IJDC | *Research Paper*

Assessing Quality Variations in Early Career Researchers' Data Management Plans

Jukka Rantasaari

University of Turku, and Åbo Akademi University

Abstract

By evaluating the real-world data management plans DMPs crafted in a multi-stakeholder RDM course, this study aims to improve understanding of quality variations in early career researchers' (ECRs) DMPs, and to identify gaps in their research data management (RDM) planning skills. We also examine the differences between DMPs in relation to several background variables (e.g., discipline, course track). The Basics of Research Data Management (BRDM) course has been held in two multi-faculty, research-intensive universities in Finland since 2020. In this study, 223 ECRs' DMPs created in the BRDM of 2020 - 2022 were assessed, using the recommendations and criteria of the Finnish DMP Evaluation Guide + General Finnish DMP Guidance (FDEG). The median quality of DMPs appeared to be satisfactory. The differences in rating according to FDEG's three-point performance criteria were statistically insignificant between DMPs developed in separate years, course tracks or disciplines. However, content analysis revealed differences in RDM best practices, such as sharing, storing, and preserving data, between disciplines or course tracks. DMPs that contained a structured data table (DtDMP) also differed highly significantly from prose DMPs. DtDMPs better acknowledged the data handling needs of different data types and improved the overall quality of a DMP. Nevertheless, more focused, further training to achieve the excellent quality is needed, especially in areas of handling personal data, legal issues, archiving, and funders' data policies.

The study provides RDM stakeholders – including researchers, institutions, funders, and publishers – with a standardized framework for the development and evaluation of DMPs. Researchers benefit from enhanced data management descriptions, boosting the integrity and reproducibility of research. Institutions can better identify DMP strengths and areas for improvement, allowing for customized support and training. Educators can leverage this framework to gauge the effectiveness of RDM training. Funders and publishers can set clear DMP standards, promoting transparency, compliance, and data sharing efficiency.

Submitted 3 March 2023 ~ *Revision received* 12 January 2024 ~ *Accepted* 6 February 2024

Correspondence should be addressed to Jukka Rantasaari: jukka.rantasaari@utu.fi

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v18i1.873
DOI: 10.2218/ijdc.v17i1.873

Introduction

If a researcher meticulously organizes, stores, processes, and preserves data, and links it to the accompanying research paper, other researchers and users can more easily understand, verify, and reuse the data. Consequently, sound RDM practices may advance data integrity, the reliability of research findings, and research reproducibility (e.g., Chiarelli et al., 2021).

For these reasons, an increasing number of funders, publishers, and policy makers started recommending or mandating researchers to write DMPs and share data during the 2010s (e.g., Academy of Finland, 2019; "Amsterdam call for action on open science," 2016; European Commission, 2018a; 2018b; European University Association, 2017; National Science Foundation, 2011; UNIFI, 2016; Wellcome, 2017). The main objective of these recommendations and mandates is to ensure that research data funded by the public is accessible and reusable whenever possible. To achieve this goal, it is important that research data, or at the very least the metadata, adhere to the FAIR principles (Findable, Accessible, Interoperable, and Reusable). [1]However, to fulfill these objectives, researchers need education, guidance, and support (e.g., European Commission, 2022).

Identifying the RDM skill gaps is crucial for determining where researchers need education and support to align with FAIR data objectives. The author has previously analyzed the outcomes of BRDM courses from 2019 to 2021, focusing on Early Career Researchers' (ECRs') self-assessments and open feedback on their RDM competencies (Author, 2022). The current study builds on that foundation by analyzing 223 DMPs formulated by ECRs during BRDM courses between 2020 and 2022, which are derived from each participant's ongoing research project. The aim is to improve understanding of the quality variations of ECRs' DMPs and to identify gaps in their RDM planning skills by evaluating the comprehensiveness and quality of the DMPs. To this end, the research questions addressed are:

1. RQ-1: What is the quality of the ECRs' DMPs, as evaluated against the Finnish DMP Evaluation Guidance (FDEG) criteria?

2. RQ-2: How do the DMPs align with the RDM best practices recommended by the FDEG and BRDM?

3. RQ-3 What differences can be found between DMPs regarding the year of the training, discipline, course track, and other specified background variables?

This study sets forth to examine ECRs' DMPs with a twofold purpose: to evaluate their quality and to unearth ECRs' educational needs in RDM. The subsequent section will delve into the DMP content analysis literature to provide a foundation for the empirical analysis that follows.

Literature Review

Earlier findings about the significance of a DMP for a researcher are described first in this section. Probable reasons for the current DMPs' low quality and steering impact on research will also be illuminated using recent review studies (Hudson-Vitale & Moulaison-Sandy; 2019; Smale et al., 2020). Finally, we will examine the results of the DMP content analyses, which are surprisingly few and not quite recent, focusing mainly on data sharing, description of data types, storage, and preservation of data.

[1] https://www.go-fair.org/fair-principles/

The Significance of a DMP for Research Practices

According to Mannheimer (2018), a DMP does not steer researchers' data management actions but is more like a representation of their current RDM practices. However, a DMP also supports considering and reflecting on the current practices and hence makes it possible to improve them (Mannheimer, 2018). Diekema et al.'s (2014) study found that researchers' views were that funders' mandates have not had an impact on their practices; that was the case, the authors said, because most of them already stored and shared their data according to mandates. Nevertheless, the RDM practices that researchers think are important are not a guarantee that they will be implemented (European Commission, 2022; Parham et al., 2016; Scaramozzino et al., 2012).

Based on the finding of Hudson-Vitale and Moulaison-Sandy's (2019) review study, current DMPs seem inefficient, the main reason being the lack of clear requirements and assessment criteria for sound RDM. Similarly, according to Smale et al.'s (2020) extensive literature review, the varied requirements of multiple stakeholders, such as an individual researcher, a funding agency, an institution, and a discipline, are not clearly delineated, which produces a DMP that does not properly meet any stakeholders' demands. Moreover, Kvale and Pharo (2021) found that too high a degree of standardization and automatization can make DMP forms inflexible and lose the autonomy of research projects and their differing needs. Nevertheless, the adoption of universal, discipline-agnostic terms and definitions is essential to enhance RDM practices across all fields. While RDM operations are fundamentally generic, their implementation may differ based on discipline, research methods, and data types (Parham et al., 2016; Coates, 2014; Lefebvre et al., 2018; Molloy and Snow, 2012; Scholtens et al., 2019; Strasser et al., 2012; Weller and Monroe-Gulick, 2014).

Other reasons for the low impact of a DMP on research practices, besides lack of detailed criteria for meeting multiple stakeholder requirements, can be lack of time (Bardyn et al., 2012; Tenopir, 2011), lack of information on sound RDM practices (Bishoff & Johnston, 2015; Mannheimer, 2018; Nicholls et al., 2014), lack of infrastructures (Berman, 2017; Diekema et al., 2014), and lack of incentives (Bierer et al., 2017; Ioannidis, 2016; Pierce et al., 2019). Furthermore, a DMP's quality has not been found to have an impact on funding decisions (Berman, 2017; Mannheimer, 2018; Mischo et al., 2014).

We will next describe the findings from DMP content analysis studies, with a focus on the descriptions of data sharing, data types, metadata, and storage and preservation strategies. Content analyses, though few and primarily from the 2010s (2012 to 2020), have mostly focused on DMPs from NSF grant proposals, as summarized in Table 1 in the Appendix. These studies typically report vague descriptions of RDM actions, often noting incomplete or absent details regarding data types, metadata, intellectual property rights (IPR), ethical considerations, data sharing methods and venues, timelines and locations for preservation, as well as defined roles and responsibilities in RDM (Nicholls et al., 2014; Parham et al., 2016; Samuel et al., 2015; Smale et al., 2020; Van Loon et al., 2017).

Data Sharing in DMP content analyses

While there is no single definition of "data sharing" in the research literature (Thoegersen & Borlund, 2022), Curty et al. (2013), Van Loon et al. (2017), and Mannheimer (2018) categorize "formal data sharing methods" as the use of repositories, file sharing services, and supplements, and "informal methods" as data shared within publications, upon request, or via web pages. In most content analyses informal data sharing methods were highlighted (Berman, 2017; Bishoff & Johnston, 2015; Curty et al., 2013; Mannheimer, 2018; Mischo et al, 2014; Van Loon et al., 2017). Yet, these methods are not necessarily adequate, reliable, or open enough to produce findable, accessible, interoperable, and reusable (FAIR) data: Publications may be located behind paywall, and they usually do not contain all necessary data needed for verification or replication of the study. Sharing data through email upon request is found to be uncertain and decline rapidly with article age (Savage & Vickers, 2009; Vines et al, 2014; Thessen et al, 2014).

Sharing through websites with non-permanent, often outdated addresses is not trustworthy either. Most of the DMPs' authors intended to share data, although the terms for data use and reuse were missing from 56 to 81% of the DMPs (Mannheimer, 2018; Parham et al., 2016; Van Loon et al., 2017).

Data Types and Metadata in DMP Content Analyses

Content analyses reveal varying levels of completeness in the description of captured, produced, or reused data types in DMPs. For instance, Mannheimer (2018;), Nicholls et al. (2014), and Parham et al. (2016) noted that data types were often well-described, yet Van Loon et al.'s (2017) study found that descriptions were complete in only 42% of DMPs, incomplete in 39%, and not addressed at all in 19%.

Additionally, metadata standards or detailed metadata descriptions were absent in a significant portion of DMPs, missing in 59 to 81% of cases across several studies (Mannheimer, 2018; Parham et al., 2016; Samuel et al., 2015; Van Loon et al., 2017).

Storing and Preserving in DMP Content Analyses

Mannheimer's (2018) study stands out, with 59% of DMPs providing complete descriptions of archiving solutions. However, other content analyses have often found the descriptions of storage and preservation, including place and time frame, to be vaguely articulated, frequently conflating these with solutions for data storing and sharing (Bishoff & Johnston, 2015; Mischo et al., 2014; Nicholls et al., 2014; Parham & Doty, 2012; Parham et al., 2016; Van Loon et al., 2017). Specifically, Van Loon et al. reported a missing time frame for preservation in 57% of DMPs, while Nicholls et al. (noted its absence in 30% of DMPs. Parham et al. (2016) observed that 43% of DMPs planned for data to be archived in a repository or data center, 28% utilized centralized or external storage media, and 29% lacked an archiving description altogether. Additionally, Bishoff and Johnston (2015) found that one-third of DMPs employed the same practices for storing, sharing, and preserving data.

