<HMC> HELMHOLTZ
Metadata Collaboration

A survey on the status quo, gaps and needs among research data professionals in the Helmholtz Association

April 2025

HELMHOLTZ

Research for grand challenges.

Call for Review

You are all invited to comment on this version. Please send your feedback by email to info@helmholtz-metadaten.de.

Version: 1.0

This document was generated in a FAIR manner. All previous versions are available on request.

Authors (ORCID): Silke Christine Gerlich (0000-0003-3043-5657), Markus Kubin* (0000-0002-2209-9385), Lucas Kulla* (0000-0002-2484-2742), Christine Lemster (0000-0001-7764-1517), Andreas Schmidt (0000-0002-9911-5881), Jan Schweikert (0000-0003-4774-2717), Sangeetha Shankar* (0000-0003-0387-7740), Karl-Uwe Stucky (0000-0002-0065-0762)

These authors contributed equally to this work.

HMC group: Working group "Task force Survey"

Editors: Sören Lorenz, Witold Arndt, Oonagh Brendike-Mannix, Constanze Curdt, Volker Hofmann, Thomas Jejkal, Marco Nolden, Emmanuel Söding, Wolfgang Süß

Contact: HMC Office
GEOMAR Helmholtz Centre for Ocean Research Kiel
Wischhofstr. 1-3
24148 Kiel, GERMANY
E-mail: info@helmholtz-metadaten.de
www.helmholtz-metadaten.de

HELMHOLTZ

Content

HELMHOLTZ

Abstract

With the HMC Data Professionals Survey 2024, the Helmholtz Metadata Collaboration (HMC) assessed the current state, challenges and requirements of Helmholtz staff in the area of research data management (RDM) and FAIR data practices. The survey focused on data professionals involved in RDM-related tasks in the Helmholtz Association. A total of 156 valid responses were collected from 16 out of 18 Helmholtz centres and all six Helmholtz research fields. This diversity ensures that the survey captures a representative range of perspectives. The results show that RDM work is spread across different stakeholders, professions and organisational levels. While most respondents are familiar with their centre's data policy, the majority have no formal training in RDM - although the data reflects widespread interest in RDM-related training. The results provide a snapshot into the implementation of the FAIR principles in the Helmholtz data ecosystem, and the collaborative use and development of various RDM tools. In particular, the survey highlights the prevalent challenges in RDM work, with many respondents reporting limited resources, gaps in technical knowledge, lack of infrastructure and lack of technical solutions. The survey results, complemented by qualitative feedback from focus groups, will guide HMC's strategic initiatives to improve the FAIR data ecosystem within the Helmholtz Association.

Kurzfassung

Mit dem HMC Data Professionals Survey 2024 hat die Helmholtz Metadata Collaboration (HMC) den aktuellen Stand, die Herausforderungen und die Bedarfe von Helmholtz-Mitarbeitenden im Bereich des Forschungsdatenmanagements (FDM) und der FAIR-Data-Praktiken erhoben. Die Umfrage konzentrierte sich auf 'Data Professionals' mit FDM-bezogenen Aufgaben in der Helmholtz Gemeinschaft. Sie schließt 156 vollständige Antworten aus 16 der 18 Helmholtz-Zentren sowie allen sechs Helmholtz-Forschungsbereiche ein. Dies stellt sicher, dass die Umfrage ein möglichst repräsentatives Spektrum von Perspektiven erfasst. Die Resultate zeigen, dass FDM-bezogene Aufgaben über verschiedene Akteure, Berufsgruppen und Organisationsebenen verteilt sind. Während die meisten Befragten mit den Daten-Policies ihres Zentrums vertraut sind, hatte die Mehrheit bisher keine formale Schulung in FDM, wobei die Umfragedaten ein großes Interesse an FDM-bezogenen Schulungen widerspiegeln. Die Ergebnisse geben einen Einblick in die Umsetzung der FAIR-Prinzipien im Helmholtz-Datenökosystem und in die kollaborative Nutzung und Entwicklung verschiedener FDM-Tools. Die Umfrage verdeutlicht insbesondere die Herausforderungen in Aufgabenbereichen des FDM, wobei viele der Befragten von begrenzten Ressourcen, Lücken bezüglich ihres Fachwissens, mangelnder Infrastruktur und fehlenden technischen Lösungen berichten. Die Ergebnisse der Umfrage, ergänzt durch qualitatives Feedback aus Fokusgruppen, fließen in die strategische Planung von HMC zur Verbesserung des FAIR-Data-Ökosystems innerhalb der Helmholtz-Gemeinschaft ein.

HELMHOLTZ

Introduction

Helmholtz Metadata Collaboration (HMC)

The Helmholtz Metadata Collaboration (HMC) platform enables researchers and data infrastructure providers to make their research data FAIR[1] - findable, accessible, interoperable, and reusable - facilitating data driven science across the Helmholtz Association and beyond. Guided by the FAIR principles, HMC empowers the six Helmholtz research fields - Aeronautics, Space, and Transport (AST); Earth and Environment; Energy; Health; Information; and Matter - to develop, share and consolidate community expertise in metadata together.

HMC aims to create a FAIR data space[2] for research data generated by the Helmholtz Association. Its roadmap comprises measures for monitoring and assessing FAIR (meta)data practices in the Helmholtz Association, for facilitating data connectivity and for transforming recommendations into implementation. With its interdisciplinary mission, structure, and services, HMC is unique in the European science community and is expected to raise the significance and practical use of metadata for research to an unprecedented level.

The Helmholtz Association

The Helmholtz Association of German Research Centers (Figure 1) is Germany's framework for federal, large-scale research facilities that contribute to solving major challenges facing society, science, and the economy through top-level scientific achievements in its six research fields.

To tackle the challenges of data-driven research, the Helmholtz Information and Data Science Framework was set up, where five platforms and numerous projects address these challenges and develop new digital solutions for cutting-edge research. The Helmholtz Metadata Collaboration is one of the platforms.

Figure 1: Helmholtz centers in Germany.

---

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

[2] Nagel, L., and Lycklama, D. Design Principles for Data Spaces - Position Paper (2021), Zenodo. https://doi.org/10.5281/zenodo.5105743

HMC Data Professionals Survey 2024

As part of its monitoring and assessment of FAIR (meta)data practices in the Helmholtz Association, and with the goal of identifying the status quo, gaps and needs of Helmholtz staff with respect to research data management (RDM) and FAIR data practices, HMC develops and runs dedicated community surveys.

In 2021 HMC ran the HMC Community Survey 2021 to address these aspects from the point of view of researchers throughout the organization.[3] The survey results were used as input to inform HMC's strategic planning by identifying gaps and needs, both general and research field-specific, allowing for designing and improving services to advance the Helmholtz FAIR data landscape.

