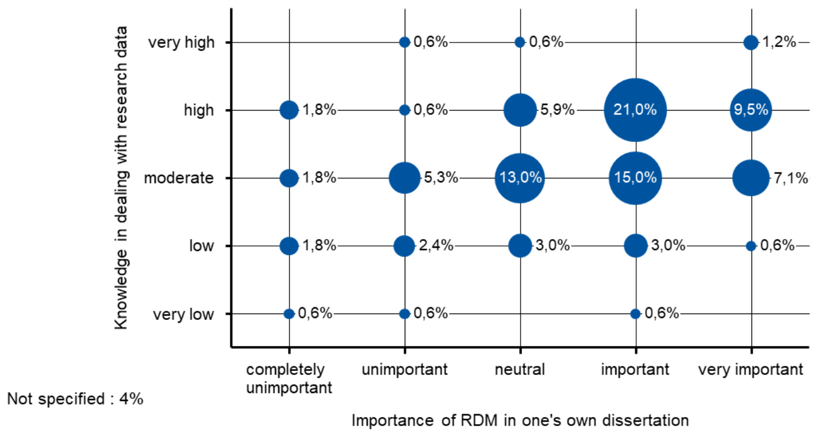

Figure 7: Perceived relevance of RDM among the participants in dependency of the perceived relevance of RDM in the researchers own dissertation

4.2 Application of RDM Related Tasks

While 58% (see figure 7) claim to find RDM important in their own dissertation, the self-assessed knowledge amongst the interviewees is mostly moderate to very low. Moreover, the claim of regular use of the ”Guidelines for Safeguarding Good Research Practice” is questioned by the answers of the interviewees in the later questions of the survey. For example: The Guidelines state that “Researchers decide autonomously [...] whether, how and where to disseminate their results.” This includes the process of determining copyrights and the control of access, which is especially important when handling data that is not shared due to reasons such as secrecy or of patent applications. In that case, a decision has to be made to control the access to only those who are allowed to access such data. However, less than 10% of the interviewees regularly determine copyrights, control access or share their data (see figure 8).

Even less make their data publicly available (<5%). To set this into perspective, 44.6% of the surveyed researchers claimed to regularly use the DFG’s ”Guidelines for Safeguarding Good Research Practice” [30]. In other words, only about one in nine researchers who regularly use this guideline "make all results available as part of scientific/academic discourse", although research data should be included "where possible and reasonable" [30] as proposed by the DFG.

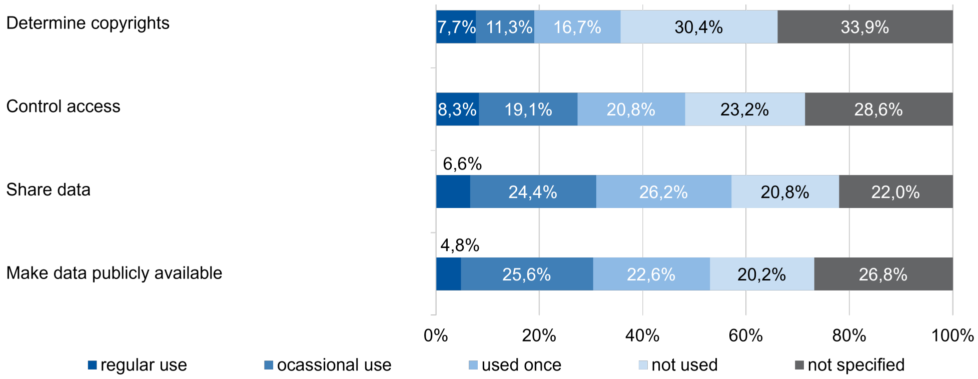

Figure 8: Data life cycle: Activities from the sharing phase

Similar low rates of regular application of research data management tasks can be observed throughout various steps of the data life cycle. This indicates the following conclusion:

2. *While the use of Guidelines like the "Guidelines for Safeguarding Good Research Practice" tend to improve the self assessed RDM knowledge among the interviewees (see figure 6), it does not necessarily imply the application of RDM connected tasks.*

The only step of the data life cycle that has a high rate of regularly performed tasks is the "prepare and analyse data" phase, as shown in figure 9. The highest rated task is "Interpret data", which scores a 38.1% regular application rate. An additional 36.3% occasional application rate is adding up to 74.4% of the researchers who at least occasionally interpret their data on their own. Taking into consideration that 16.1% of the interviewees are professors or academic councillors, this initially rather low rate of data interpretation among researchers becomes clearer.

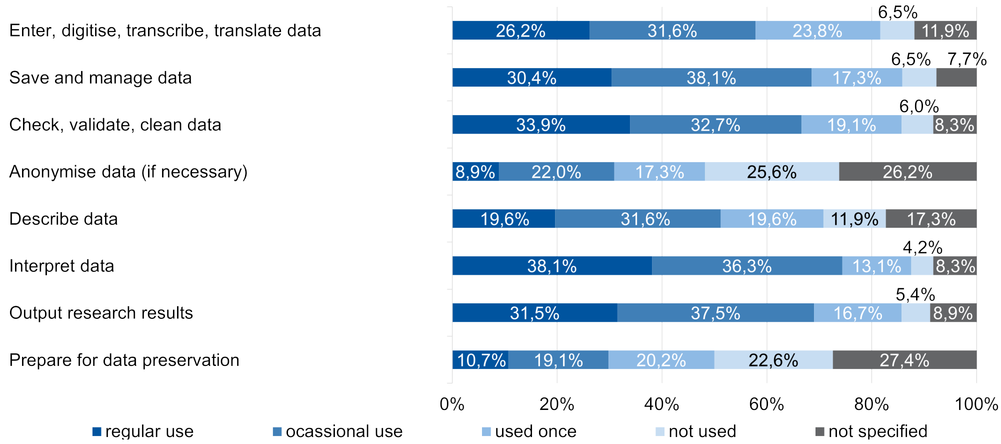

Figure 9: Data life cycle: Activities from the prepare and analyse data phase

This leads to the next conclusion this paper draws:

3. *RDM-related tasks that are not directly part of the everyday research activity (like determining copyrights) are much less likely to be carried out than those who are mandatory to receive results from data, such as transcribing, preparing, interpreting or validating data.*

4.3 Data Sharing with Third Parties

Another set of questions asked about the willingness to share research data with third parties and the reuse of third party research data. This set of questions however seems to be inappropriately specified, as the results are inconsistent. One participant gave feedback on this topic:

”The questions [regarding sharing research data with third parties] are flawed, as the attitude towards any third party is different than within the institute or a network.”