Differences Between Disciplines in DMP Content Analyses

Studies of disciplinary differences in RDM practice descriptions within DMP content analyses have been limited. They have mostly focused on differences in sharing and preserving methods. Parham et al. (2016) found that biology disciplines more commonly share data and employ metadata standards or a detailed description of different data types, suggesting a stronger culture of data sharing compared to other fields. Biology, along with the geosciences and social sciences, showed a preference for formal data sharing methods, particularly using discipline-specific repositories, indicating a potentially more sophisticated approach to data management. In contrast, the information sciences and engineering were less likely to use these formal methods. Additionally, researchers in the mathematical and physical sciences, as well as those in liberal arts, tended to favor supplements as a sharing mechanism, pointing to different cultural practices in data dissemination. Van Loon et al. highlight that engineering researchers more frequently included detailed data sharing policies, particularly concerning reuse and protection of sensitive data, than their liberal arts counterparts. They also noted that engineering DMPs were more likely to specify time frames for data sharing, underscoring a discipline-specific attention to detail in planning for data usage and preservation.

In conclusion, the DMP content analysis studies delineate a landscape where RDM practices are often incompletely described in DMPs. To address this, the subsequent methodology section details our empirical investigation into the quality of ECRs' DMPs, with the intention of identifying precise areas for pedagogical intervention.

Methods

This study aims to enhance our understanding of the quality variations in ECRs' DMPs and identify gaps in their RDM planning skills by evaluating the comprehensiveness and quality of the DMPs from their ongoing research projects. For this purpose, we assessed 223 DMPs created in the BRDM course between 2020 and 2022. Next, we briefly describe the BRDM course. For complete background information and a description of the course, as well as the results of the post-course survey and course feedback, please see Rantasaari (2022). Finally, we describe the assessment methods used in this study.

BRDM

The 3 ECTS credit, multi-stakeholder BRDM course for doctoral students and postdoc researchers has been arranged in two Finnish universities: the University of Turku (UTU) since 2019 and the Åbo Akademi University (ÅAU) since 2020. The course, structured into four tracks, encompasses an introductory lecture, seven modules, a voluntary Q&A session, and a final assignment, with general and module-specific learning objectives.[2]
The teachers are academic and research support professionals. The idea behind the four-track-based division is that the type of RDM actions needed and applied depend partly on data type, research methods, and discipline. Thus, the Clinical Health Sciences course track (CHSct) is primarily aimed at researchers using clinical methods in health sciences and the Natural Sciences course track (NSct) at researchers using the natural science approach. Survey Research (SRct) is similarly directed at researchers using survey methods and Qualitative Research (QRct) at researchers using the qualitative approach irrespective of their disciplines (Figure 1 in the Appendix).

All participants developed their own research plan in module one. After that, they started to write a DMP using the Finnish modification of DMP-online, DMPTuuli tool.[3] Each module's pre-class assignment was to write a draft of the DMP's relevant section; the post-class assignment was to update the section, informed by the discussion in the module's workshop. Everyone returned their DMP with an abstract of their research project and gave a structured, anonymous peer-review report of another participant's DMP at the end of the course. Finally, we gave a general level and personal feedback to the ECRs. We recommended the participants use the Finnish DMP Evaluation Guidance + General Finnish DMP Guidance (FDEG) (Aalto et al. 2021) as an aid to prepare their DMP and to review another participant's DMP. The three-point DMP performance criteria presented in the FDEG were developed by the Finnish DMP-Tuuli consortium's working group chaired by the author of this study during the Spring of 2021.

The assessment of the DMPs

By attending BRDM, participants consented to the use of their anonymized DMPs and assignments for research and curriculum development. The assessment evaluated the quality of DMPs as an indicator of participants' understanding and application of RDM concepts and principles. While we cannot conclusively ascribe the quality of the DMPs and the adherence to RDM best practices solely to the BRDM course, it is reasonable to assert that the analysis of 223 DMPs enables us to draw indicative conclusions about the impact of the RDM training.

The project's characteristics such as data types, sensitive data, and the number of collaborators, influence the required data management measures. The evaluation criteria of the FDEG consider varying levels of adequacy, categorizing them as insufficient, adequate, or excellent for different project types. For instance, Table 2 in the Appendix outlines the

[2] https://doi.org/10.5281/zenodo.3692224
[3] https://dmptuuli.fi/

evaluation criteria for rights management in DMP Section 2.2, contrasting multi-participant consortium research with single-researcher projects.

The participants were asked to attach an abstract of their project with the DMP to facilitate the assessment. Furthermore, if needed, the author could also consult the research plan returned in BRDM's Module One. This study's author read, assessed, and rated each DMP, applying FDEG's three-point performance criteria to all 11 sections of a General Finnish DMP template.[4] The sections were rated from 0 to 2: DMPs with median values between 0-0.66 were rated 'poor'; 0.67-1.33 were 'sufficient/satisfactory'; and 1.34-2 were 'good/excellent.' The maximum score for a DMP was 22 points. See the FDEG (Aalto et al., 2021) for complete information on the criteria and their contents.

The author read a DMP a second time after assessing it, using another participant's peer-review report as a reference. The author read and assessed a DMP third time if the scores of the first and second read differed markedly.

Besides rating the DMPs according to the FDEG performance criteria, the author defined categorical distinctions of the RDM best practices based on the recommendations by the FDEG (Aalto et al., 2021) and BRDM (Table 3 in the Appendix).

Data table (Dt)

A data table (Dt) is one of the RDM best practices. Both the BRDM course and the FDEG guidelines (Aalto et al., 2021) recommend its inclusion in Section 1.1 of a DMP. A data table should provide a comprehensive list or table of all data types that are reused, collected, and produced, detailing at minimum their formats and volumes, and ideally, other significant characteristics. (See Table 4 in the Appendix for an example of a data table).

Analysis

SAS JMP Pro 16 statistical software was used to analyze the results. Descriptive and inferential statistics with medians, custom quantiles, and p-values (Wilcoxon signed rank tests) were produced for ratings. Frequencies and p-values (Pearson's Chi-square test) were produced for the RDM best practices. A significance level of 0.05 (two-tailed) was used. When differences between DMPs in relation to their rating, best practices, or frequencies are discussed, they are statistically (highly) significant unless otherwise stated. See Rantasaari (2024) for the underlying data of the results.

Results

The Rating and the Quality of the DMPs According to the FDEG

In total, 201 doctoral students and 22 postdoc researchers developed a DMP in the 2020-2022 BRDM courses. The social sciences, business, and economics (SSBE) were the biggest disciplines, health sciences (HS) the second biggest, science and engineering (SE) the third biggest, and the humanities, psychology, theology (HPT) were the smallest disciplines (Figure 1).

[4] https://www.dmptuuli.fi/template_export/476471047.pdf

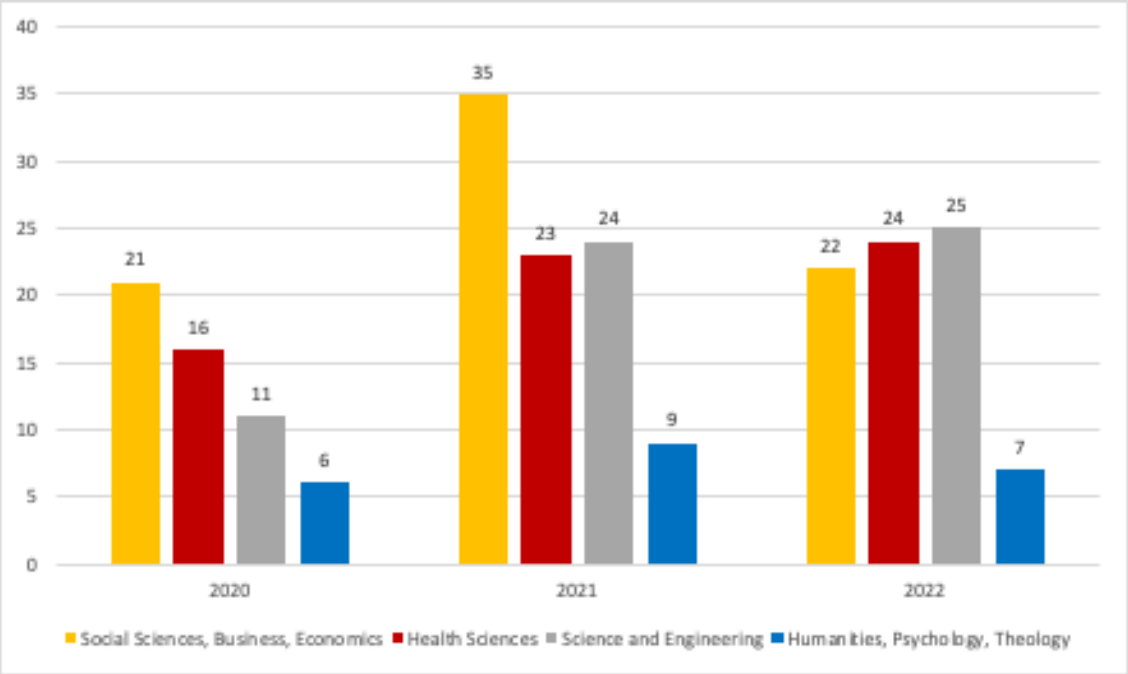

Figure 1. DMPs by disciplines in the BRDM 2020-2022 courses

Upon applying the FDEG's three-point criteria to evaluate the DMPs, we found that 27% received a 'good/excellent' rating, 64% were deemed 'sufficient/satisfactory', and 9% were considered 'poor.' The sections of the DMPs that were best described included 4.1 'Storage and security', with 65% rated 'good/excellent'; 4.2 'Related data security policies', with 41% rated as such; and 6.1 'Roles and responsibilities', also at 41% (Figure 2). Conversely, Section 5.2 'Data preservation' saw only 12% achieving a 'good/excellent' rating and 22% being rated 'poor'. Additionally, Section 6.2 'Budgeting and resourcing' had merely 8% rated 'good/excellent', while a significant 43% fell into the 'poor' category.