In 2024 HMC conducted a complementary survey, the 'HMC Data Professionals Survey 2024'. With this HMC targeted 'data professionals', Helmholtz staff working on RDM-related tasks, and investigated the status quo, gaps and needs with respect to RDM and FAIR data practices in Helmholtz, as expressed by this group.

In the current document, we present the results of this HMC Data Professionals Survey 2024. These results inform the strategic planning of HMC by identifying gaps, thereby enabling the design and improvement of tools and services to enhance the FAIR data landscape in the Helmholtz Association.

Preliminary survey results were discussed with the Helmholtz data professionals community in a December 2024 workshop and insights from these focus groups were collected to qualitatively enrich the survey data, as included in the this report.

---

[3] Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022). A survey on research data management practices among researchers in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_05

Survey Design and Analysis

Target group

The HMC Data Professionals Survey 2024 was targeted at all staff of the Helmholtz Association of German Research Centers who work on tasks that relate to the management of research data. The target group includes Helmholtz staff working on one or more of the following RDM-related tasks.

- Research data management strategy for the Helmholtz centre or institute

- Provision of research data infrastructure

- Technical support for research data infrastructure

- Creation of data management plans (DMP)

- Development of software for (meta)data collection and/or processing

- Enrichment of data and/or software with metadata

- Publication of data and/or software in a FAIR-compliant way

- Identification of data management needs

- Networking with other RDM-related personnel and other relevant organisations

- Training of employees on RDM topics

- Provision of long-term data archiving solutions

- Technical support for long-term data archiving solutions

- Check compliance with policies and guidelines

- Consultation on RDM topics

- Other tasks related to the management of research data

This target group was expected to be comprised of research data managers, research data professionals and data stewards, librarians, software developers and researchers. They were expected to have varying levels of expertise and dedicate different amounts of their working time to RDM activities.

Survey goals

The overarching goals of the HMC Data Professionals Survey 2024 were (1) to identify research data professionals working in the Helmholtz Association; (2) to analyze the status quo, gaps and needs expressed by data professionals with respect to their RDM and FAIR data practices; and (3) to derive strategic measures, to be potentially addressed by HMC.

The following goals and sub-goals were formulated to address the overarching goal in a holistic manner:

1. Identify and connect to RDM communities in the Helmholtz Association.

    1.1. What is their role in Helmholtz?

    1.2. What RDM-related tasks are they working on?

HELMHOLTZ

      1.3.   Have they had useful training for these tasks?

2. Identify the status quo of RDM in Helmholtz.

      2.1.   Self-assessment of FAIR data practices.

      2.2.   Identify tools used in the centers and across centers.

3. Identify action items for HMC.

      3.1.   Identify successful tools that might be enhanced by HMC.

      3.2.   Identify gaps, problems and needs mentioned by RDM-related personnel.

      3.3.   Suggest priorities for action items based on quantitative analysis.

      3.4.   Identify factors limiting progress towards a FAIR data landscape in Helmholtz.

Working hypotheses

The following working hypotheses were formulated to be probed by the survey:

1. RDM-related work is done on different institutional levels (center, institute(s), departments, research groups) within Helmholtz.

2. Work packages among RDM-related personnel are streamlined across the centers.

3. At least 50% of staff with RDM-related tasks have had a training for these tasks.

4. At least 50% of staff with RDM-related tasks in Helmholtz align their work with the FAIR data principles.

5. At least 50% of staff with RDM-related tasks align their work based on the policies and guidelines in their center.

6. Tools commonly used by RDM-related personnel are also used by others in the respective communities.

7. At least 50% of staff with RDM-related tasks report that (technical) infrastructure for implementing the FAIR data principles is available.

8. RDM-related personnel have gaps and needs with respect to the FAIR compliance of data.

9. RDM-related personnel are interested in the services offered by HMC.

Survey questionnaire

A survey questionnaire was developed on the basis of the survey goals and hypotheses listed above. All survey questions and their logical dependencies are shown in Figure 2. Target group-oriented language was chosen for designing the questionnaire to allow participants of all expertise levels to understand questions and answers. In this questionnaire, questions PERBG1, PERBG2, RDMPR5A, SERVC3 and SERVC10 were reused or adapted from questions in the HMC Community Survey 2021. [4]

---

[4] Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022). A survey on research data management practices among researchers in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_05

Personal background

entry to survey

PERBG1
Which Helmholtz center do you typically work in? * (SC)

PERBG2
Please select the Helmholtz research field(s) you associate yourself with. * (MC)

PERBG11
Which of the following titles best matches your current job? * (SC)

PERBG13
Which of the following research data management (RDM) tasks do you work on? * (MC)

≥ 1 task selected   no RDM-related task selected

PERBG13A
Which organizational level(s) do you address with your research data management tasks? (MC)

PERBG13B
On average, what percentage of your working time do you spend on RDM-related tasks? (numeric)

PERBG14
Have you had any training for your tasks in research data management? (SC)   — no —

subset of selected tasks

yes

PERBG14A
Did the training help you in accomplishing the following tasks? (Matrix)

Research data management practices

RDMPR54
Please rate the following statements concerning the data(sets) you process or make publicly accessible. (Matrix)

RDMPR23
The data(sets) are assigned a globally unique, persistent and resolvable identifier (PID) when deposited in a digital repository.

RDMPR24
When data(sets) are deposited in a digital repository, descriptive metadata (creator, title, description, etc.) is provided to make the data(set) findable, understandable and reusable to others.

RDMPR25
The data(sets) are deposited in a repository, which provides descriptive metadata (creator, title, description, etc.) in both human-readable and machine-readable ways.

RDMPR26
Access to the data(sets) is controlled, while ensuring that the data is as open as possible and as closed as necessary.

RDMPR27
Metadata includes license information under which the data can be reused.

RDMPR28
Metadata of the data(sets) remains available over time, even if the data(set) is no longer accessible.

RDMPR29
The metadata describing the data(sets) is based on controlled vocabularies (e.g. taxonomies, ontologies).

RDMPR5A
Internationally used metadata standards, templates and schemas are used to describe the data.

RDMPR30
The metadata describing the data(sets) includes provenance information about the collection or generation of data.

RDMPR31
The metadata describing the data(sets) is based on the specifications of a community-endorsed standard.

RDMPR32
The data(sets) use an open file format.

RDMPR33
Curation of the data is assured.

RDMPR34
Digital long-term archival is assured for the data(sets).

RDMPR41
Are you aware of any official guidelines or policies, provided by your organization on how to manage research data? (SC)

no   yes

RDMPR53
How much do you agree with the following statements? (Matrix)

RDMPR42
The guidelines or policies are easy to understand.

RDMPR43
Assistance for policy implementation is sufficiently provided by my organization for all projects I am involved in.

RDMPR44
I fully align my work to these policies and guidelines.

RDMPR45
Compliance with these guidelines is sufficiently audited within my organization for all projects I am involved in.