Anticipating focus group interviews that took place months after the survey with different participants, it can be said that this definition of ”third parties” harshly varies in the understanding of researchers. The questions in this survey aimed towards the interpretation of third parties as ”not related to the research project in any way”. This however seems to be misinterpreted by some of the participants. The questions that asked for the data life cycle, specifically the ones for the sharing data phase, show that 57.2% shared data at least once, which was shown in figure 8.

When asked for the actual possibility for third parties to access one’s own research data, this value raises to 65.5%. This can be explained in two ways:

1. The additional 8.3% of interviewees did not specify an answer in the corresponding question set at the data life cycle section of the survey.
2. The surveyed researchers interpreted the expression ”third party” as ”involved in the actual research project, but not part of the own institute”.

It is unclear which of the two applies in this case. It has to be noted that, although the expression ”third parties” is used in the ”Guidelines for Safeguarding Good Research Practice”, it is never specified in the document itself [32].

4.4 Usage of RDM Tools and Services

The next part focused on tools and services. A distinction is made between usage and awareness of tools. The term usage refers to the following options: ”regular use” and ”occasional use”. Awareness means the tool is either ”known by name” or has at least a ”one-time use”. Respectively, unawareness refers to the option ”unknown”. A ”not specified” option was given as well.

More than 70% of all responses are ”unknown”. A further 19% are assigned to the answer option ”not specified”. It has to be noted that this distribution also applies if only the answers of those are taken into consideration, who have stated to have a high or very high self accessed RDM knowledge. In this case, 69.3% answered ”unknown” and 20.3% answered ”not specified” or did not answer the question at all. In general, the answers of the respondents are strongly polarised. A few tools stand out due to regular use, while others are almost completely unknown.

Literally the most prominent example is Git, with 72% awareness among respondents. Almost 30% use the tool regularly and 25% occasionally. 7% have used Git at least once and 10% are familiar with it by name. No other tool has a similar level of awareness and use among researchers. Although mySQL is better known than Git (78.5%), it is used much less frequently (regularly 12% and occasionally 22.6%) and is limited to one-time use (28%).

An overview of awareness (”known by name” and all mentions of useage) and usage (sum of the mentions of ”occasional” and ”regular use”) is given in table 5, sorted by the proportion of respondents who state multiple uses. Due to the large number of tools surveyed, only those used more than once by at least 5% of the respondents are mentioned below for the sake of clarity.

Tool/Service	Category	Awareness [%]	Usage [%] ▽
Git	Data organisation	72.0	55.0
mySQL	Databases and repositories	78.6	34.5
DOI Citation Formatter	Citation	45.8	30.4
KeePass	Password help	44.6	26.8
TIB PID Competence Centre	Persistent identifiers	35.1	22.0
Microsoft Project	Collaborative work	64.3	20.4
NoSQL	Databases and repositories	42.9	14.9
TortoiseSVN	Data organisation	34.5	14.9
TortoiseGit	Data organisation	32.7	11.9
PostgreSQL	Databases and repositories	29.8	8.9
Google Dataset Search	Find research data	32.7	8.33
STD-DOI	Citation	17.3	8.33
Apache Subversion	Data organisation	23.8	7.7

Table 5: Awareness and use of tools among researchers sorted by use among respondents

As shown in table 5, of the 90 tools and services surveyed, only 13 have been used more than once by at least 5% of the respondents. Seven of those 13 come from the field of software development, i.e., they are directly or indirectly related to programming. Those can be recognised by the categories ”Data organisation” and ”Databases and repositories”. The remaining six tools/services are two tools for citation (DOI Citation Formatter and STD-DOI), one for persistent identifiers (TIB PID Competence Centre), one for finding research data (Google Dataset Search), a password organiser (KeePass) and a tool for collaborative working (Microsoft Project).

Tool/Service	Category	Awareness [%]	Usage [%] ▽
Microsoft Project	Collaborative work	88.9	45.8
mySQL	Databases and repositories	69.4	43.1
Git	Data organisation	40.3	31.9
KeePass	Password help	31.9	22.2
NoSQL	Databases and repositories	48.6	20.8
TortoiseGit	Data organisation	34.7	20.8
DOI Citation Formatter	Citation	30.6	20.8
TortoiseSVN	Data organisation	33.3	19.4
Google Dataset Search	Find research data	36.1	18.1
TIB PID Competence Centre	Persistent identifiers	26.4	15.3
PostgreSQL	Databases and repositories	37.5	13.9
Apache Subversion	Data organisation	26.4	9.7
STD-DOI	Citation	15.3	9.7
GNU Arch	Data organisation	30.6	5.6

Table 6: Awareness and use of tools among researchers who have stated to have a high or very high self accessed RDM knowledge sorted by use among respondents

As shown in table 6, a similar distribution can be observed when only reviewing the answers of researchers who have stated to have a high or very high self accessed RDM knowledge. Here, 14 have been used more than once by at least 5% of the respondents. The same focus on software development becomes apparent with eight of the 14 listed tools related to this area.

The majority of the best-known or most-used tools have in common that they offer solutions to researchers' everyday problems (compare finding 3). For example, the versioning tool Git offers a possibility to version source code, which can hardly be kept manageable without versioning. The added value of Git is known and is also passed on to other researchers, at least in groups that regularly deal with source code. The immediate applicability is what separates those best-known and most-used tools from especially the less-used RDM tools.

Such RDM tools that should mainly accompany the research process, are virtually unknown and unused. The majority of respondents thus lacks knowledge about suitable programs, supporting tools or services in the context of RDM. Therefore, such programs, tools or services are not used by the majority of respondents, which is another core finding of this paper:

4. *Researchers lack awareness about existing solutions for RDM specific problems and therefore the knowledge and ability to use those solutions.*

4.5 Usage of File Formats

The survey also asked about the frequently used file formats. 31 file formats as well as opportunities for free text answers were given. The interviewees could choose whether or not they use that file format. File formats cover the MS Office family, PDF and common image and video formats as well as formats for quantitative data and text-based formats. The later ones also contain file formats for source code such as .py or .cpp.