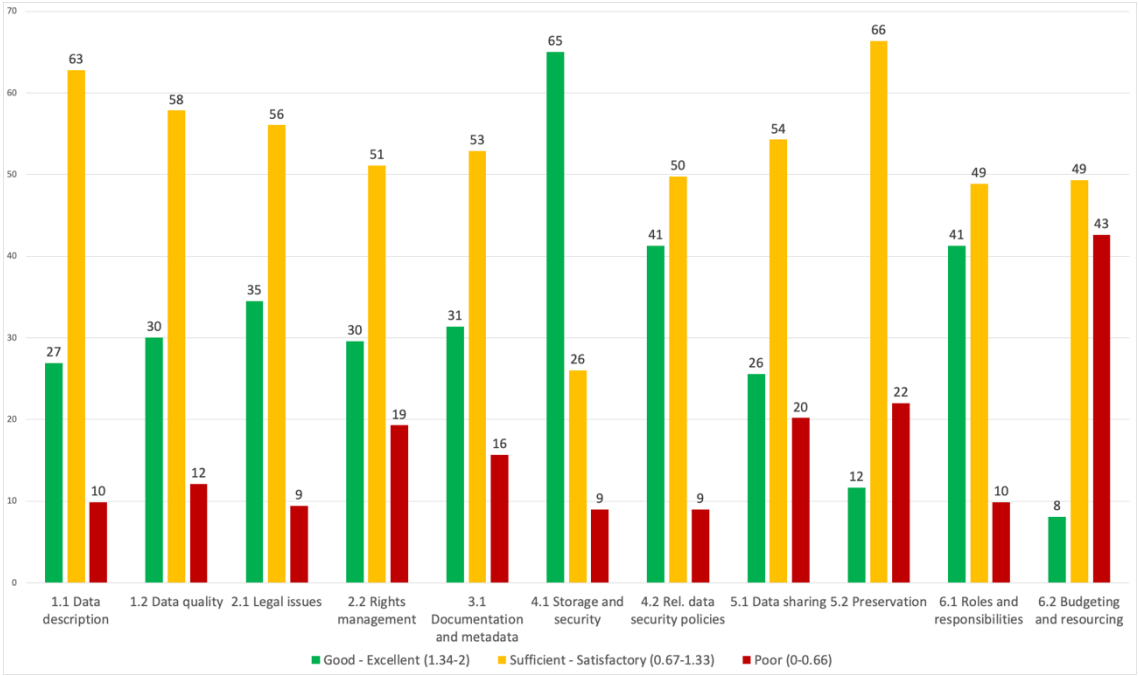

Figure 2. Percentual distribution of the scoring of the DMPs' sections from 0 to 2

In our comparative analysis of all DMPs, prose DMPs, and DMPs with a data table (DtDMPs), it was found that DtDMPs significantly differed from prose DMPs ($p<0.0001$, Wilcoxon Rank Sum Test). Specifically, in the overall assessment:

- Only Section 4.1 'Storage and security' achieved a 'good/excellent' rating in all DMPs, with the rest rated as 'sufficient/satisfactory'.
- For prose DMPs, apart from Section 4.1 'Storage and security' which was rated 'good/excellent', all other sections were deemed 'sufficient/satisfactory', except for Section 6.2 'Budgeting and resourcing', which was categorized as 'poor'.
- In contrast, DtDMPs performed notably better in several sections. The 'good/excellent' rating was attained in Sections 1.1 'Data description', 1.2 'Data quality', 2.1 'Legal issues', 4.1 'Storage and security', 4.2 'Related data security policies', and 6.1 'Roles and responsibilities', while the remaining sections were rated as 'sufficient/satisfactory'.

The overall median ratings highlight this disparity, with DtDMPs achieving a median of 1.09 (representing 45% of the sample), compared to 0.98 for prose DMPs (55% of the sample), and a median of 1.01 for all DMPs (n=223). This indicates that DtDMPs consistently received the highest median ratings across all assessed sections. For a visual representation of these findings and for a detailed breakdown of medians, custom quantiles, and p-values, see Figure 2 and Table 5 in the Appendix.

The RDM Best Practices

The assessed RDM best practices were based on recommendations derived from the FDEG and BRDM (Figure 3). In the DMPs, descriptions of RDM practices were typically rated as 'sufficient/satisfactory,' whereas data storage strategies, particularly for personal data (n=166), received a 'good/excellent' rating, with 68% being safe and secure. Despite the median for data sharing not achieving 'good/excellent,' a substantial proportion of ECRs were ready to share data – 80% planned to share metadata (n=178) and 61% some research data (n=136) through formal venues, including repositories and file-sharing services. Moreover, 46% of DMPs that planned formal data sharing, specified a reuse license. In terms of outlining RDM roles and responsibilities, 41% were rated as 'good/excellent' and 49% as 'sufficient/satisfactory.'

However, further learning needs also remained. For example, only 7% (n=16) of the DMPs' authors acknowledged a funder's or a publisher's data sharing policy. Additionally, only 19% (n=32) of DMPs with personal data (n=166) identified a data controller, and just 25% (n=42) stated the legal basis for data processing, both mandates of the EU's GDPR. Of those intending to share anonymized data (n=107), only 33% (n=35) indicated awareness of the needed permission for sharing and reuse. Conversely, less than half, 46% (n=58), of the authors who did not plan to share metadata or data (n=126) provided reasons. Finally, the ownership was clearly described and justified in only 42% (n=93) of the DMPs.

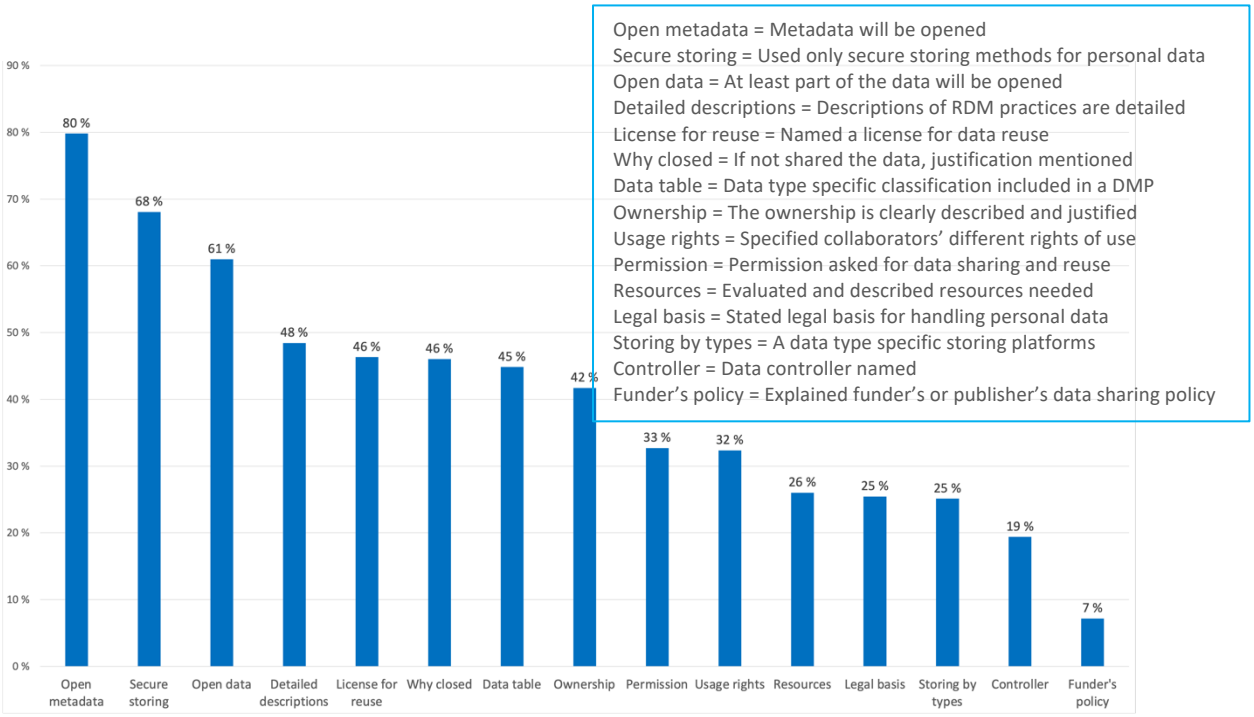

Figure 3. The RDM best practices in DMPs 2020-2022

The differences between the year of the training, roles (doctoral students or postdoc researchers), BRDM course tracks, and disciplines were statistically insignificant when DMPs were rated according to the FDEG's three-point criteria. However, DtDMPs differed from prose DMPs regarding seven RDM best practices in which a majority of the DtDMPs but a minority of the prose DMPs had sound descriptions for RDM operations (Figure 4).

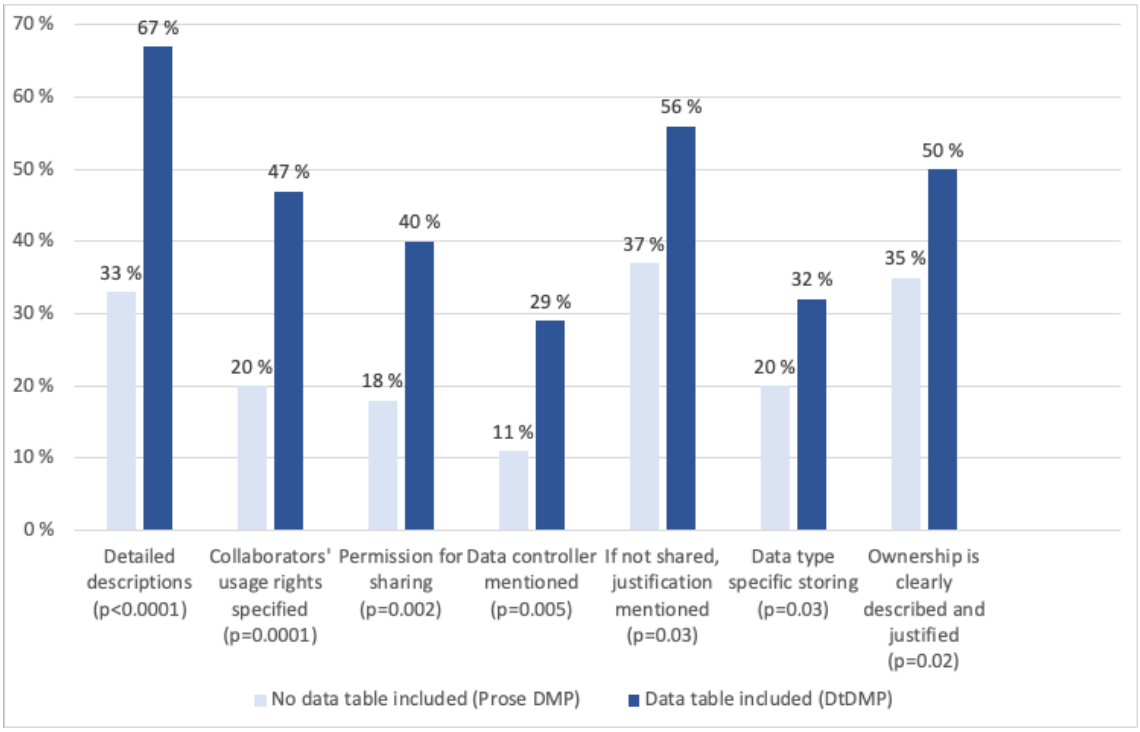

Figure 4. Differences in the RDM best practices between DtDMPs and prose DMPs with p-values (Pearson's Chi-square test)

Consequently, DtDMPs typically offered a more comprehensive, detailed, and justified account of RDM practices compared to prose DMPs. Additionally, the explicit specification of data types within a research project contributed to the structured nature of DtDMPs. This likely facilitated the authors' ability to delineate and articulate the specific practices required for different data types, such as the selection of appropriate storage methods.