RDMPR46
Do you align your work with other third-party policies or guidelines? (SC + free-text)

Service formats

SERVC10
Which service formats would you be interested in? Please specify topics. (MC + free text)

SERVC3
You are almost there! You are welcome to formulate questions, wishes or suggestions in the following free-text field. (free text)

end of survey

RDMPR47
Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work.

RDMPR48
Which of the following tool categories are (co)developed in your organizational level/unit? (MC + free text)

RDMPR51
What obstacles or difficulties do you encounter when working on research data management-related tasks? (MC + free text)
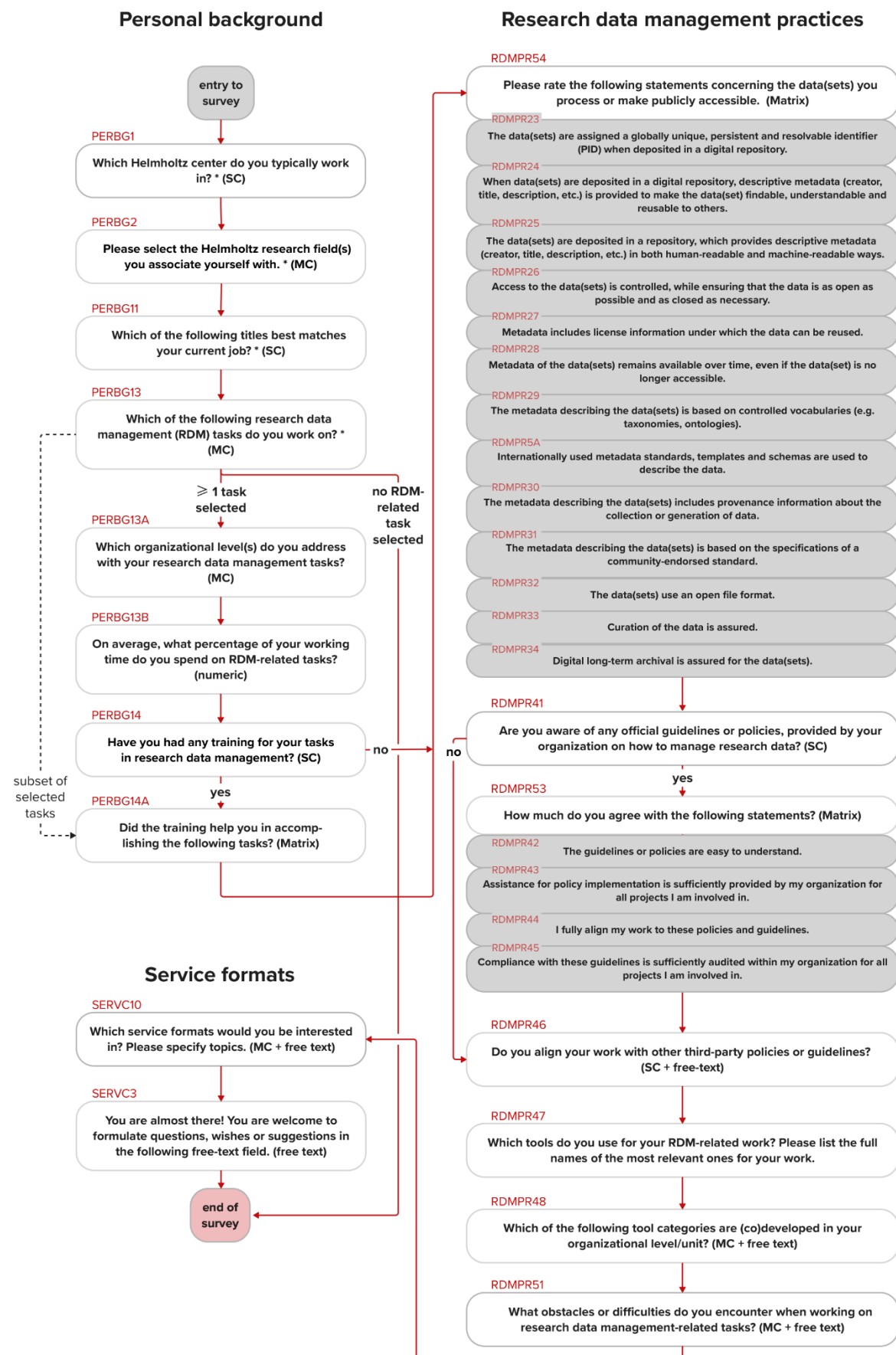
Figure 2: Sequence of survey questions. Mandatory questions are marked with asterisks, question codes in red, single- and multiple choice as "SC" and "MC", respectively.

HELMHOLTZ

Question types

The following types of questions were included in the survey questionnaire:

- In "single choice" (SC) questions respondents were allowed to choose up to one answer option from a predefined list of answers.

- In "multiple choice" (MC) questions respondents were allowed to choose a set of answers from a predefined list, ranging from none to all answer options. Some multiple choice questions contained exclusive answer options, for example, "I do not work on RDM-related tasks". Upon selecting these options, all other answer options were unchecked and disabled.

- In "array" questions, also known as "matrix" questions, for each answer category (sub-question) respondents could select an answer option from a predefined list in a single-choice manner.

- In "slider" questions respondents could choose a numeric value from a predefined range of numbers by moving a "slider" on the screen.

- In "free text" answers respondents could enter text into one or more input field(s).

Question groups

The question group "Personal Background" (question code PERBG) aims at understanding the background of the respondents such as the Helmholtz center they are employed at, their research field(s), their job title and the activities related to research data management they are working on.

The question group "Research Data Management Practices" (question code RDMPR) aims at understanding respondents' RDM and FAIR data practices. A central element of this question group is a matrix question designed for the participants' self-assessment of FAIR data practices at their center. These questions (see Figure 2, RDMPR54 and following questions) are derived from the FAIR-Aware questionnaire, as developed within the European FAIRsFAIR project[5]. Further questions in this question group address the respondents' awareness of data policies at their research centers, RDM-related tools used by the respondents, and tools (co-)developed at their centers, as well as the challenges faced in respondents' RDM-related work. Overall, this question group aims at capturing the status quo of the FAIR data landscape in Helmholtz, while the identification of challenges can reveal potential areas of improvement.

The question group "Services" (question code SERVC) aims at understanding the respondents' reported need for support in various formats and provides an opportunity to provide general feedback. This question group aims to inform strategic decision making in HMC. Combined with respondent background data, this set of questions can inform the development of tailored RDM solutions and service formats.

Question logics

Question logics were implemented in the questionnaire to dynamically adapt questions to the respondents' previous answers. The final question catalog contains 17 (sub-)questions, grouped into three subject areas (question groups), as described above.

---

[5] Data Archiving and Networked Service (2021). FAIR-Aware: Your first step towards your FAIR data(set). Retrieved January 30, 2025, from https://fairaware.dans.knaw.nl/.