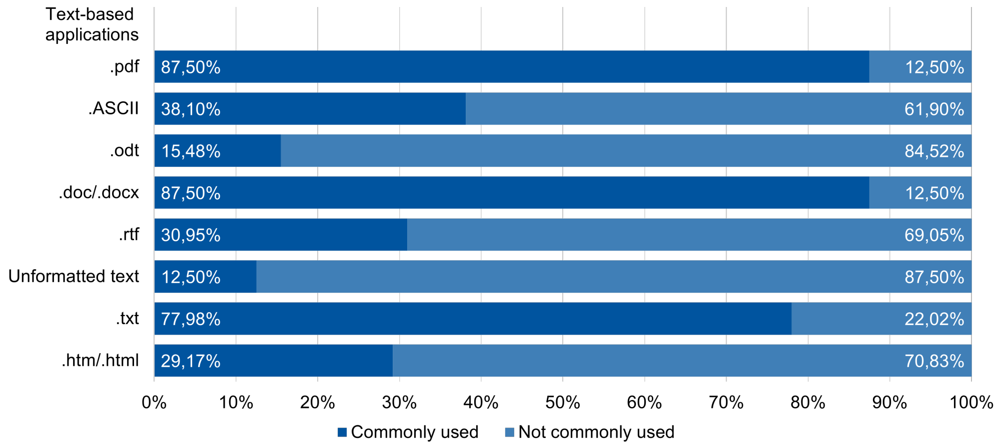

Figure 10: Common usage of text-based file formats among interviewees

When reviewing the results for file formats in text-based applications, a strong distinction between commonly used and not commonly used formats is possible (see figure 10). MS Word files (.doc or .docx), just like PDF documents, are frequently used by 87.5% of the respondents. With 78.0%, .txt is the most frequently used format for unformatted text. Other file formats are commonly used by a minority of the interviewees as shown in figure 10.

MS Excel files (.xls or .xlsx) are used by 87.5% of the respondents (see figure 11). Close behind (86.3%) is .csv, another file format usable in Excel. Again, other file formats are much less commonly used than the aforementioned, making the distinction between commonly used file formats and not commonly used file formats very unambiguous.

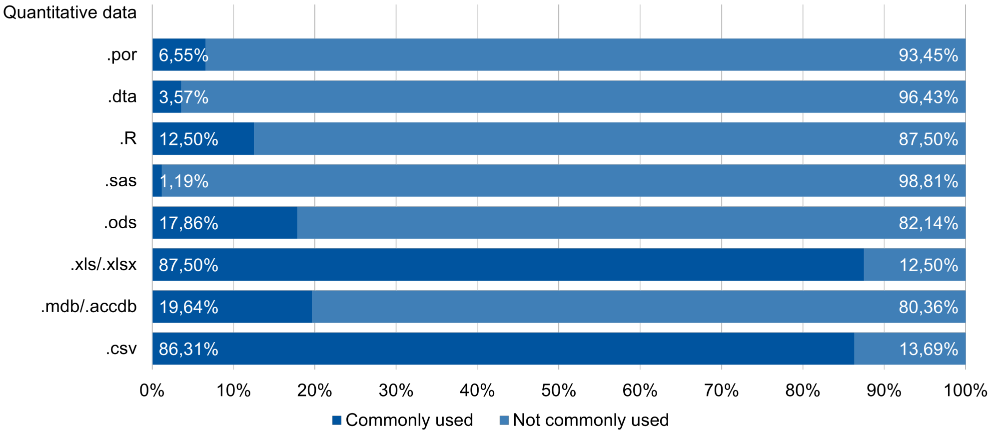

Figure 11: Common usage of file formats for quantitative data among interviewees

For media files (image, audio and video files), the spread in the answers given is not nearly as pronounced as for example in quantitative data. However the aforementioned formats .jpg/.jpeg, .png, .mp3 and .mp4 are predominant for their respective category (see figure 12).

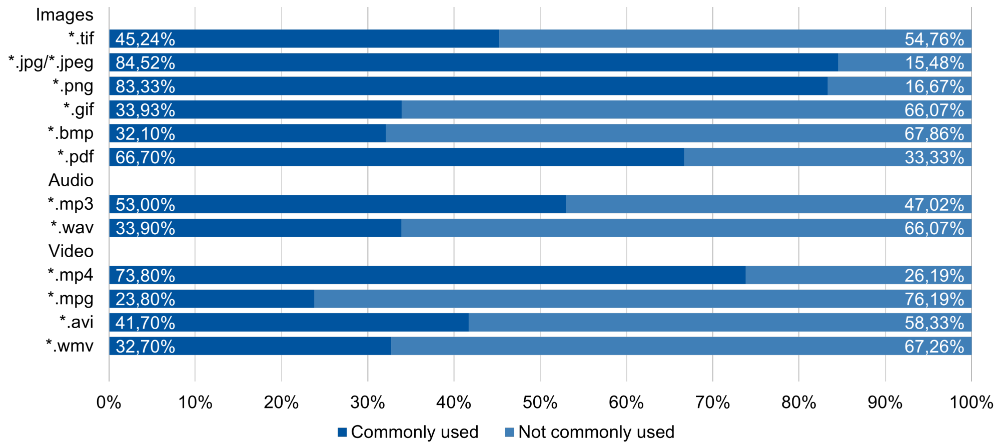

Figure 12: Common usage of file formats among interviewees

The commonality of the aforementioned formats is their general widespread use, familiarity and the resulting usability. All these can be used on a standard Windows PC with MS Office installed, without the need for further installations. The latter is a factor not to be neglected. On the one hand an installation of further programmes may have to be carried out by corresponding IT departments, which is associated with personnel and time expenditure. On the other hand, depending on the file format, there are licence fees for associated programmes. The latter becomes more important if there are free or already available alternatives in the work environment.

This relation is expressed most strongly in the processing of quantitative data, e.g. table-based evaluation of data through Excel. MS Office, including Excel, is one of the standard installations on Windows PCs, as already mentioned above. Therefore, the use of .csv, .xls and .xlsx files is possible on the majority of Windows PCs; these formats are used by 87.5% of the respondents. In contrast, the use of the .por format, which was developed by IBM for the statistical programme SPSS and is only used by 6.6% of respondents, is only possible in this very programme [33].

For other formats in the field of quantitative data, the usage rates are hardly higher and formats usable with Excel seem to be the only option. In contrast, only 15.5% of respondents use the .odt format, although this can also be opened and edited in licence-free and openly available programmes.

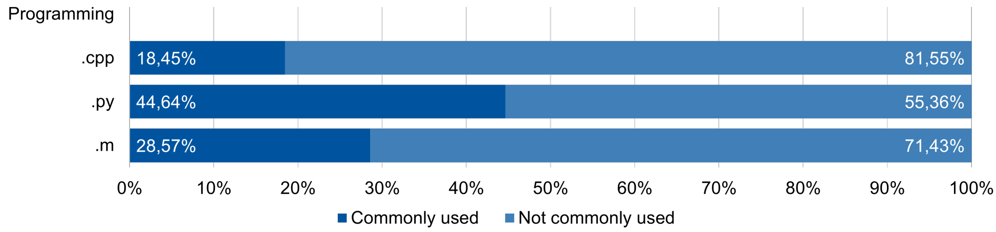

Figure 13: Common usage of file formats used in programming among interviewees

The usage of file formats is primarily based on programmes and tools available and the usability of the formats. The usability is partly dependent on the availability of programmes or their corresponding licences. It is unclear why specific programming languages and file formats (see figure 13) are used in software development. The reasons for or against an approach are not part of the survey, as researchers should be supported in everyday research and not forced into new directions. The collected knowledge about the used file formats used does not provide any direct recommendations for action to advance RDM. It rather shows the heterogeneous file formats that need to be taken into account when working with research data.