Next, we will take a closer look at data sharing, storing, and preserving methods and differences between disciplines, course tracks, and DtDMPs vs. prose DMPs. Depending on the nature of different data types used in a project, an ECR could mention more than one sharing, storing, or preserving method: For example, part of the data could be shared through a generalist repository (e.g., Zenodo, Harvard Dataverse) and part through a file sharing service (e.g., GitHub, Dropbox); hence, the chosen venues and their counts are overlapping because choosing one venue does not exclude other venues.

Data sharing

In our assessment of data sharing practices, we found that 78% of the DMPs included plans for either formal or informal data sharing. 'Open data', as represented in Figure 3, specifically excludes informal methods such as sharing upon request or within a publication due to their less reliable nature – a concern elaborated in the literature review. For a comprehensive understanding, however, these informal methods are accounted for in Figure 5 below and Table 6 in the Appendix.

A discipline-specific repository was the primary sharing method for 26% of researchers, particularly within the social sciences, business, and economics (SSBE) fields, and those conducting qualitative or survey research (QRct, SRct). Conversely, researchers from the science and engineering (SE) sectors, and those utilizing natural science methodologies (NSct), showed a preference for generalist repositories. Publication as a sharing method was more common among STEM disciplines and clinical or natural science method users (CHSct, NSct). DtDMPs referenced publication sharing method in 14% (n=14) of cases, compared to 24% (30) for prose DMPs – a notable difference, albeit not reaching statistical significance (p=0.05).

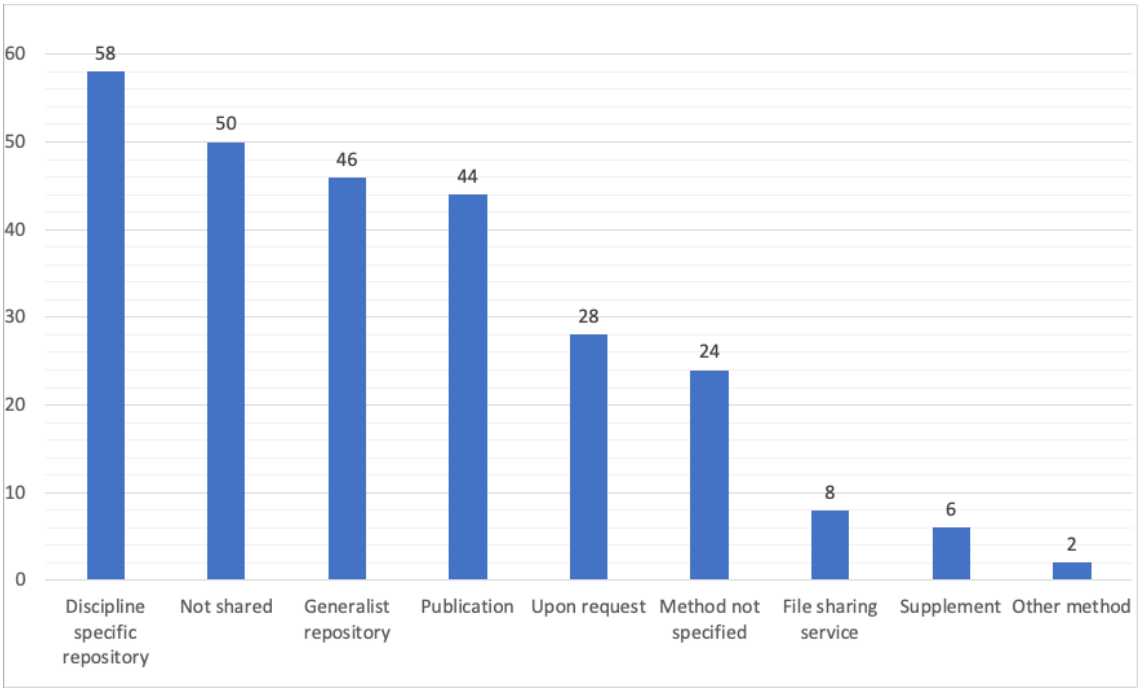

Figure 5. Data sharing venues (n=223)

Data storing

Data storing, defined per FDEG guidelines (Aalto et al., 2021), is the retention of data during the active research phase. As depicted in Figure 6, 82% of researchers (183) employed institutional network drives or cloud services. External drives were a secondary preference among natural science researchers (NSct), while health science researchers (HS) and those using clinical methods (CHSct) often utilized data collection or analysis software for storage (Table 7 in the Appendix). Storage on lab or department servers was significantly more referenced in prose DMPs than in DtDMPs (11% (14) versus 2% (2), p=0.007).

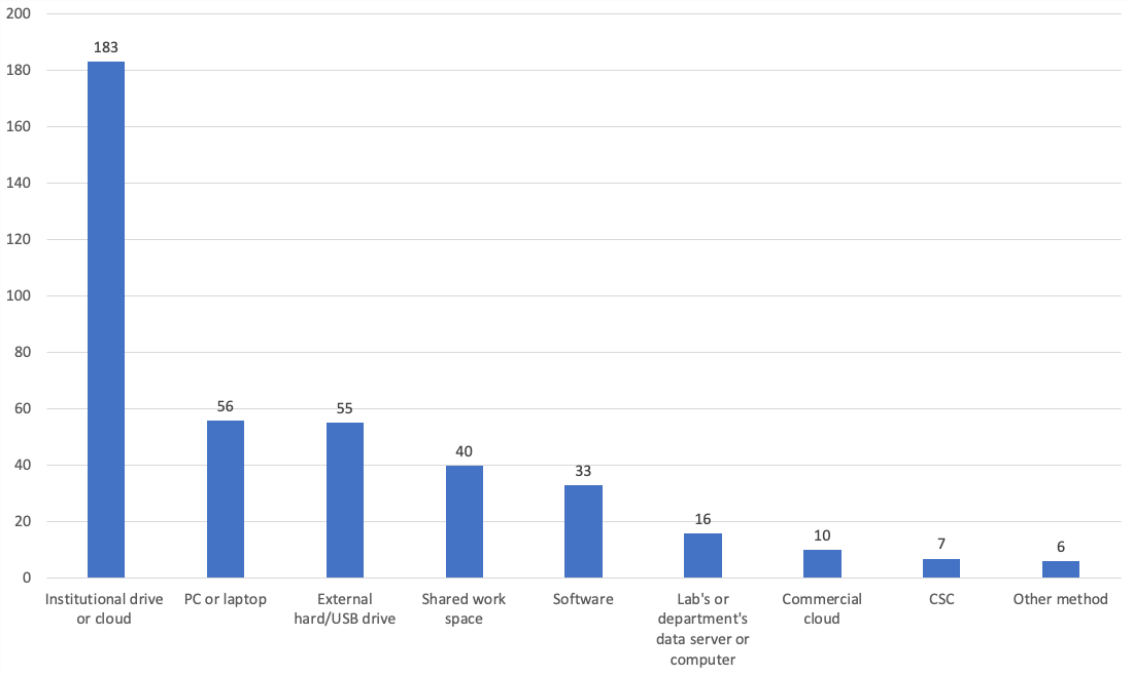

Figure 6. Data storing platforms (n=223)

Data preserving venues

Regarding data preservation, which entails the retention of data beyond the active phase of research, 31% of researchers (69) utilized institutional drives, clouds, or servers, with this method being prevalent among health sciences (HS) and clinical method researchers (CHSct) (Figure 7, Table 8 in the Appendix). Humanities and social sciences researchers (SSBE, HPT), as well as those using qualitative or survey methods (QRct, SRct), often chose discipline-specific archives, while generalist archives were favoured by science and engineering (SE) researchers and natural science method (NSct) users.

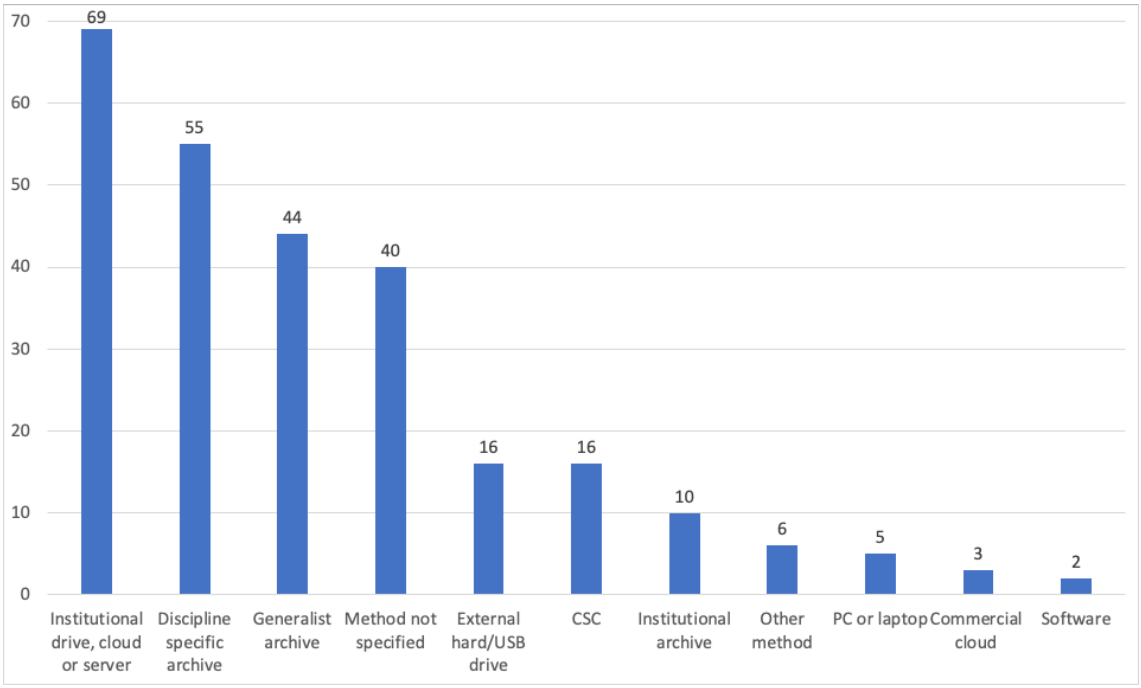

Figure 7. Data preserving venues (n=223)

Data preserving time frame

Lastly, the term 'archiving' in our study refers to the long-term preservation of data. According to FDEG guidelines, data should be preserved for verification for 5-15 years, for potential reuse for about 25 years, and indefinitely if it has permanent value (Aalto et al., 2021). The most common preservation timeframe was 5-15 years, favored by 40% (90) of researchers, with health science (HS) researchers and those using clinical methods (CHSct) particularly endorsing this period for data preservation (Figure 8, Table 9 in the Appendix). DtDMPs more frequently suggested preserving data for potential reuse for 25 years compared to prose DMPs (21% (21) versus 11% (13)), a statistically significant difference (p=0.03).