HELMHOLTZ

To tailor the questionnaire to the target group of the survey, question logics were implemented such that respondents who answered "I do not work on RDM-related tasks" to question PERBG13 were directed to the end of the survey.

Implementation and dissemination

The survey questionnaire was implemented using LimeSurvey, a web-based open source software for conducting online surveys.[6] For data protection reasons, two separate instances were hosted and the survey results of both instances were later combined for joint analysis - one instance for general, Helmholtz-wide deployment was run from April 22nd to October 11th, 2024, and a second instance for exclusive deployment at the German Aerospace Center (DLR) was run from August 28th to October 11th, 2024. The setup was chosen such that survey participation was anonymous and voluntary. No incentives were provided for participating in the survey.

The survey was disseminated and advertized via target group specific mailing lists, social media channels, and targeted emails to Helmholtz staff known to be involved in the management of research data.

Data processing and analysis

Raw data was pre-processed in the following way:

1. Data from both survey instances was merged for joint analysis, as presented in this report.

2. Survey replies were omitted if respondents did not submit their survey response on the last page of the questionnaire.

3. Respondents who could not be associated with the target group of this survey were removed based on their answer "I do not work on RDM-related tasks" to question PERBG13.

4. Free text answers were mapped to existing answer categories, wherever applicable. If unforeseen clusters of answer categories were identified therein, this information was mapped to the category "other". Where not applicable, free text answers were mapped to common themes in a two-step process. In the first step, a customized version of the large language model Llama 3.3[7] was used to precategorize common categories. In the second step, these mappings were manually revised by the project team.

Data analysis was performed using an adoption of the Python-based "hifis-surveyval" framework,[8] originally developed by the Helmholtz Federated IT-Services (HIFIS). All questions and answer options were described in the YAML format required by the HIFIS Surveyval framework. Software developed for the analysis of the HMC Community Survey 2021[9] was reused and further customized.

---

[6] LimeSurvey (GitHub repository). Retrieved January 31, 2025, from https://github.com/LimeSurvey/LimeSurvey.

[7] Ollama, Llama 3.3, Retrieved April 15, 2025, from https://ollama.com/library/llama3.3

[8] HIFIS (2023). hifis-surveyval version 1.5.3. Python Package Index. https://pypi.org/project/hifis-surveyval/.

[9] Helmholtz Metadata Collaboration (2022). HMC Community Survey - Analysis (GitLab repository). Retrieved April 15, 2025, from https://codebase.helmholtz.cloud/hmc/hmc-public/surveys/hmc-community-survey-2021/analysis

HELMHOLTZ

Survey Results

Data professionals in Helmholtz

The HMC Data Professionals Survey 2024 was answered by a total of 384 respondents, out of which 200 responses (52.1%) were found to be complete. 184 responses (47.9%) contained no data or were incomplete. Survey replies submitted by 44 respondents could not be associated with the target group of this survey and were removed based on their answer "I do not work on RDM-related tasks". The final data set analyzed in this report contains 156 responses.

The analyzed survey data includes 156 complete survey responses from staff located at the following sixteen out of eighteen Helmholtz centers, which ensures that the survey results capture diverse perspectives and practices across the Helmholtz Association:

- Alfred-Wegener-Institute (AWI)

- Deutsches Elektronen-Synchrotron (DESY)

- German Cancer Research Center (DKFZ)

- German Aerospace Center (DLR)

- German Center for Neurodegenerative Diseases (DZNE)

- Forschungszentrum Jülich (FZJ)

- GEOMAR Helmholtz Centre for Ocean Research Kiel

- German Research Centre for Geosciences (GFZ)

- Helmholtz Centre for Heavy Ion Research (GSI)

- Helmholtz-Zentrum Hereon

- Helmholtz Zentrum München - German Research Center for Environmental Health (HMGU)

- Helmholtz-Zentrum Berlin für Materialien und Energie (HZB)

- Helmholtz-Zentrum Dresden-Rossendorf (HZDR)

- Karlsruhe Institute of Technology (KIT)

- Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)

- Helmholtz Centre for Environmental Research (UFZ)

As shown in Figure 3, relative contributions from these sixteen centers range between 0.64% (1 response) and 17.3% (27 responses) per center.
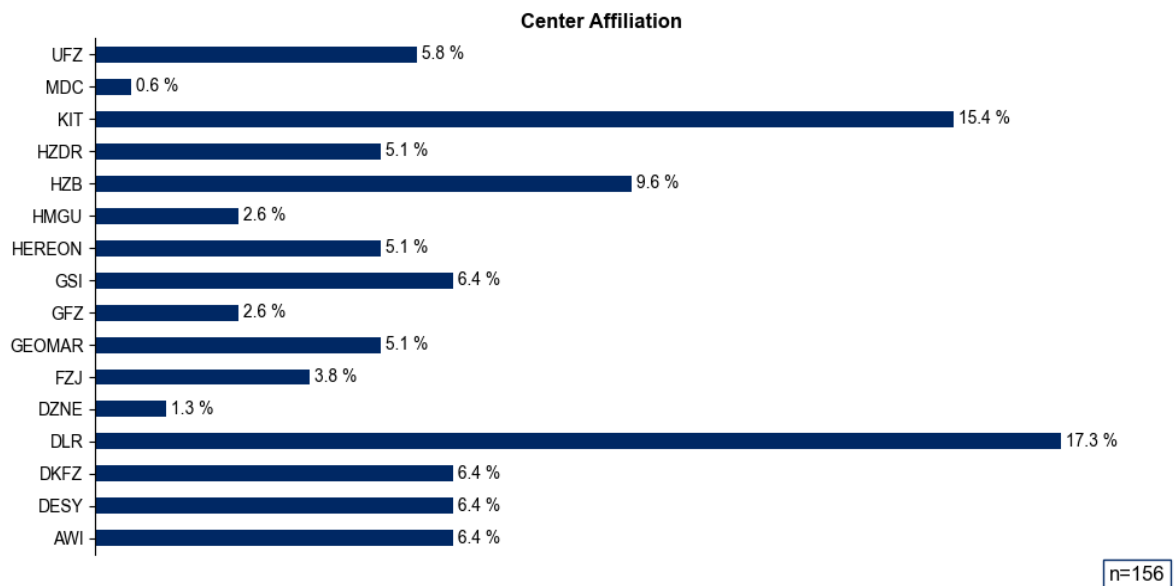
Center Affiliation

Figure 3: "Which Helmholtz center do you typically work in?" (Question type: single choice. Available to all respondents. Number of respondents who answered this question: n = 156, relative amounts refer to n)