4.6 Specifications and Support Structures

The last question set is directed at the requirements and support structures for RDM that are specified or offered by the respondents' respective institution. Those include, but are not limited to, RDM-Teams at universities, available tools for RDM or specific support at institutions. The exact question was ”Is there support within your organisation in the area of RDM?”.

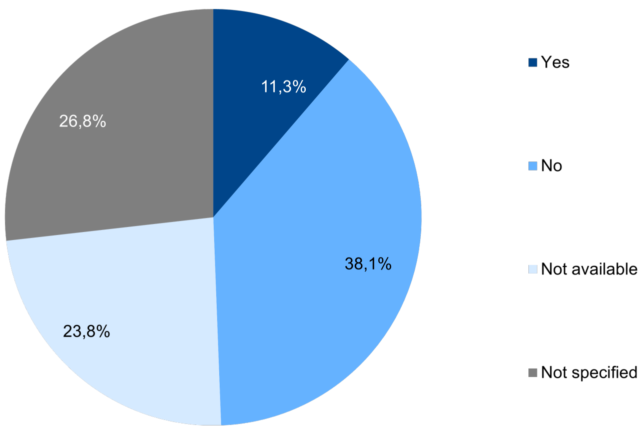

Figure 14: Usage of offered support structures at the interviewees' own institution

Shown in figure 14 are the responses of researchers asked if they use offered support structures at their organisation. Only about one tenth of the surveyed researchers have used offered support structures while almost a quarter states there was no support available at their institution. The survey did not include any questions regarding why support structures are not used by researchers.

However, there might be two reasons for this. Firstly, support structures are available but not known, which is relevant only for the 23.8% of researchers who claim that there are none. Secondly, the benefit of such structures is not perceived as important enough to be worth the expense. One third of researchers who know about support structures do not use them despite having the opportunity to do so. This, in turn, might be a result of either insufficient support structures (may it be in terms of offered service, format or content) or lack of knowledge about how and why such structures could improve the interviewees RDM. The survey also asked for an evaluation of the offered support structures with the results being shown in figure 15.

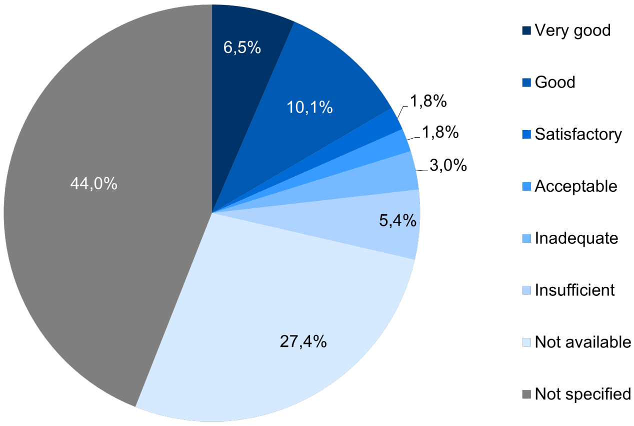

Figure 15: Evaluation of the offered support structures

Combining the data basis from figure 15 with figure 14, there are several groups of researchers to be identified, clustered by their access to and their usage of RDM support, shown in table 7.

Group of researchers who have...	Respondents [%]
... access to RDM support and use it.	11.3
... access to RDM support and do not use it.	17.9
... no access to RDM support, but would like to use it.	23.8
... no access to RDM support and do not criticise its absence.	23.8
... not specified it.	27.6

Table 7: Groups of researchers clustered by their access to and their usage of RDM support structures

4.7 Further Open Questions

In further open questions, respondents were given the opportunity to mention possible reasons that might prevent researchers from RDM in the form of free text answers. Most interesting are the answers on the question ”What reasons could prevent researchers from sustainable RDM?”, which 39 of the 168 interviewees (23%) answered. A detailed list of quotes of the respondents can be found in the Appendix. The effort or workload for the establishment and operation of RDM is with 16 mentions the most recognisable reason against proper RDM usage. Likewise, the lack of clear standards or guidelines for RDM is cited twelve times, closely followed by the lack of awareness of RDM among researchers (nine mentions). This last statement is specified: RDM is primarily perceived as an additional expense, there is no incentive to use it and no necessity for RDM is seen. The lack of necessity is justified by the time-limited nature of projects and their isolation in research environments. Other reasons against RDM application are a lack of knowledge (seven mentions), the concern of data misuse or data usage without permission or citation (six mentions) and problems with missing or complicated support structures, which five interviewees mentioned.

The feeling that the own data can only be used for the own projects prevails for many. Contrarily, others who consider their data to be usable, fear data misuse. In this case the protection of the own research is seen as more important than a provision of data within the framework of RDM. This is expressed, for example, in the following quote from one of the respondents:

> ”Real data, e.g. from production, is not easy to obtain. Those who have such data sets have an advantage. Therefore, data is not shared, although it would be useful to promote scientific progress and test results for reproducibility.”

Many of the interviewees’ statements can be condensed into the following statement (adapted in wording for the purpose of anonymisation), which was formulated by an interviewee:

> “ Besides the most obvious reason - lack of knowledge - I think [RDM] just meets [ignorance] by and large. One Example: For [research] I have collected publicly available data. Of course I maintain and cherish my data and go through large parts of the data life cycle, but for that I don’t need thousands of tools that nobody else [in my organisation] uses. It is also likely that others will not (be able to) continue to use this data - which is why it makes sense to maintain it sustainably. It is similar with research projects. The more isolated and smaller the project is, the less sense there really is in elaborate management [...]. This is not only true for the data. Furthermore, it is unfortunately inherent in the research system that I could suffer great professional damage if I give out my data beyond a certain level. In applied research projects the situation is certainly different, but even here I need (at least initially) a more or less exclusive use of data so that I can firstly secure my livelihood. Furthermore, there are often confidentiality clauses that do not allow me to pass on the data. ”

The free-text answers allow the following conclusions to be drawn:

5. *The interviewees see the effort of RDM in terms of initialisation, familiarisation with it and everyday work as a reason that prevents researchers from sustainable RDM.*
6. *The interviewees name the lack of clear guidance through the RDM process like guidelines, standards or processes as a reason that prevents researchers from sustainable RDM.*
7. *The interviewees perceive that RDM as a topic does not receive enough awareness yet, which is a reason that prevents researchers from sustainable RDM.*
8. *The interviewees see a lack of knowledge among themselves and other researchers, which is a reason that prevents researchers from sustainable RDM.*
9. *The interviewees consider the risk of data misuse and data usage without citation or permission as a reason that prevents researchers from sustainable RDM.*
10. *The interviewees see the lack or quality of support structures as a reason that prevents researchers from sustainable RDM.*

The acceptance of the reuse of data among the respondents is limited. Thus, the ”not-invented-here syndrome” [34] is cited by the respondents. This effect describes the rejection of ideas and inventions not founded in one’s own institution for reasons other than monetary ones. For example, openly available data might not be reused because it is not trusted as it is of other origin as the own institution. As a result, the subsequent use of existing data is omitted and additional work is done, since data must be collected by the institution itself [34].