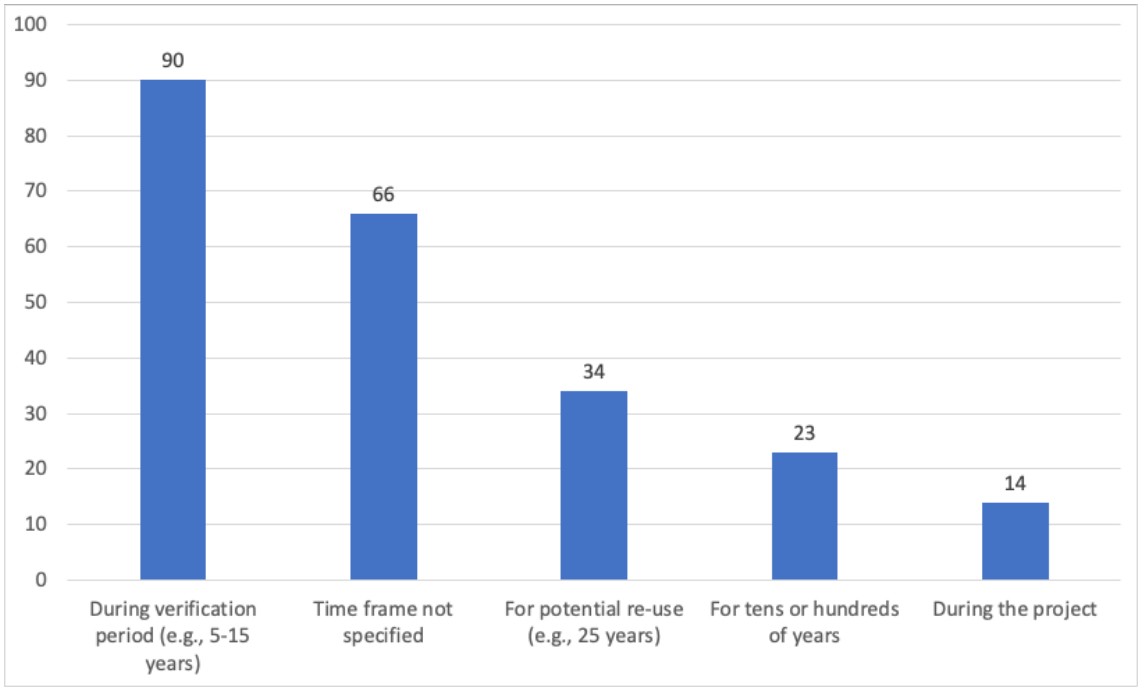

Figure 8. Preserving time frame (n=223)

Limitations

In this study, we have based our analysis on a convenience sample, which may not encompass the full diversity of ECRs' DMPs. As such, the findings might not be generalizable to all ECRs. Additionally, the DMPs assessed were developed as part of an RDM training course, which may differ in content and emphasis from those created for grant proposals, potentially affecting the applicability of our results to other contexts. For instance, the focus on FAIR principles in the course might have influenced the DMPs to reflect less disciplinary variation than might be observed among more established researchers with ingrained discipline-specific practices.

Moreover, while the DMPs provide a blueprint of intended research practices, we cannot ascertain if researchers will follow through with the actions described, a limitation that underscores the need for follow-up research to evaluate the actual implementation of these practices. The current evaluation of DMPs and the earlier self-assessment of the competencies before and after the BRDM course (Rantasaari, 2022), although indicative, does not allow us to definitively determine to what extent the quality and the applied RDM best practices are attributable to the course itself or pre-existing skills among the participants.

Despite these limitations, the assessment of 223 DMPs generated during the BRDM course provides a solid basis for drawing indicative conclusions about the impact of RDM training. It sheds light on the current state of the ECRs' RDM planning skills and the quality of their DMPs. Moreover, it helps to identify areas where further education is required.

Discussion

Our study examined early career researchers' DMPs using the FDEG's criteria to enhance our understanding of the quality variations in the DMPs and identify gaps in ECRs' RDM planning skills. Analyzing 223 DMPs from the BRDM course, we address critical questions regarding their quality, alignment with RDM best practices, and variability across different variables. We then interpret these results within the broader framework of RDM stakeholder expectations.

The Quality of the DMPs

In terms of the FDEG's three-point criteria, over a quarter of the DMPs received a rating of 'good/excellent', almost two-thirds were rated 'sufficient/satisfactory', and nearly one-tenth were rated 'poor.' The median of all DMPs was satisfactory 1.01, DtDMPs 1.09 and prose DMPs 0.98 out of 2. The satisfactory median result is in line with the findings of the author's (2022) earlier research, where participants in the BRDM course reported improved self-assessment scores, advancing from 'little competence' to 'somewhat competent' in 2020–2021.

RDM Best Practices

When examining the RDM best practices that enhanced the quality of a DMP, we found that ECRs demonstrate robust knowledge and application in data storage, evidenced by 65% of DMPs receiving 'good/excellent' ratings. This performance is likely bolstered by UTU and ÅAU's infrastructure providing net drives or cloud storage, with 82% utilization rate among ECRs.

Moreover, 61% of the researchers planned to share at least a portion of their data through formal methods, while 78% considered using either formal or informal methods. Of the DMPs indicating formal data sharing, just under half (46%) specified a license for data reuse, compared to 19 to 41% in earlier DMP content analysis studies (Mannheimer, 2018; Parham et al., 2016; Van Loon et al., 2017). Notably, informal data sharing methods were mentioned less frequently

than in previous studies (Berman, 2017; Bishoff & Johnston, 2015; Curty et al., 2013; Mischo et al., 2014; Parham et al., 2016; Van Loon et al., 2017), suggesting a trend towards formal sharing mechanisms. A comparison with these previous studies reveals a lower intent for data sharing, potentially reflecting the unique context of these DMPs, which were not primarily created for grant applications that mandate data dissemination.

Regarding the delineation of roles and responsibilities, 41% of the DMPs were rated as 'good/excellent.' This represents a higher rate compared to earlier DMP content analyses by Nicholls et al. (2014), Samuel et al. (2015), and Van Loon et al. (2017), where such descriptions were less frequent. A contributing factor to this improvement may be the emphasis on identifying responsible parties within the RDM process in the FDEG and BRDM.

Conversely, the DMPs frequently fail to mention funders' or publishers' data sharing policies – 93% lacked such mentions, possibly indicating a gap in the authors' familiarity with these requirements or an absence of intent to procure external funding. Additionally, descriptions for handling of personal data were often faint, with 81% omiting the naming of a data controller and 75% bypassing the legal basis, contrary to GDPR mandates. These oversights signal a critical need for enhanced training on legal, contractual, and data protection issues in RDM.

Other notable omissions include the budgeting and resourcing of RDM actions, with 74% of DMPs lacking in details, suggesting the future training may need to place greater emphasis on these aspects. Furthermore, the lack of data sharing permissions (68%), absence of reasons for withholding data (54%), and unclear definitions and justifications for data ownership (58%) were additional areas of concern.

Differences Between DMPs

Contrary to the negligible variations in quality across training year, discipline, course track, or roles as per FDEG ratings, the difference between DtDMPs and prose DMPs was statistically highly significant. DtDMPs were notably more comprehensive in detailing data types, their characteristics, and needed operations. DtDMPs also more effectively addressed rights management and personal data handling.

Additionally, differences especially in data sharing and preservation methods emerged when comparing various disciplines. Course tracks also showed distinct preferences in these areas. Researchers in the social sciences, business, and economics, as well as those conducting qualitative or survey research, favored discipline-specific repositories for sharing and preserving data. In contrast, researchers in science and engineering, and those employing natural science methods, preferred generalist repositories. The former preference may be partly due to the Finnish Social Sciences Data Archive (FSD), the most-used single discipline-specific repository, which has broadened its scope from exclusively social sciences data to include any research data involving human subjects. Thus, FSD exhibits characteristics akin to those of generalist repositories. Furthermore, UTU and ÅAU have only one discipline-specific institutional repository for history, culture, and arts data. It is also possible that many ECRs in STEM disciplines are unaware of their disciplines' repositories.

Second, researchers in STEM disciplines and those using clinical or natural science methods often chose publication as a data sharing venue. This preference might indicate uncertainty about suitable repositories for their data, or some researchers may have conflated open access publication of research articles with open sharing of research data (see also Mischo et al., 2014; Parham et al., 2016).

Third, approximately one-third of all researchers – and nearly half of those in the health sciences and using clinical health methods – intended to preserve their data on an institutional net drive or cloud service. However, such methods are not always suitable for long-term preservation due to the potential disposal of the net drive or cloud account when a researcher leaves the organization.

Fourth, about one-fourth of the researchers left the preservation method unspecified. This is comparable to Parham et al.'s (2016) finding that 29% of DMPs did not include an archiving description.

A likely reason for these preservation issues is that UTU and ÅAU offer viable, user-friendly institutional solutions only for data storage during the active phase of a research project, not for preservation after the project's completion. Researchers are instead advised to use international generalist or discipline-specific repositories, which they may find unfamiliar or challenging to use.

Differences in preferred storage methods, aside from data sharing and preservation, were also identified. Variations were primarily found in the use of laptops, external drives, shared workspaces, and other platforms, which often served as secondary or tertiary storage options. Notwithstanding these differences, a significant majority (82%) of researchers consistently stored or backed up their data on institutional network drives or cloud services.

While these variations in the descriptions of RDM practices demonstrate areas of strength and weakness, they must be contextualized within the broader ecosystem of stakeholder requirements and expectations.

The Structure and Content of a DMP

Stakeholders have varied requirements for DMPs, as noted by Smale et al. (2020). However, Kvale and Pharo (2021) argue that the primary function of DMPs is to aid researchers in organizing and documenting their data for easier sharing. Considering this, which approach is more effective: standardizing and automating DMPs with machine-readable elements, or enhancing them with autonomy, flexibility, and open-ended questions that enable researchers to create a DMP tailored to their project, with support from peers and research services? Smale et al. (2020) criticize current DMP forms as trying to please too many stakeholders simultaneously and end up pleasing no one. Their proposed solution is a strict specification of the requirements of the multiple stakeholders and differentiating a current multi-purpose DMP to these separate needs. Kvale and Pharo propose that a DMP should be developed foremost as a tool for researchers by lowering standardization, such as close-ended questions, and increasing flexibility, autonomy, and open-ended questions that allow a researcher to plan the organization and documentation of data from their own premises. Thus, by preserving the possibility for unique features of research projects, instead of standardizing and simplifying them away, a DMP best serves also other stakeholders, such as research institutions, other researchers, research support services, policy makers, and funders (Kvale and Pharo, 2021).