The survey data includes responses from all six Helmholtz research fields. As shown in Figure 4, contributions from the six research fields range from 13.5% (Health) to 29.5% (Earth and Environment). While this question was implemented as a multiple choice question, the majority of respondents (76.2%) chose a single research field.
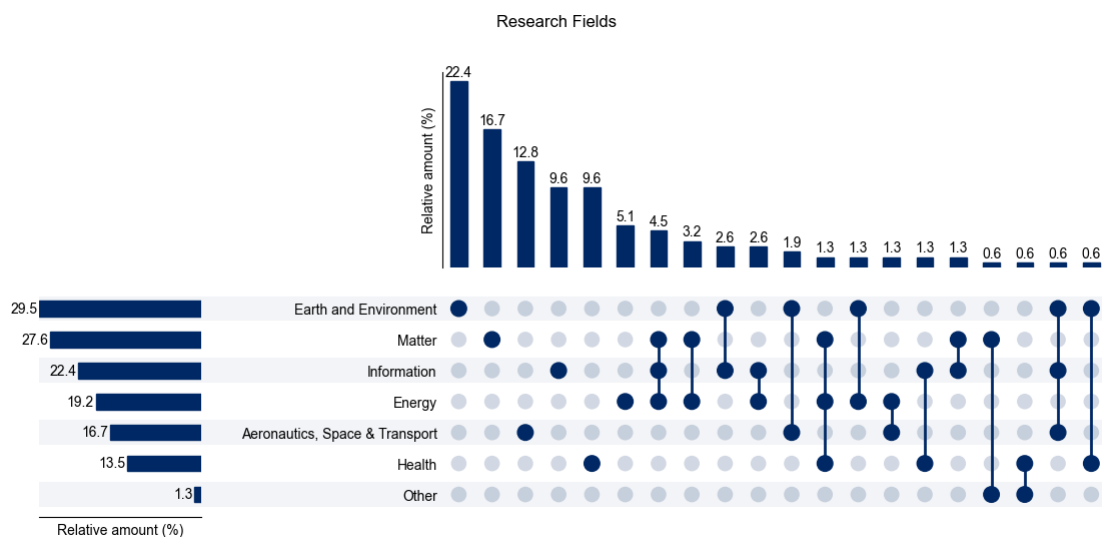
Figure 4: "Please select the Helmholtz research field(s) you associate yourself with." (Question type: multiple choice. Available to all respondents. Number of respondents who answered this question: n = 156, relative amounts refer to n)

With respect to their self-explained job descriptions, the survey respondents represent a diverse cross-section of professionals involved in research data management-related tasks (see Figure 5). With 38.5% researchers, the largest group of respondents, is followed by research data professionals at 30.8%, software developers at 16.7%, and librarians at 2.6%.
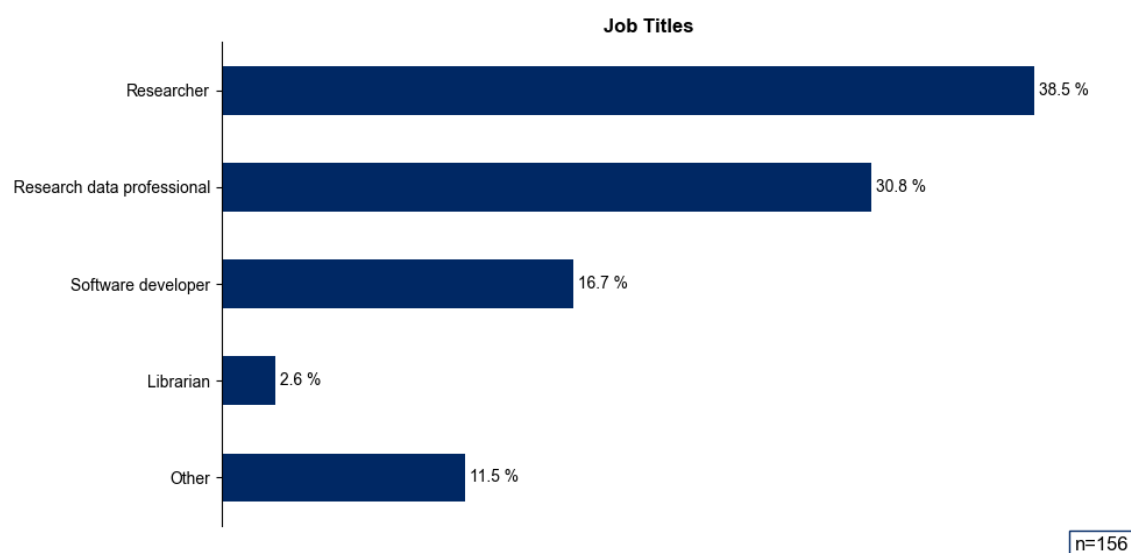
HELMHOLTZ

Job Titles

Researcher	38.5 %
Research data professional	30.8 %
Software developer	16.7 %
Librarian	2.6 %
Other	11.5 %

n=156

Figure 5: "Which of the following titles best matches your current job?" (Question type: single choice. Available to all respondents. Number of respondents who answered this question: n = 156, relative amounts refer to n)

With respect to the main RDM-related activities, the results shown in Figure 6 reveal a consistent pattern across the participating centers. According to the survey data, respondents focus primarily on the following key tasks: metadata annotation (60.3%), FAIR-compliant publication of data and/or software (56.4%), and identification of data management needs (56.4%), followed by the provision of RDM infrastructure (49.4%), software development for metadata collection and/or processing (48.1%) and networking initiatives (47.4%). Staff from 14 to 16 out of 16 centers reported engagement in each of the tasks listed in Figure 6.
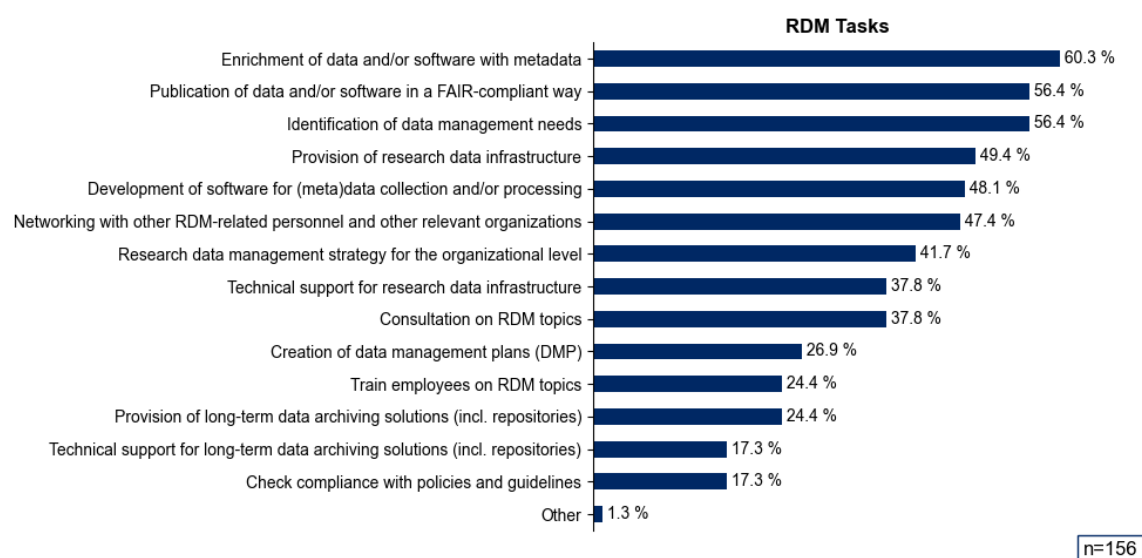
RDM Tasks