5 Discussion

Within this paper ten conclusions could be drawn, derived from the data of the survey results. While these ten hypotheses do only provide a qualitative approach to the topic of RDM usage and application, the survey still provided conclusions regarding main issues in the context of RDM and opened the possibility to derive potential measures. The knowledge, awareness and usage of RDM has to be fostered to enhance the management and therefore FAIRness [10] of research data. To achieve this, researchers firstly need to know what to do when starting managing research data (see hypotheses 4., 5. & 8.). An appropriate approach needs to be handed to them with a clear entry point and a structured and adaptable process needs to be defined (see hypothesis 6.). When questions occur, those have to be answered right away (see hypotheses 5. & 10.). Also, training materials to the very topic of the question have to be provided and suitable tools have to be introduced (see hypotheses 1. & 4.). Those materials should be light-weight and focused on applicability. Light-weight in this context means that provided information should only focus on the very specific problem of the researcher. A huge amount of additional and unapplicable instructions will compromise the will of researchers to use RDM and cause frustration. The process of RDM has to be embedded within everyday research (see hypothesis 3).

Incentivation for RDM usage needs to be provided as the requirements of, for example the DFG, are not sufficient to enhance the application of RDM (see hypothesis 2.). Also, the awareness for RDM has to be broadened (see hypothesis 7.). Suitable measures could be the requirements of RDM in connection with dissertations or bachelor/master theses.

Opposing to the incentivation is the fear of data misuse or missing citations of the own work (see hypothesis 9.). This could be addressed by the possibility of storing data in closed repositories and clear instructions of how data can be made publicly available in a way that it is unambiguously recognisable who the author is and to whom the data belongs. Access management and licensing has therefore to be taken into consideration, granting the possibility of a controlled reuse of data.

To conclude this paper, a comparison of the hypotheses to the findings of the literature review shall be given, orderd by the number of hypotheses listed above. This comparison is drawn to different disciplines and countries than the scope of this survey. Yet there are some similarities and common challenges that form a reoccurring pattern in the nature of RDM.

For instance, hypothesis 1 is supported by several papers. The ”lack of trainers in RDM practices” [12], ”lack of knowledge/training” [24], a lack of ”data sharing skills” [22], or the need of training as stated by Elsayed and Saleh [20] is represented in many papers. The only contradiction found in literature by Costanzo et al. states that ”Lack of RDM Knowledge [is a] low barrier” [13].

Costanzo and Cooper support hypothesis 2, describing the ”lack of institutional understanding and awareness of [...] expectations” [13]. Wilms et al. state, a ”requirement to comply with possible guidelines” [15] is not enough incentive for researchers to adhere to good RDM practices.

The third hypothesis is not supported by any findings in the literature. Therefore, this hypothesis could benefit from a revision in the future. However, Palsdottir states that RDM ”is not a normal practice” in the researchers work [17]. Still, the reasons for the usage of tools should be clarified. The hypothesis can not be supported by literature but is still a finding of this paper.

While Björnmalm et al. see the problem in too many generic and yet too few specific RDM tools [12], Israel et al. state that ”respondents continue to rely on [...] paper laboratory notebooks” [18] instead of electronic laboratory notebooks. While there are many tools available for RDM activities both generic and specific [18], the ”lack of knowledge” [24] about these tools can be seen as the actual challenge RDM is facing in this context. This also supports hypothesis 4.

Hypothesis 5 is also represented within the literature. RDM is seen as ”a significant burden” [14] as ”the amount of time it takes” [24] is a ”perceived increased workload” [15] connected to RDM, opposing a ”lack of resources (time, budget, personnel etc.)” [13].

Connected to the effort required for RDM, the lack of guidance (hypothesis 6) is found both in the answers of this survey as well as the literature. Björnmalm et al. found a lack of ”specific instructions (or links to relevant guidelines)” [12], which is supported by Costanzo et al. regarding the ”lack of institutional understanding and awareness of [...] expectations” [13] as well as the findings of Borghi and Van Gulick that there is missing guidance through ”lack of best practices” [24]. The ”large number of tools and methods” [18] and ”complexity in data structures [,] formats [and] documentation” [18] is a challenge yet to be faced. As ”processes are not yet clearly defined, let alone standardised” [23] ”researchers needed assistance” [23] in RDM, which is also supported by [17]. Additionally, ”establishing [...] guidelines” can improve RDM [19].

Many papers also address hypothesis 7, however some support it while others oppose it. While Björnmalm et al. see ”too few incentives for researchers that reward and incentivise implementation of RDM practices into everyday workflow” [12], Wilms et al. see that the ”overall acceptance of RDM policies is low” [15]. According to Austin et al. there is a ”need to demonstrate to researchers the value of data management” [14]. Simmilarly, Borghi and Van Gulick point out that the importance of RDM is not commonly known [24]. These four statements support hypothesis 7. Israel et al. point out that ”making data FAIR needs to start most importantly, awareness” [18], also supporting hypothesis 4 to some extend. However, Vilar and Zabukovec oppose these theories, stating that researchers are rather convinced by RDM [16]. Ortloff et al. also argue in their spotlight investigation that ”most of the partners are strongly aware of the benefits provided” [19] by RDM. The incentivation of RDM, as for example brought up by Borghi and Van Gulick, has to be addressed by funding organisations, universities and institutions. However, it is not part of this paper, as the focus lies on the researchers perspective on RDM. Still, the topic of incentives has to be considered from all sides, from making funding dependent on concrete RDM practices to the demanded RDM in the context of a dissertation.

While hypothesis 8 is not directly supported or opposed by the literature, it is to some extend a consequence from hypotheses 1 and 4. Palsdottir states the ”limited knowledge” and that RDM ”is not a normal practice” as well as an ”urgent need to increase the researcher’s knowledge and understanding of the importance of data managenent" [17]. However, it can neither be contradicted nor be proven that the lack of knowledge hinders the application of RDM. The lack of knowledge has been stated several times, both in this survey and the literature. A plausible outcome might be the hindering of (sustainable) RDM.