While this study does not address machine actionability or the automation of DMPs, it reveals that DMPs using structured data tables are of higher quality than those based solely on prose text. This underpins the advantage of DMP templates that prompt for generic, discipline-specific, or data type-specific information. A structured DMP, categorizing data types and addressing their unique storage, processing, preservation, and sharing requirements, meets the needs of stakeholders more effectively, including the needs of researchers. However, it is crucial to incorporate prose responses in order to maintain the potential for distinct characteristics of research projects, as highlighted by Kvale and Pharo (2021).

Thus, the viewpoints of the DtDMPs' and prose DMPs' different strengths offer critical insights for both RDM policy and training. Policies could be developed to promote a hybrid approach, leveraging the strengths of both structured and narrative styles. Concurrently, training programs could be tailored to equip researchers with skills to navigate these styles effectively. Emphasizing the importance of narrative for capturing the nuances of research processes, while advocating for structured formats for ease of data retrieval and compliance, can foster more robust and comprehensive RDM strategies.

Conclusions

Prior studies have highlighted several factors that compromise DMP quality and effective RDM, such as unclear assessment criteria and ambiguous stakeholder requirements (Hudson-Vitale &

Moulaison-Sandy, 2019; Smale et al., 2020), and constraints related to time, skills, infrastructure, and incentives (Bardyn et al., 2012; Berman 2017; Perrier et al., 2020; Tenopir et al., 2011; Van den Eynden & Bishop, 2014).

Our study has identified quality variations in ECRs' DMPs, revealing a commendable adherence to RDM best practices, yet it also highlights critical gaps that can be addressed through targeted policy interventions and training programs.

Key Findings

While ECRs' DMPs meet an acceptable quality standard based on FDEG criteria, our assessment found that they often lack specificity how they describe RDM practices, particularly in areas such as data sharing details, including permissions, scope, timing, and preservation methods. Our study has found that a structured tool, such as a data table, prompts researchers to more precisely define their RDM strategies, thereby enhancing the DMP's overall quality. Additionally, the study sheds light on how discipline, research method, and data type influence RDM choices, affecting decisions on data sharing, storage, and preservation. These findings can be utilized to develop individual DMPs and RDM policies and practices in general.

Practical Implications

This study provides RDM stakeholders – researchers, institutions, funders, and publishers – a standardized FDEG framework model for developing and evaluating DMPs that is independent of the research discipline, methods used, data types, and the nature of the project.

For researchers, the implications are clear: They can improve the descriptions of data management actions, for example, through a FDEG's three-point criterion and recommended RDM best practices such as a structured data table. This can not only improve the data management of the research but also contribute to the integrity, reliability, and reproducibility of the entire study.

Institutions can utilize the FDEG framework model to discern both strengths and potential improvements in DMPs, and to tailor services and training to the researchers' needs, taking into account the unique impacts of research discipline, method, and data type. This, in turn, allows for more targeted development of support services, training, and infrastructures. Furthermore, educators can apply the method developed and used in this study to assess the impact of RDM training on the quality of real-world DMPs as a practical tool in the evaluation of the outcomes of their RDM training.

Finally, funders and publishers can exploit the FDEG framework by specifying their requirements for the maintenance of the DMP, the quality of action descriptions, and adherence to recommended best RDM practices. For example, a funder could require that the descriptions in the first, preliminary version of the DMP be at least 'sufficient/satisfactory' level, but the final version's descriptions should be at 'good/excellent' level. This would improve the transparency and integrity of data management and the entire research process, compliance with laws and ethical standards, and thereby also enable more efficient data sharing and reuse.

Future Directions

To catalyze cultural shifts towards robust RDM, we must extend beyond training to a tighter integration of RDM within the research lifecycle. Our findings suggest that enhancing RDM knowledge and practices through training is a critical step, yet alone it is insufficient for the necessary cultural transformation. Embedding research support with research activities and promoting the use of structured DMPs, coupled with stronger incentives for data sharing, preservation, and reuse, could advance this shift.

Our research highlights the tangible benefits of RDM training and the employment of structured data tables in DMPs, significantly advancing the integration of data management theory with practical application. This facilitates more informed and strategic research

processes. In light of these advances, future studies should aim to refine our understanding of the training's effectiveness by comparing the DMPs of ECRs who have received training with those who have not. Furthermore, research should assess the extent to which researchers actualize the practices outlined in their DMPs within their daily research activities.

In future work, we will expand upon this research by delving deeper into RDM competencies within the context of the analyzed DMPs. This will involve making connections between RDM practices, BRDM modules, established best practices, and participants' self-assessment of their competency development.

Acknowledgements

I would like to thank my supervisors, Professor Gunilla Widén at the Åbo Akademi University and Professor Isto Huvila at the Uppsala University, who read and commented on several drafts of this study; and Carolyn Abbott for proofreading the manuscript.

Data

The quantitative data underlying this study can be accessed through Zenodo (Rantasaari, 2024).

References

Aalto, S., Ahokas, M., Friman, J., Fuchs, S., Korhonen, T., Kuusniemi, M. E., Laakso, K., Lennes, M., Manninen, S., Ojanen, M., Rantasaari, J., Virtanen, M. E., Xu, Q. (2021) *Finnish DMP evaluation guidance + General Finnish DMP guidance.* [Working paper]. Zenodo. https://doi.org/10.5281/zenodo.4729831

Academy of Finland. (2019). *The Academy of Finland's funding terms and conditions 2019-2020* (Vol. 1, Issue September 2019). https://www.aka.fi/globalassets/10rahoitus/liiteet/rahoitusehdot_2019-2020_en.pdf

Amsterdam call for action on open science. (2016, April 5). In W*ikipedia*. https://en.wikipedia.org/wiki/Amsterdam_Call_for_Action_on_Open_Science

Bardyn, T. P., Resnick, T., & Camina, S. K. (2012). Translational researchers' perceptions of data management practices and data curation needs: Findings from a focus group in an academic health sciences library. *Journal of Web Librarianship*, *6*(4), 274–287. https://doi.org/10.1080/19322909.2012.730375

Berman, E. A. (2017). An exploratory sequential mixed methods approach to understanding researchers' data management practices at UVM: Findings from the qualitative phase. *Journal of EScience Librarianship*, *6*(1), 5. https://doi.org/10.7191/jeslib.2017.1097

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data authorship as an incentive to data sharing. *New England Journal of Medicine*, *376*(17), 1684–1687. https://doi.org/10.1056/NEJMSB1616595/SUPPL_FILE/NEJMSB1616595_DISCLOSURES.PDF

Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication*, *3*(2), eP1231. https://doi.org/10.7710/2162-3309.1231

Carlson, J., Fosmire, M., Miller, C. C., & Sapp Nelson, M. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Portal: Libraries and the Academy*, *11*(2), 629–657. https://doi.org/10.1353/pla.2011.0022

Chiarelli, A., Loffreda, L., & Johnson, R. (2021). *The art of publishing reproducible research outputs: Supporting emerging practices through cultural and technological innovation*. [Report]. Zenodo. https://doi.org/10.5281/ZENODO.5521077

Coates, H. L. (2014). Ensuring research integrity: The role of data management in current crises. *College and Research Libraries News, 75*(11), 598–601. https://doi.org/doi:10.5860/crln.75.11.9224

Curty, R., Kim, Y., & Qin, J. (2013). What have scientists planned for data sharing and reuse? A content analysis of NSF awardees' data management plans. In *Research data access & preservation summit 2013*. https://surface.syr.edu/ischoolstudents/2/

Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *The Journal of Academic Librarianship*, *40*(3–4), 322–331. https://doi.org/10.1016/J.ACALIB.2014.04.010

European Commission. (2018a). *OSPP-REC: Open science policy platform recommendations*. OSPP-REC. https://doi.org/10.2777/958647

European Commission. (2018b). Commission recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information. *Official journal of the European Union* L 134, p. 12–18 . http://data.europa.eu/eli/reco/2018/790/oj

European Commission. (2022). *European research data landscape: Final report*. https://op.europa.eu/en/publication-detail/-/publication/03b5562d-6a35-11ed-b14f-01aa75ed71a1/language-en/format-PDF/source-275372809

European University Association. (2017). *Towards full open access in 2020: Aims and recommendations for university leaders and national rectors' conferences*. http://www.eua.be/Libraries/publications-homepage-list/towards-full-open-access-in-2020-aims-and-recommendations-for-university-leaders-and-national-rectors-conferences

Hudson-Vitale, C., & Moulaison-Sandy, H. (2019). Data management plans: A review. *DESIDOC Journal of Library and Information Technology*, *39*(6), 322–328. https://doi.org/10.14429/djlit.39.6.15086

Ioannidis, J. P. A. (2016). Anticipating consequences of sharing raw data and code and of awarding badges for sharing. *Journal of Clinical Epidemiology*, *70*, 258–260. https://doi.org/10.1016/J.JCLINEPI.2015.04.015

Kvale, L., & Pharo, N. (2021). Understanding the mata management plan as a boundary object through a multi-stakeholder perspective. *International Journal of Digital Curation*, *15*(1), 16-16. https://doi.org/10.2218/ijdc.v16i1.746

Lefebvre, A., Schermerhorn, E., & Spruit, M. (2018). How research data management can contribute to efficient and reliable science. In *26th European Conference on Information Systems: Beyond Digitization – Facets of Socio-Technical Change, ECIS 2018* (pp. 1–15). Association for Information Systems. Retrieved from https://aisel.aisnet.org/ecis2018_rp/35

Mannheimer, S. (2018). Toward a better data management plan: The impact of DMPs on grant funded research practices. *Journal of EScience Librarianship*, *7*(3), e1155. https://doi.org/https://doi.org/10.7191/jeslib.2018.1155

Mischo, W. H., Schlembach, M. C., & O'Donnell, M. N. (2014). An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *Journal of EScience Librarianship*, *3*(1), e1060. https://doi.org/10.7191/jeslib.2014.1060

Molloy, L., & Snow, K. (2012). The data management skills support initiative: Synthesising postgraduate training in research data management. *International Journal of Digital Curation, 7*(2), 101–109. https://doi.org/10.2218/ijdc.v7i2.233

National Science Foundation. (2011). *Grant proposal guide* (NSF 11-1). https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp

Nicholls, N. H., Samuel, S. M., Lalwani, L. N., Grochowski, P. F., & Green, J. A. (2014). Resources to support faculty writing data management plans: Lessons learned from an engineering pilot. *International Journal of Digital Curation*, *9*(1), 242–252. https://doi.org/10.2218/IJDC.V9I1.315

Parham, S. W., & Doty, C. (2012). NSF DMP content analysis: What are researchers saying? *Bulletin of the American Society for Information Science and Technology*, *39*(1), 37–38. http://hdl.handle.net/1853/48707