Enrichment of data and/or software with metadata	60.3 %
Publication of data and/or software in a FAIR-compliant way	56.4 %
Identification of data management needs	56.4 %
Provision of research data infrastructure	49.4 %
Development of software for (meta)data collection and/or processing	48.1 %
Networking with other RDM-related personnel and other relevant organizations	47.4 %
Research data management strategy for the organizational level	41.7 %
Technical support for research data infrastructure	37.8 %
Consultation on RDM topics	37.8 %
Creation of data management plans (DMP)	26.9 %
Train employees on RDM topics	24.4 %
Provision of long-term data archiving solutions (incl. repositories)	24.4 %
Technical support for long-term data archiving solutions (incl. repositories)	17.3 %
Check compliance with policies and guidelines	17.3 %
Other	1.3 %

n=156

Figure 6: "Which of the following research data management (RDM) tasks do you work on?" (Question type: multiple choice. Available to all respondents. Number of respondents who answered this question: n = 156, relative amounts refer to n)

The survey also addresses the self-reported, relative number of working hours that survey respondents allocate for the RDM-related tasks (see Figure 7). The responses showed a relatively uniform distribution, ranging from 5 to 100%.
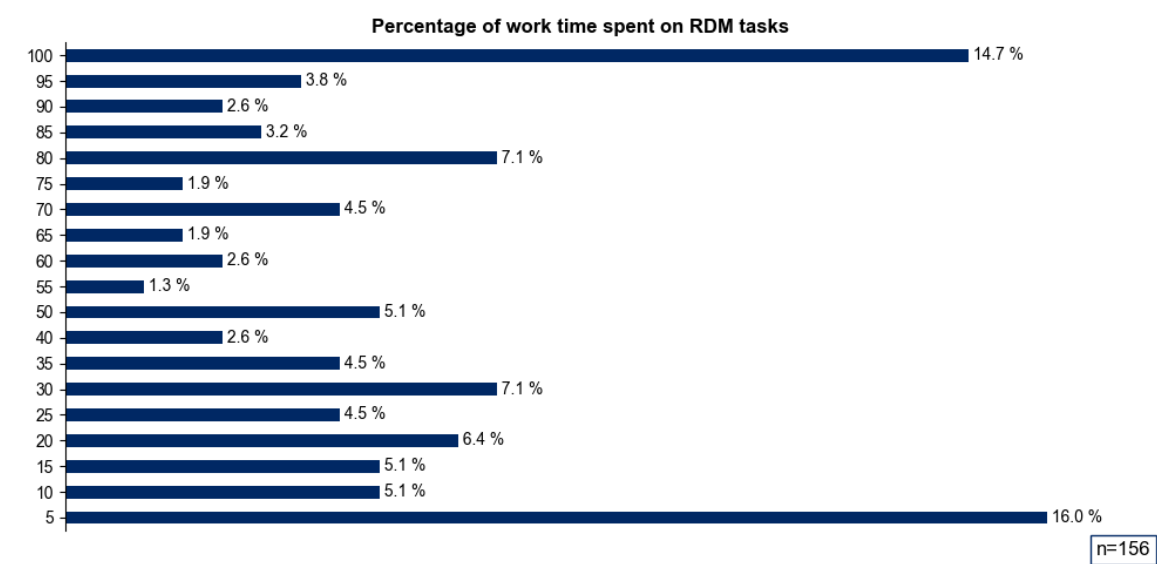
HELMHOLTZ

Figure 7: "On average, what percentage of your working time do you spend on RDM-related tasks?" (Question type: numerical slider. Available to all respondents. Number of respondents who answered this question: n = 156, relative amounts refer to n)

As shown in Figure 8, respondents report to address RDM tasks across various organizational levels, from the division level (65.2%), department level (52.9%), center level (51.6%), to overarching levels (38.1%). Around 65% of survey respondents indicate engagement with multiple organizational levels.

Figure 8: "Which organizational level(s) do you address with your research data management tasks?" (Question type: multiple choice. Available to all respondents. Number of respondents who answered this question: n = 155, relative amounts refer to n)

Data policies and guidelines

The survey data shown in Figure 9 reveals a significant awareness among Helmholtz staff working on RDM-related tasks with respect to official data management policies or guidelines at their centers. The high level of awareness (75.4%) suggests that communication of these documents is relatively well-established.

HELMHOLTZ
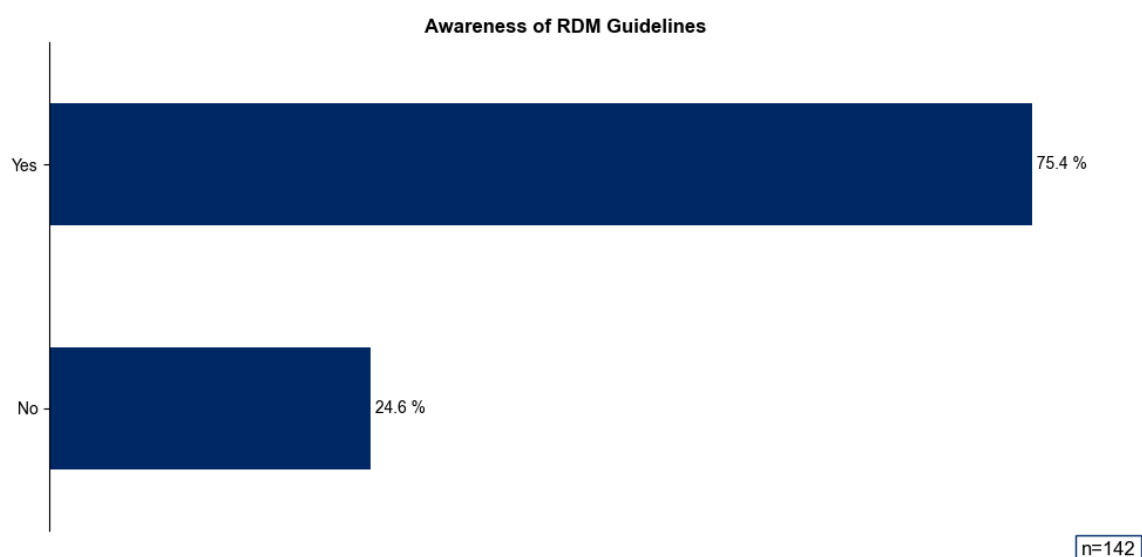
Awareness of RDM Guidelines

Figure 9: "Are you aware of any official guidelines or policies, provided by your organization on how to manage research data?" (Question type: single choice. Available to all respondents. Number of respondents who answered this question: n = 142, relative amounts refer to n)