The ninth hypothesis is addressed by five papers. Austin et al. state that more than half of the involved partners in the projects rejected data sharing [14]. This is mostly based on the "concerns regarding IP protection" [19] respectively "intellectual property rights" [21] and the "fear of losing control" [15]. The "partner's consent for publication was the biggest hurdle" [23].

Lastly, hypothesis 10 is supported by some papers. Elsayed and Saleh see a need for support [20] as well as [17], while Björnmalm et al. see a lack of "support at a faculty level" [12], similar to the "lack of availability of support materials" [13] stated by Costanzo et al. Wuchner et al. also see a need for support, but on a more immediate level. While the aforementioned papers focus on generic support, Wuchner et al. see a direct assistance needed for "data publications – especially FAIR ones [because they are] are a major challenge for researchers" [23]. This last statement excluded, all papers revolve around the lack of support, which is partially true, but might also be a consequence of the lack of knowledge and awareness, as stated in hypotheses 1, 4 and 8.

6 Summary and Outlook

This paper has shown the results of a survey that took place from October to December 2020. 168 researchers were interviewed and the results were derived from their answers to the 216 questions within the survey. Main topics of the survey as well as (sub)sections within this paper were "RDM Knowledge and Perceived Relevance of RDM", "Application of RDM Related Tasks", "Data Sharing with Third Parties", "Usage of RDM Tools and Services", "Usage of File Formats", "Specifications and Support Structures" and responses to "Further Open Questions".

A key finding is the need of researchers in engineering sciences for guidance and support regarding RDM in their everyday research. This results from the main reasons against RDM, namely missing knowledge about guidelines, tools and support in RDM as well as the additional effort connected. Guidance should be provided in form of use case related processes that integrate into everyday research and support researchers with knowledge and tool support when needed.

Future research could further elaborate on RDM requirements of researchers, integration of RDM into everyday research, general feasibility and practices resulting. The applicability and usability of RDM should be fostered to facilitate the needed cultural change in engineering sciences.

Additionally, the authors would like to point out that a complete statistical analysis of the linked data could result in further findings. The linked data is specifically intended to be reused.

7 Appendix

Interviewees were asked ”What reasons could prevent researchers from sustainable research data management?”. Their answers on this questions can be found below. The statements are split up into the following categories:

- Effort
- Guidelines and Standards
- General Acceptance, Discipline and Awareness of RDM
- RDM Knowledge
- Data Misuse and Permissions
- Support Structures
- Longer Statements

Some statements contained content that would fit into multiple of these categories. Such statements were split into two or more parts and listed in the corresponding category if the meaning was untouched by the split. If a concrete distinction between two parts cannot be made within one statement, the quote will be listed in multiple categories.

7.1 Effort

One of the main concerns of the interviewed researchers is the effort connected to RDM. 16 of the 39 free-text answers mentioned the effort or time expenditure as a reason to not manage research data.

- *”Time-limited projects that one works on alone. Sustainable and systematic data storage usually **only additional effort.**”*
- *”**Time required** for upkeep”*
- *”**Much too elaborate**, no predefined structures. Clear specifications must be applicable and clear”*
- *”Time expenditure”*
- *”Effort”*
- *”**Effort during set-up**”*
- *”Lack of time”*
- *”Effort and time”*
- *”**Additional effort is considered too high - regardless of the desire for implementation. Familiarisation with formats is too time-consuming, as step-by-step introduction along the daily work routine is not available.**”*
- *”Too much effort”*
- *”**High organisational and training costs** with low capacities”*

- *"**Too complicated**, no infrastructure, no advice, no support, importance is not rewarded"*
- *"**Increased documentation effort**, restrictions in the use of file formats and systems for data storage"*
- *"lack of processes - lack of contact persons - **time expenditure** / "inertia" –> initially no direct benefit for the person who has to do RDM - lack of IT infrastructure - lack of know-how regarding data migration, data security, data representation, etc."*
- *"**Sustainable RDM takes time** and goes beyond use in own promotion - joint effort needed."*
- *"Ignorance and carelessness, **additional effort if there are no clear rules** from the beginning"*
- *"Extensive/varied software to support - lack of standardisation? - Lack of knowledge? - **High effort in the life cycle (pre-planning, ..., archiving)**"*

7.2 Guidelines and Standards

The following twelve quotes make statements about guidelines and standards not being sufficient or too ambiguous.

- *"Lack of awareness, **no existing or communicated guidelines**"*
- *"**Ambiguities** in the specifications"*
- *"Ignorance and carelessness, **additional effort if there are no clear rules** from the beginning"*
- *"Much too elaborate, **no predefined structures. Clear specifications must be applicable and clear**"*
- *"The **lack of time** to deal with new formats/tools and to carry out extensive data preparation."*
- *"Missing or unclear specifications."*
- *"Researchers are **not aware of what proper research data management should look like.**"*
- *"**No information culture** regarding RDM exists. Framework conditions are completely unknown"*
- *"Lack of knowledge. Non-existent guidelines in the organisation"*
- *"Too complicated, no infrastructure, no advice, no support, importance is not rewarded"*
- *"**lack of processes** - lack of contact persons - time expenditure / "inertia" –> initially no direct benefit for the person who has to do RDM - lack of IT infrastructure - lack of know-how regarding data migration, data security, data representation, etc."*
- *"Extensive/varied software to support - **lack of standardisation?** - Lack of knowledge? - High effort in the life cycle (pre-planning, ..., archiving)"*

7.3 General Acceptance, Discipline and Awareness of RDM

Nine researchers referred to general acceptance of RDM as well as discipline and awareness issues.

- ”*Own evaluations paired with expertise”*
- ”***Lack of awareness. Silo thinking****”*
- ”***No sense of necessity****”*
- ”***Negligence, workload, ignorance,*** *too much variety of options”*
- ”***Benefits not always easily recognisable*** *for others”*
- ”*Meaning-making. Knowledge of the tools”*
- ”***No more recognisable added value in relation to the effort involved*** *in familiarisation when it also works with self-structured Excel files.”*
- ”*In my opinion, it is much more important that the generated data can also be reproduced by third parties. Therefore, for me, providing the code in conjunction with a sandbox environment is much more important than the data itself.”*
- ”*Agreement on duration of employment/project duration. A large part of the data is only generated towards the end of the project duration/employment contract period, as the experimental facilities must first be set up and put into operation. And: Lack of state positions/permanent positions and high additional workload due to teaching/relocation”*

7.4 RDM Knowledge

Seven quotes addressing RDM knowledge issues are listed below.