Parham, S. W., Carlson, J., Hswe, P., Westra, B., & Whitmire, A. (2016). Using data management plans to explore variability in research data management practices across domains. *International Journal of Digital Curation*, *11*(1), 53–67. https://doi.org/10.2218/ijdc.v11i1.423

Perrier, L., Blondal, E., & MacDonald, H. (2020). The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PLoS ONE*, *15*(2), 1–21. https://doi.org/10.1371/journal.pone.0229182

Pierce, H. H., Dev, A., Statham, E., & Bierer, B. E. (2019). Credit data generators for data reuse. *Nature 2021 570:7759*, *570*(7759), 30–32. https://doi.org/10.1038/d41586-019-01715-4

Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, *63*(2), 123–134. https://doi.org/10.1515/libri-2013-0010

Qin, J., D'ignazio, J., & D'ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata*, *10*(2–3), 188–204. https://doi.org/10.1080/19386389.2010.506379

Rantasaari, J. (2021). Doctoral students' educational needs in research data management: Perceived importance and current competencies. *International Journal of Digital Curation*, *16*(1), 36. https://doi.org/10.2218/IJDC.V16I1.684

Rantasaari, J. (2022). Multi-stakeholder research data management training as a tool to improve the quality, integrity, reliability and reproducibility of research. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *32*(1), 1–54. https://doi.org/10.53377/LQ.11726

Rantasaari, J. (2024). Assessing quality variations in early career researchers' data management plans: Quantitative data of the content analysis [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10620761

Rantasaari, J., & Kokkinen, H. (2019). Closing the skills gap: The basics of the research data management (BRDM) course: Case University of Turku. *Proceedings of the IATUL Conferences*, 1–8. https://docs.lib.purdue.edu/iatul/2019/fair/5/

Samuel, S. M., Grochowski, P. F., Lalwani, L. N., & Carlson, J. (2015). Analyzing data management plans: Where librarians can make a difference. *ASEE Annual Conference and Exposition, Conference Proceedings*, *122nd ASEE*, 1–21. https://peer.asee.org/analyzing-data-management-plans-where-librarians-can-make-a-difference

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLOS ONE*, *4*(9), e7078. https://doi.org/10.1371/JOURNAL.PONE.0007078

Scaramozzino, J. M., Ramírez, M. L., & McGaughey, K. J. (2012). A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries*, *73*(4), 349–365. https://doi.org/10.5860/crl-255

Scholtens, S., Anbeek, P., Böhmer, J., Brullemans-Spansier, M., Geest, M. van der, Jetten, M., Staiger, C., Slouwerho, I., & Gelder, C. W. G. van. (2019). *Life sciences data steward function matrix, version 1.1.* https://10.5281/ZENODO.2561723

Smale, N., Unsworth, K., Denyer, G., Magatova, E., & Barr, D. (2020). A review of the history, advocacy and efficacy of data management plans. *International Journal of Digital Curation*, *15*(1), 1–29. https://doi.org/10.2218/ijdc.v15i1.525

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). *Primer on data management.* Retrieved from: http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), 1–21. https://doi.org/10.1371/journal.pone.0021101

Thessen, A. E., Mcginnis, S., & North, E. W. (2015). Lessons learned while building the Deepwater Horizon Database: Toward improved data sharing in coastal science. *Computers & Geosciences*, *87*, 84–90. https://doi.org/10.1016/j.cageo.2015.12.001

Thoegersen, J. L., & Borlund, P. (2022). Researcher attitudes toward data sharing in public data repositories: A meta-evaluation of studies on researcher data sharing. *Journal of Documentation*, *78*(7), 1–17. https://doi.org/10.1108/JD-01-2021-0015/FULL/PDF

UNIFI. (2016). *Open science and data: Action programme for the Finnish scholarly community*. https://www.unifi.fi/wp-content/uploads/2019/04/UNIFI_Open_Science_and_Data_Action_Programme.pdf

Van den Eynden, V., & Bishop, L. (2014). *Sowing the seed: Incentives and motivations for sharing research data, a researcher's perspective*. https://www.knowledge-exchange.info/projects/project/research-data/sowing-the-seed

Van Loon, J. E., Akers, K. G., Hudson, C., & Sarkozy, A. (2017). Quality evaluation of data management plans at a research university. *IFLA Journal*, *43*(1), 98–104. https://doi.org/10.1177/0340035216682041

Van Tuyl, S., & Whitmire, A. L. (2016). Water, water, everywhere: Defining and assessing data sharing in academia. *PLoS One*, *11*(2), e0147942. https://doi.org/10.1371/journal.pone.0147942

Wellcome. (2017, July 10). *Data, software and materials management and sharing policy*. https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy

Weller, T., & Monroe-Gulick, A. (2014). Understanding methodological and disciplinary differences in the data practices of academic researchers. *Library Hi Tech*, *32*(3), 467–482. https://doi.org/10.1108/LHT-02-2014-0021

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J. S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/J.CUB.2013.11.014

Appendix

Table 1. DMP Content analyses

Author	Topic	Method	Number of analyzed DMPs	Context
Berman (2017)	Data management behaviors of researchers, especially data sharing and preservation practices and barriers researchers have encountered	Content analysis; Semi-structured interviews	35	DMPs included in NSF's funded grant proposals at the University of Vermont
Bishoff and Johnston (2015)	Mainly STEM discipline researchers' different data sharing approaches	Content analysis	182	DMPs included in funded NSF's grant proposals at the University of Minnesota
Curty et al., (2013	Data sharing, archiving, and reuse descriptions	Content analysis; Faculty survey	68	NSF awardees' DMPs at several U.S. universities
Mannheimer (2018)	The impact of DMPs on grant awards and on principal investigators' data management and sharing practices	Content analysis; Semi-structured interviews	186	Montana State University researchers' awarded and declined NSF's grant proposals that included a DMP
Mischo et al., (2014)	The impact of DMPs' proposed storage, preservation, and sharing venues on funding decisions	Content analysis	1260	DMPs submitted in NSF's proposals at the University of Illinois
Nicholls et al., (2014)	Examination and comparison of the DMPs to the NSF requirements and the Engineering Directorate	Content analysis; Faculty survey	104	Engineering faculty's DMPs from accepted NSF proposals at the University of Michigan
Parham and Doty (2012)	Descriptions and references in DMPs to the institutional repository services	Content analysis	181	NSF DMPs at Georgia Tech
Parham et al. (2016)	DART rubric in assessing and rating data sharing, discovery and reuse, and the planned use of data curation infrastructure	Content analysis	465	NSF DMPs at five U.S. research institutes
Samuel et al. (2015)	Evaluation of the DMPs using three different rubrics	Content analysis	29	DMPs from accepted NSF proposals at the University of Michigan
Smale et al. (2020)	Analysis of the	Content analysis; A	834	A random sample of

Author	Topic	Method	Number of analyzed DMPs	Context
	DMPs concerning their benefit for researchers, institutions or funding bodies	review		completed DMPs from the major Australian institution's DMP database
Van Loon et al. (2017)	Evaluation of the content of DMPs using modified version of the rubric previously used by Nicholls et al., (2014) and Samuel et al., (2015)	Content analysis	119	Funded and unfunded NSF proposals between 2012 and 2014
Van Tuyl and Whitmire (2016)	Evaluation of the compliance of the STEM disciplines' data sharing plans in practice	Content analysis; A review	33	NSF awardees' DMPs at Oregon State University

CLINICAL HEALTH SCIENCES	SURVEY RESEARCH	QUALITATIVE RESEARCH	NATURAL SCIENCES	TEACHERS
Introductory Lecture: Background and concepts; Characteristics of a high quality research plan; Course practicalities				Head of library services; Data librarian; Grant writer
Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Researchers; Lectors; University teachers
Data management plan (DMP): data life cycle; DMP-tool; roles; resources; metadata;documentation				Self-study module; Data librarian
IPR, agreements and licenses (in Finnish)		IPR, agreements and licenses (in English)		Head of legal affairs unit; Assistant legal counsel; Data librarian
Data privacy: privacy notice; risk analysis; (in Finnish)		Data privacy: privacy notice; risk analysis; anonymization (in English)		Data protection officer; Data archive specialist
RedCap: building a form based database	RedCap: building a survey form	NVivo: organizing and coding data	RedCap: building a form based database	Head of biostatistician team; Lector
Data storage, protection, processing, describing and IT Service solutions				IT system architect
Data preservation, sharing and citing; Open data repositories				Data librarian
A voluntary Q&A Session				Module teachers
DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	Head of library services; Data librarian
a general level feedback on DMPs				Head of library services

Figure 1. The structure, contents, and teachers of the 16-week, four-track BRDM course

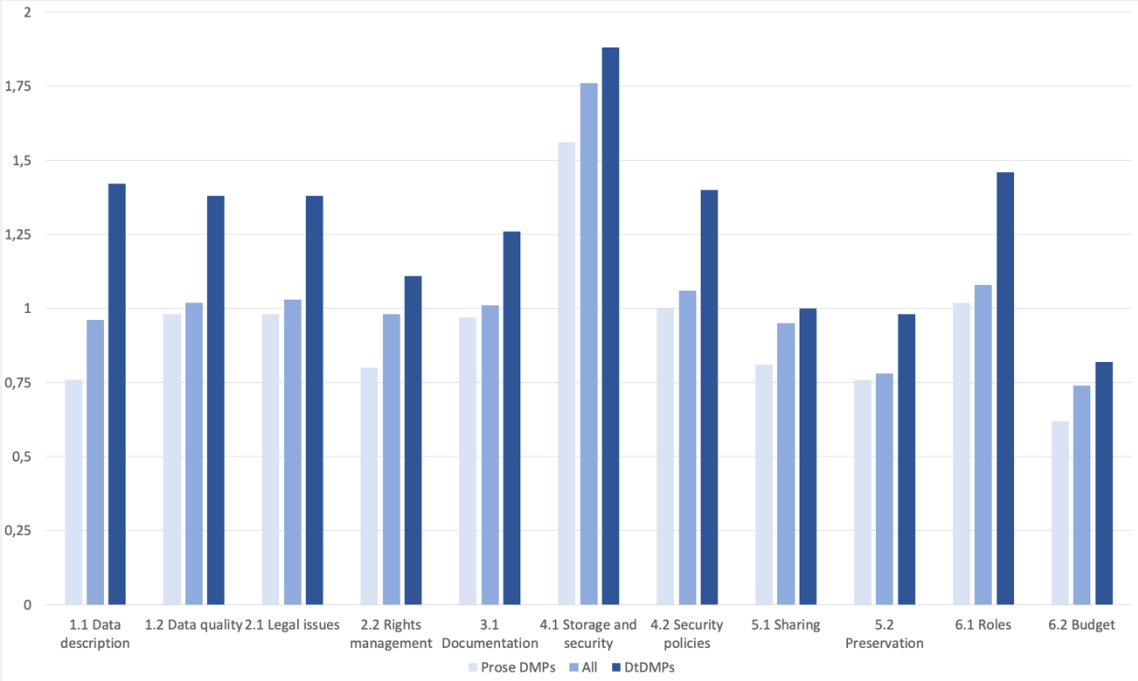

Figure 2. Median ratings of the DMP sections of all DMPs (n=223), DtDMPs (n=100), and prose DMPs (n=123). See the titles of the sections in full in Table 6 in The Appendix

Table 2: An example of the assessment of rights management measures in different types of projects according to the FDEG criteria

	Poor	Sufficient / Satisfactory	Good / Excellent
Consortium research project	Author does not specify participants' roles, responsibilities, and ownership, and how legal and ethical aspects affect the measures.	Author generally describes how relevant legal and ethical aspects are considered in the project without detailing them, and state that roles, responsibilities, and ownership have been or will be specified in a consortium agreement.	Author clearly specify what legal and ethical aspects concern the data, how they are considered in the project, and describe how different participants' roles, responsibilities, and ownership are detailed in a consortium agreement.
Single researcher's project	Author does not describe who or what entity owns the data and what and how legal/ethical aspects are considered.	Author identify the owner, relevant laws and ethical aspects that concern the project's data and generally describe how they are considered, or justify why they do not concern this project's data.	Author identifies the owner/rights holder and justifies it. In addition, they identify and specify how relevant laws and ethical aspects are considered and taken into account in data handling, for example, in various agreements.