Those respondents who reported knowing about their center's data management policy or guidelines were asked four follow-up questions. The respective data is shown in Figure 10, where the following paragraph takes into account positive replies by respondents who "rather" or "fully agreed" with the follow-up statements. With 70.4% the majority reports aligning their work with their center's policies and guidelines and 57.1% of respondents find these documents easy to understand. However, only 29.4% believe that sufficient assistance for implementation is provided and only 25.6% agree that compliance is adequately audited.
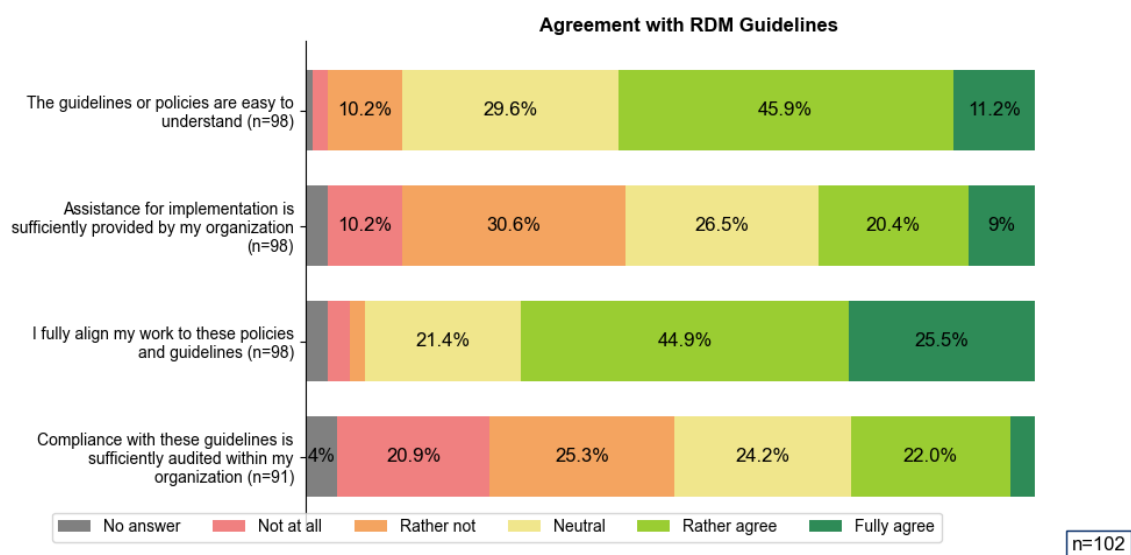
Agreement with RDM Guidelines

Figure 10: "How much do you agree with the following statements?" (Question type: array. Available to all respondents who said they are aware of guidelines or policies at their center. Relative amounts refer to the number of answers n for each subquestion, as shown next to each subquestion title.)

The data shown in Figure 11 reveals a significant amount (54.6%) of respondents across the Helmholtz Association aligning their research data management-related practices to external policies or guidelines.

Third-party guidelines followed
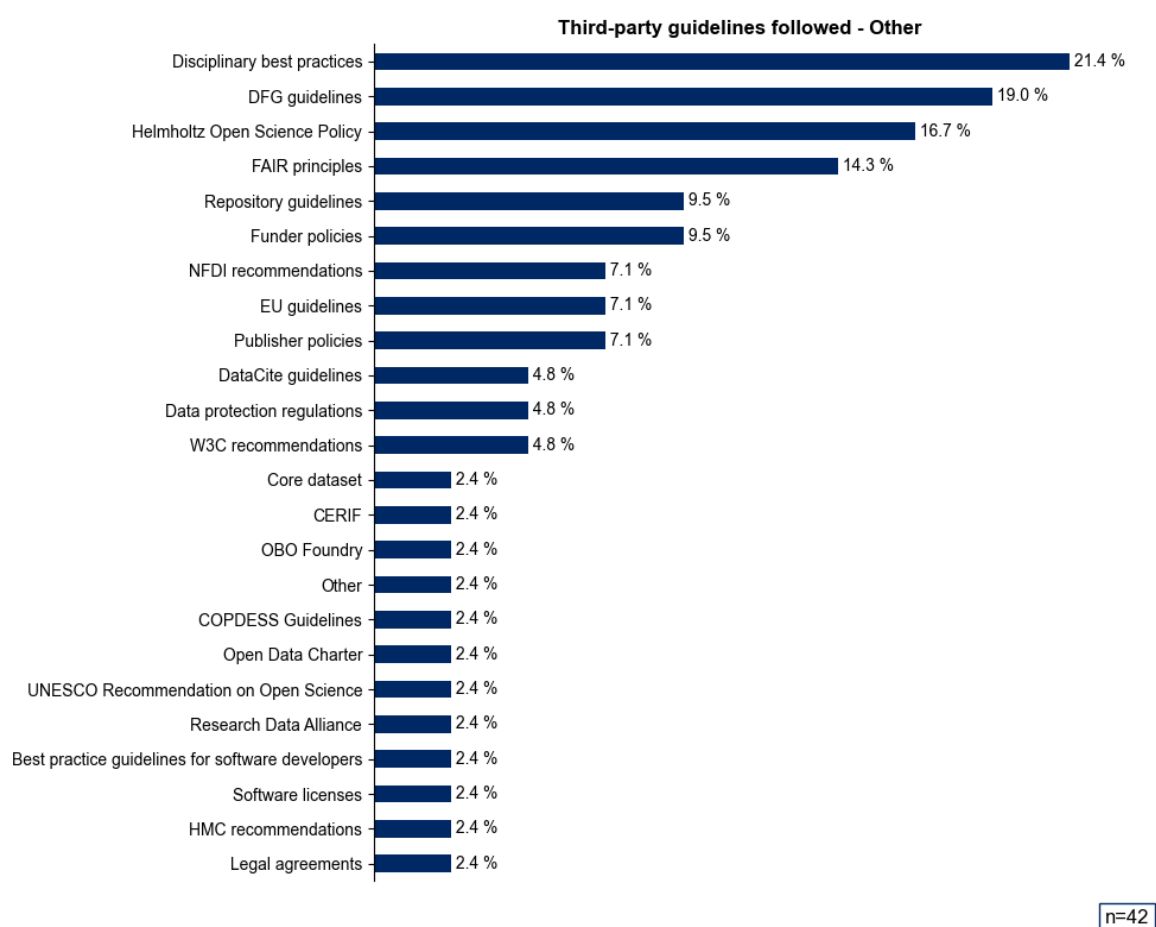

Figure 11: "Do you align your work with other third-party policies or guidelines? Please specify." (Question type: single choice with comments. Available to all respondents. Number of respondents who answered this question: n = 108, relative amounts refer to n)

Figure 12 further reveals a prevailing commitment to disciplinary best practices, followed by DFG guidelines and the Helmholtz Open Science Policy; also, the FAIR data principles, funder policies, NFDI recommendations, EU guidelines, and publisher policies were mentioned.