- ”***Too little own expertise*** *and* ***too much effort*** *for familiarisation.* ***Offers and tools not sufficiently known****. Especially the technological progress: Often standard software from 10 years ago no longer runs on new operating systems, media for persistent storage lose their functionality in the medium term, necessary software and the knowledge to use this software could no longer be available after a few years.”*
- ”*There are many tools but* ***too little experience to choose the appropriate*** *ones.”*
- ”***Excessive number of tools. No clear place to save.****”*
- ”*No information culture regarding RDM exists. Framework conditions are completely unknown”*
- ”*Lack of knowledge. Non-existent guidelines in the organisation”*
- ”*Extensive/varied software to support - lack of standardisation? - Lack of knowledge? - High effort in the life cycle (pre-planning, ..., archiving)”*
- ”*lack of processes - lack of contact persons - time expenditure / ”inertia” –> initially no direct benefit for the person who has to do RDM - lack of IT infrastructure -* ***lack of know-how regarding data migration, data security, data representation, etc.****”*

7.5 Data Misuse and Permissions

Another concern of researchers is the fear of data misuse or data usage without permission or citation, mentioned six times.

- *"**Protection of own research**, as not everything has been published yet"*
- *"Fear of **data misuse (publication without naming the source or similar)**"*
- *"Fear for **data sovereignty**"*
- *"Data loss, violation of DFG rules"*
- *"Fear that third parties could overtake you in your own research. Worry that one's own data has not been collected or analysed cleanly enough. (But hey, others only boil with water, too)"*
- *"Real data, e.g. from production, is not easy to obtain. Those who have such data sets have an advantage. Therefore, data is not shared, although it would make sense to do so in order to promote scientific progress and check results for reproducibility."*

7.6 Support Structures

Last but not least, five of the quotes contain comments on support structures etc. and what reasons against RDM are connected to those.

- *"There is little support [at my institute]. **Training and education on tools and possibilities** would be particularly useful, as would an institute-wide standard. Solutions for individual projects are currently failing due to the IT department and the administration. (Topic licences, accesses, installations)"*
- *"Much too elaborate, **no predefined structures. Clear guidelines must be applicable and clear**"*
- *"**Non-existent or impractical to use infrastructure.**"*
- *"Too complicated, no infrastructure, no advice, **no support**, importance is not rewarded"*
- *"lack of processes - lack of contact persons - time expenditure / "inertia" –> initially no direct benefit for the person who has to do RDM - **lack of IT infrastructure** - lack of know-how regarding data migration, data security, data representation, etc."*

7.7 Longer Statements

As wrap up, two rather long statements that address multiple of the topics listed above may be cited:

> *"Lack of tool support. **Unclear what "research data" comprises.** The DFG definition is very broad and thus not very clear. Classically, it was measurement and observation data, interview data and the like. In the meantime - and this is also well reflected in some of the questions in this survey - the term encompasses practically every piece of information that a researcher comes across in his or her life. But this is difficult because everyone (if one takes the principle of assignability of ideas*

> *strictly seriously) would have to keep a complete documentation of all conversations, impressions, experiences in the professional and private environment because it cannot be ruled out that a remark made by a third party during small talk, remembered by chance weeks later, provides the decisive push to get ahead with a problem in a completely different context. Lack of awareness - It is now common knowledge that primary data must be kept secure. What primary data is is more of a question, especially in disciplines that are more constructive and less observational/measuring. Not only in data management, but also there: "Not invented here" syndrome (especially in software-heavy projects a widespread nuisance, partly forced by too tight copyright / too tight patent protection)."*

> *"Apart from the most obvious reason - lack of knowledge - I believe that it simply encounters a lot of irrelevance in various fields on the whole. Ex: I collected publicly available data for my dissertation. Of course I maintain and care for my data and go through large parts of the data life cycle, but for that I don't need thousands of tools that no one else at the [institute] uses. Also, others will probably not (be able to) continue to use this data - this also results in the meaninglessness of sustainable maintenance. It is similar to research projects. The more isolated and smaller the project, the less sense there really is in complex management around it. This does not only apply to the data. Moreover, it is unfortunately inherent in the research system that I could suffer great professional damage if I give out my data beyond a certain level. In applied research projects the situation is certainly different, but here, too, I need (at least initially) a more or less exclusive use of data so that I can initially secure my livelihood. Furthermore, there are often confidentiality clauses that do not allow me to pass on the data."*

8 Acknowledgements

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713 [32].

9 Roles and contributions

Tobias Hamann: Conceptualization and Methodology of the survey evaluation, Writing

Amelie Metzmacher: Conceptualization, Methodology and Execution of the survey

Patrick Mund: Conceptualization and Methodology of the survey

Marcos Alexandre Galdino: Writing - Review

Anas Abdelrazeq: Writing - Review

Robert Schmitt: Idea, Supervision, Funding acquisition

References

[1] Statista IDC. "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes)." (2021), [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/ (visited on 03/15/2023).

[2] G. Bell, T. Hey, and A. Szalay, "Computer science. Beyond the data deluge," eng, *Science (New York, N.Y.)*, vol. 323, no. 5919, pp. 1297–1298, 2009, Journal Article. DOI: 10.1126/science.1170411. eprint: 19265007.

[3] T. Hey and A. Trefethen, *The Data Deluge: An e-Science Perspective*, Mar. 11, 2003. DOI: 10.1002/0470867167.ch36.

[4] D. Williams and H. Tang, *Data quality management for industry 4.0: A survey*. 2020. [Online]. Available: https://search.proquest.com/openview/f7aaf15f64ccb032852e4edbd1ad16c9/1?pq-origsite=gscholar&cbl=25782 (visited on 03/15/2023).

[5] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data - ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2021, ISSN: 1041-4347. DOI: 10.1109/TKDE.2019.2946162.

[6] J.-M. Rodriguez, "Aktuelle Softwareplattformen für Forschungsdatenrepositorien auf dem Prüfstand," *ABI Technik*, vol. 39, no. 1, pp. 23–25, 2019, ISSN: 0720-6763. DOI: 10.1515/abitech-2019-1004.

[7] E. Böker. "Was ist Forschungsdatenmanagement? | Informieren und Planen | Themen | Forschungsdaten und Forschungsdatenmanagement." de. (2023), [Online]. Available: https://forschungsdaten.info/themen/informieren-und-planen/was-ist-forschungsdatenmanagement/ (visited on 02/15/2023).

[8] forschungsdaten.info. "Glossar | Praxis kompakt | Forschungsdaten und Forschungsdatenmanagement." de. (2023), [Online]. Available: https://forschungsdaten.info/praxis-kompakt/glossar/#c269821 (visited on 02/15/2023).