Table 3. RDM best practices

Category	Definition
Controller	Data controller named
Data table	Data type specific classification included in a DMP
Detailed descriptions	Descriptions of RDM practices are detailed
Funder's policy	Explained funder's or publisher's data sharing policy
Legal basis	Stated legal basis for handling personal data
License for reuse	Named a license for data reuse
Open data	At least part of the data will be opened
Open metadata	Metadata will be opened
Ownership	The ownership is clearly described and justified
Permission	Permission asked for data sharing and reuse
Resources	Evaluated and described resources needed
Secure storing	Used only secure storing of personal data
Storing by types	A data type specific storing platform
Usage rights	Specified collaborators' different rights of use
Why closed	If not shared the data, justification mentioned

Table 4. An example of a data table

Data type and source	File format	Personal or sensitive data	Ownership and agreements	Metadata documentation	Storage during project	Sharing data after the project	Long-term archiving	Estimated size

Lab notes (Data produced)	.doc .txt .pdf	Yes. Subject to IPR check	PI and group	Programme generates metadata by itself	Electronic lab notebook (eLabJournal)	Project team	Discarded after 15 years	< 10 MB
RNA sequences (Data produced)	raw: FASTA, BAM, .xlsx	no	PI	Readme.txt	UTU's network drive and cloud	European Nucleotide Archive	no	< 1 GB
MRI images (Data reused)	DICOM, .nii, .tiff	Yes, record keeper: xx	PI	Readme.txt	Database x at TYKS, backup	NITRC after anonymization	no	< 1 GB
Question-naire forms (Data collected)	Paper forms	Yes, record keeper: xx	PI	Readme.txt	Locked filing cabinets in PI's office	No, metadata shared in Zenodo/Etsin	Discarded 5 years after publication	

Table 5. Medians, Custom quantiles, and p-values of all DMPs, prose DMPs, and DtDMP

DMP Section	Median, all	Custom quantiles Q1; Q3	Median, prose DMPs	Q1; Q3	Median, DtDMPs	Q1; Q3	p-value, Wilcoxon Rank Sum test
1.1 Data description	0.96	0.84; 1	0.76	0.75; 0.78	1.42	1.28; 1.50	<0.0001
1.2 Data quality	1.02	0.99; 1.05	0.98	0.97; 1	1.38	1.1; 1.48	<0.0001
2.1 Legal issues	1.03	1; 1.07	0.98	0.94; 1	1.38	1.1; 1.49	<0.0001
2.2 Rights management	0.98	0.95; 1.02	0.8	0.76; 0.95	1.11	1.04; 1.43	<0.0001
3.1 Documentation and metadata	1.01	0.99; 1.05	0.97	0.92; 1	1.26	1.06; 1.45	<0.0001
4.1 Storage and security	1.76	1.55; 1.90	1.56	1.34; 1.78	1.88	1.75; 1.92	0.02
4.2 Related data security policies	1.06	1.02; 1.24	1	0.96; 1.05	1.4	1.18; 1.54	0.0005
5.1 Data sharing	0.96	0.85; 0.99	0.81	0.76; 0.96	1	0.95; 1.07	<0.0001
5.2 Preservation	0.78	0.76; 0.98	0.76	0.72; 0.80	0.98	0.94; 1.02	<0.0001
6.1 Roles and responsibilities	1.08	1.04; 1.24	1.02	0.98; 1.06	1.46	1.20; 1.64	<0.0001
6.2 Budgeting and resourcing	0.74	0.70; 0.78	0.62	0.51; 0.76	0.82	0.74; 0.95	0.004
Median sections	1.01	0.98; 1.05	0.98	0.79; 0.99	1.09	1.04; 1.43	<0.0001

Green = good/excellent

Yellow = sufficient/satisfactory

$p < 0.05$ = statistically significant difference

$p < 0.01$ = statistically highly significant difference between prose and DtDMPs

Red = Poor

Table 6. Data sharing venues by disciplines and course tracks (frequencies and Pearson's Chi-Square test)

	Discipline specific repository	Generalist repository	Publication	Upon request	Method not specified	Supple-ment	File sharing service	Other method	Not shared
Discipline									
HS	19 %	16 %	29 %	19 %	10 %	5 %	0 %	0 %	27 %
HPT	23 %	14 %	9 %	23 %	23 %	0 %	0 %	0 %	23 %
SE	17 %	37 %	28 %	13 %	7 %	3 %	10 %	0 %	15 %
SSBE	40 %	14 %	9 %	4 %	12 %	1 %	3 %	3 %	27 %
	p=0.007	p=0.005	p=0.004	p=0.02	p=0.21	p=0.51	p=0.01	p=0.29	p=0.34
Course Track									
CHSct	13 %	18 %	24 %	13 %	8 %	0 %	3 %	0 %	34 %
QRct	37 %	12 %	11 %	9 %	11 %	1 %	1 %	0 %	29 %
NSct	20 %	35 %	32 %	20 %	11 %	6 %	9 %	0 %	6 %
SRct	27 %	16 %	14 %	8 %	14 %	3 %	0 %	5 %	30 %
	p=0.02	p=0.007	p=0.01	p=0.18	p=0.89	p=0.20	p=0.04	p=0.02	p=0.001

Discipline:	Course Track:
HS=Health Sciences	CHSct=Clinical Health Sciences
HPT=Humanities, Psychology, Theology	QRct=Qualitative Research
SE=Science and Engineering	NSct=Natural Sciences
SSBE=Social Sciences, Business, Economics	SRct=Survey Research

Red color = difference is statistically highly significant

Yellow color = difference is statistically significant

Table 7. Data storing venues by disciplines and course tracks (frequencies and Chi-Square tests)

	Institutional net drive or cloud	PC or laptop	External hard/USB drive	Shared work space	Software	Lab or depart-ment's data server or computer	Comm-ercial cloud	CSC	Other method
Discipline									
HS	83 %	21 %	24 %	5 %	32 %	16 %	3 %	2 %	2 %
HPT	82 %	23 %	23 %	18 %	5 %	0 %	9 %	0 %	0 %
SE	77 %	28 %	28 %	30 %	10 %	10 %	7 %	5 %	7 %
SSBE	86 %	27 %	23 %	19 %	8 %	0 %	3 %	4 %	1 %
	p=0.58	p=0.75	p=0.89	p=0.004	p=0.0001	p=0.001	p=0.44	p=0.57	p=0.16
Course Track									
CHSct	82 %	11 %	13 %	11 %	42 %	13 %	3 %	0 %	5 %
QRct	82 %	32 %	28 %	21 %	5 %	0 %	6 %	4 %	0 %
NSct	79 %	27 %	37 %	26 %	12 %	14 %	6 %	5 %	3 %
SRct	89 %	22 %	8 %	5 %	14 %	5 %	0 %	3 %	5 %
	p=0.62	p=0.09	p=0.004	p=0.04	p<0.0001	p=0.005	p=0.40	p=0.62	p=0.23

Table 8. Data preserving venues (frequencies and Pearson's Chi-Square test)

	Institutional net drive, cloud or server	Discipline specific archive	Generalist archive	Method not specified	External hard/USB drive	CSC	Institutional archive	Other method	PC of laptop	Software	Commercial cloud
Discipline											
HS	40 %	13 %	13 %	22 %	8 %	5 %	3 %	6 %	2 %	3 %	2 %
HPT	23 %	32 %	5 %	23 %	5 %	9 %	18 %	0 %	5 %	0 %	5 %
SE	33 %	15 %	40 %	15 %	5 %	10 %	2 %	2 %	2 %	0 %	2 %
SSBE	23 %	40 %	14 %	15 %	9 %	6 %	4 %	3 %	3 %	0 %	0 %
	p=0.15	p=0.0004	p<0.0001	p=0.61	p=0.78	p=0.69	p=0.01	p=0.33	p=0.85	p=0.16	p=0.42
Course Track											
CHSct	50 %	8 %	16 %	18 %	8 %	0 %	5 %	8 %	3 %	5 %	0 %
QRct	23 %	38 %	10 %	20 %	6 %	6 %	7 %	0 %	4 %	0 %	1 %
NSct	27 %	15 %	36 %	20 %	9 %	12 %	2 %	2 %	2 %	0 %	2 %
SRct	32 %	30 %	16 %	11 %	5 %	8 %	3 %	8 %	0 %	0 %	3 %
	p=0.03	p=0.0007	p=0.0006	p=0.67	p=0.87	p=0.14	p=0.36	p=0.03	p=0.62	p=0.02	p=0.79

Table 9. Data preserving time frame by disciplines and course tracks (frequencies and Pearson's Chi-Square test)

	Preserving during verification period (e.g., 5-15 years)	Preserving time frame not specified	Archived for potential re-use (e.g., 25 years)	Archived for tens or hundreds of years	Preserving during the project
Discipline					
HS	60 %	29 %	13 %	2 %	3 %
HPT	36 %	27 %	9 %	18 %	9 %
SE	33 %	35 %	12 %	13 %	7 %
SSBE	31 %	27 %	22 %	13 %	8 %
	p=0.002	p=0.75	p=0.25	p=0.05	p=0.65
Course Track					
CHS	58 %	32 %	13 %	0 %	3 %
QR	26 %	34 %	20 %	11 %	10 %
NS	44 %	30 %	12 %	14 %	3 %
SR	49 %	16 %	14 %	14 %	8 %
	p=0.003	p=0.25	p=0.60	p=0.13	p=0.27