[9] R. H. Schmitt, V. Anthofer, S. Auer, *et al.*, *Nfdi4ing - the national research data infrastructure for engineering sciences*, 2020. DOI: 10.5281/ZENODO.4015201.

[10] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," eng, *Scientific Data*, vol. 3, no. 1, p. 160 018, 2016, Journal Article, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. eprint: 26978244.

[11] T. Todorova, M. Garvanova, M. Peteva, and V. Avramova, "Comparative findings from data literacy survey in three bulgarian universities," in *EDULEARN19 Proceedings*, L. Gómez Chova, A. López Martínez, and I. Candel Torres, Eds., ser. EDULEARN Proceedings, IATED, 2019, pp. 932–940. DOI: 10.21125/edulearn.2019.0304.

[12] Mattias Björnmalm, Federica Cappelluti, Alastair Dunning, *et al.*, *Advancing research data management in universities of science and technology*, 2020. DOI: 10.5281/ZENODO.3665372.

[13] L. Costanzo and A. Cooper, *Setting the foundations for stronger partnerships and collaborations for developing institutional rdm strategies in canada*, 2023. DOI: 10.5281/ZENODO.8010869.

[14] T. Austin, K. Bei, T. Efthymiadis, and E. Koumoulos, "Lessons learnt from engineering science projects participating in the horizon 2020 open research data pilot," *Data*, vol. 6, no. 9, p. 96, 2021. DOI: 10.3390/data6090096.

[15] K. L. Wilms, S. Stieglitz, B. Ross, and C. Meske, "A value-based perspective on supporting and hindering factors for research data management," *International Journal of Information Management*, vol. 54, p. 102 174, 2020, ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2020.102174.

[16] P. Vilar and V. Zabukovec, "Research data management and research data literacy in slovenian science," *Journal of Documentation*, vol. 75, no. 1, pp. 24–43, 2019, ISSN: 0022-0418. DOI: 10.1108/JD-03-2018-0042.

[17] A. Palsdottir, "Data literacy and management of research data – a prerequisite for the sharing of research data," *Aslib Journal of Information Management*, vol. 73, no. 2, pp. 322–341, 2021, ISSN: 2050-3806. DOI: 10.1108/AJIM-04-2020-0110.

[18] H. Israel and M. M. Becker, "What does data stewardship mean in physics?," 2023. DOI: 10.17192/bfdm.2023.2.8570.

[19] D. Ortloff, S. Anger, and M. Schellenberger, "An empirical study of the state of research data management in the semiconductor manufacturing industry: An analysis of industry and research institutions in the idev4.0 project: Ing.grid preprint," vol. 2023, 2023. [Online]. Available: https://preprints.inggrid.org/repository/view/19/ (visited on 10/20/2023).

[20] A. M. Elsayed and E. I. Saleh, "Research data management and sharing among researchers in arab universities: An exploratory study," *IFLA Journal*, vol. 44, no. 4, pp. 281–299, 2018, ISSN: 0340-0352. DOI: 10.1177/0340035218785196.

[21] M. Cheung, *Data culture in canada: Perceptions and practice across the disciplines*, 2021. DOI: 10.5281/ZENODO.6760995.

[22] W. D. Chawinga and S. Zinn, "Global perspectives of research data sharing: A systematic literature review," *Library & Information Science Research*, vol. 41, no. 2, pp. 109–122, 2019, ISSN: 07408188. DOI: 10.1016/j.lisr.2019.04.004.

[23] A. Wuchner, M. Robrecht, P. Kehl, and H. R. Schmitt, "Challenges in publishing research data – a fraunhofer case study: Ing.grid preprint," 2023. [Online]. Available: https://preprints.inggrid.org/repository/view/20/ (visited on 10/31/2023).

[24] J. A. Borghi and A. E. van Gulick, "Data management and sharing in neuroimaging: Practices and perceptions of mri researchers," *PloS one*, vol. 13, no. 7, e0200562, 2018. DOI: 10.1371/journal.pone.0200562.

[25] WGP. "Über uns - wgp." (2022), [Online]. Available: https://wgp.de/de/ueber-uns/ (visited on 01/02/2023).

[26] WiGeP. "Über uns – WiGeP – Wissenschaftliche Gesellschaft für Produktentwicklung." (2022), [Online]. Available: https://wigep.de/wigep/#wigep (visited on 01/02/2023).

[27] Fraunhofer. “Fraunhofer-Verbund Produktion.” Copyright: Copyright. (2022), [Online]. Available: https://www.produktion.fraunhofer.de/ (visited on 01/02/2023).

[28] E. Internet of Production. “”The World becomes a Lab” - die Vision des IoP - RWTH AACHEN UNIVERSITY Exzellenzcluster Internet of Production - Deutsch.” (2022), [Online]. Available: https://www.iop.rwth-aachen.de/cms/Produktionstechnik/Das-Exzellenzcluster/~rgqb/Leitbild/ (visited on 01/02/2023).

[29] forschungsdaten.info. “Datenlebenszyklus | Informieren und Planen | Themen | Forschungsdaten und Forschungsdatenmanagement.” de. (2022), [Online]. Available: https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/ (visited on 11/29/2022).

[30] Deutsche Forschungsgemeinschaft, “Guidelines for Safeguarding Good Research Practice. Code of Conduct,” de, 2022. DOI: 10.5281/ZENODO.6472827.

[31] L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, and G. Tutz, *Statistik, Der Weg zur Datenanalyse* (Springer-Lehrbuch), ger, 8., überarbeitete und ergänzte Auflage. Berlin and Heidelberg: Springer Spektrum, 2016, 581 pp., Fahrmeir, Ludwig (VerfasserIn) Heumann, Christian (VerfasserIn) Künstler, Rita (VerfasserIn) Pigeot, Iris (VerfasserIn) Tutz, Gerhard (VerfasserIn), ISBN: 978-3-662-50371-3. DOI: 10.1007/978-3-662-50372-0.

[32] DFG. “Dfg - gepris - nfdi4ing – national research data infrastructure for engineering services.” (2022), [Online]. Available: https://gepris.dfg.de/gepris/projekt/442146713?language=en (visited on 11/29/2022).

[33] fileinfo.com. “Spss portable file.” (2011), [Online]. Available: https://fileinfo.com/extension/por (visited on 01/19/2023).

[34] H. Mehrwald, *Das “Not Invented Here”-Syndrom in Forschung und Entwicklung*. Wiesbaden: Deutscher Universitätsverlag, 1999, ISBN: 978-3-8244-0483-4. DOI: 10.1007/978-3-663-08337-5.