Third-party guidelines followed - Other

Figure 12: "Do you align your work with other third-party policies or guidelines? Please specify." (Question type: free text (clustered), available to all respondents who answered "yes" in the main question. Number of respondents who answered this question: n = 42, relative amounts refer to n)

HELMHOLTZ

Research data management and FAIR practices

The survey data provides valuable insight into FAIR data practices from the perspective of Helmholtz staff with research data management related tasks. The data shown in Figure 13 indicate variable levels of adherence with respect to specific aspects of the FAIR principles.
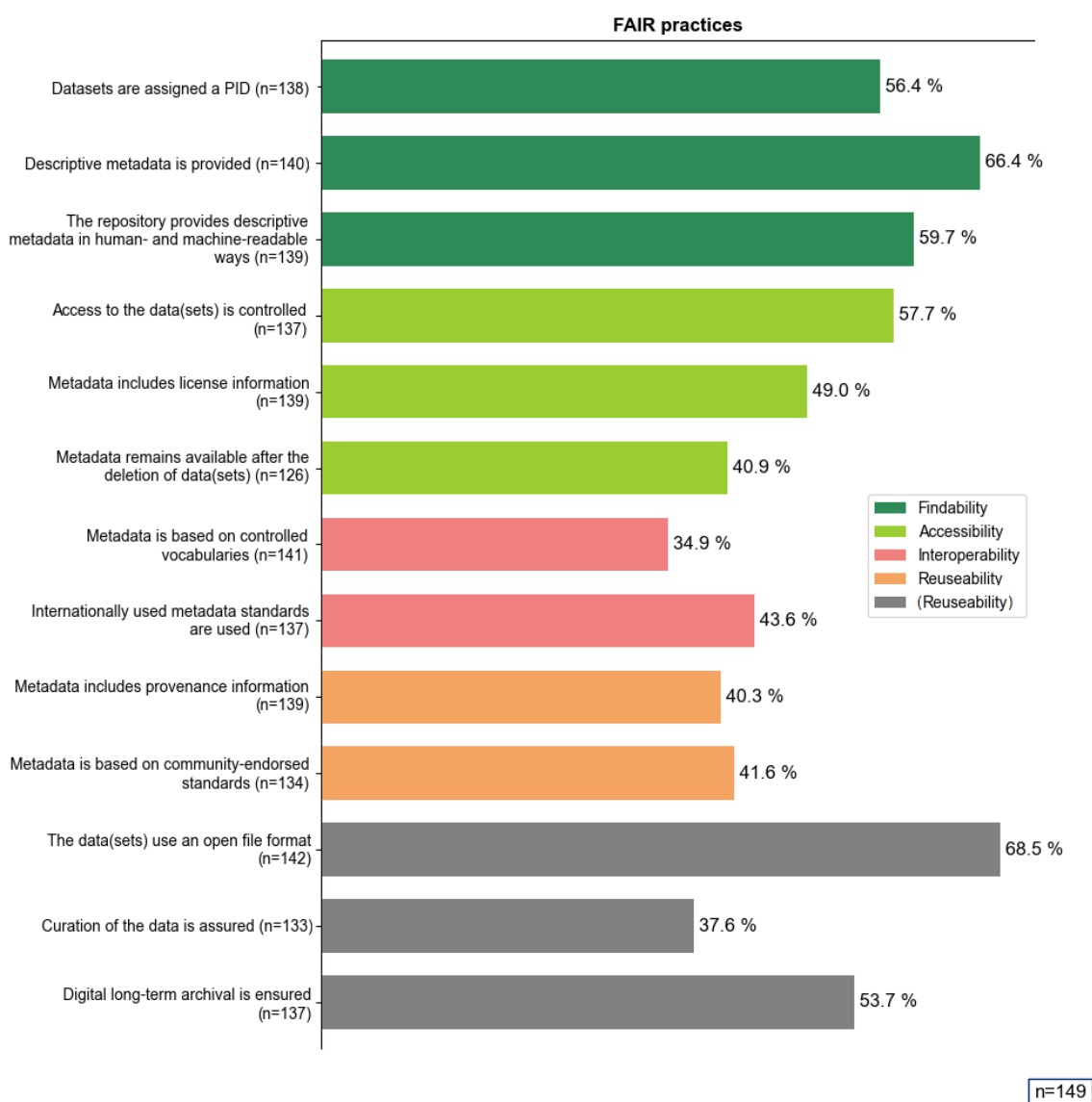

Figure 13: "Please rate the following statements concerning the data(sets) you process or make publicly accessible." (counting answer options "Always" and "Usually") (Question type: array. Available to all respondents. Relative amounts refer to the number of answers n for each subquestion, as shown next to each subquestion title.)

Practices related to the findability aspect of FAIR show rather promising adoption rates: with 66.4% of respondents usually or always providing descriptive metadata to ensure the findability of data, 59.7% usually or always ensuring that repositories provide such descriptive metadata in human and machine-readable ways, and 56.4% usually or always assigning a persistent identifier (PID) to datasets which they process or make publicly available.

Accessibility-related FAIR practices demonstrate moderate adoption rates: with 57.7% of respondents saying that access to the data they process and/or make publicly available is always or usually controlled, 49.0% saying that license information is always or usually included in the

metadata, and only 40.9% express that always or usually metadata is maintained after the deletion of data.

Interoperability-related practices show lower adoption rates: 43.6% of respondents use international metadata standards, and only 34.9% make use of controlled vocabularies in the metadata of data they process or make publicly available.

Similarly, low adoption rates are also observed regarding reusability-related FAIR practices where 41.6% always or usually base metadata on community-endorsed standards and 40.3% always or usually include provenance information in the metadata.

The last three bars at the bottom of Figure 13 show data on reusability-related questions that were adopted from the FAIR-Aware questionnaire and are related to but not directly addressed in the FAIR principles. 68.5% of respondents always or usually use open file formats for the data they process and/or make publicly available. Concerning the long-term management of data, 53.7% always or usually ensure digital long-term archival, and 37.6% report an assured curation plan for the data.

Ecosystem of RDM-related tools

Figure 14 reveals a diverse range of tools used for research data management (RDM) across the Helmholtz Association. Version control software emerges as the most widely adopted tool category, used by 84.0% of respondents, followed by PID services (63.2%) and controlled vocabularies (46.5%). Repository software (32.6%) and electronic lab notebooks (25.0%) show moderate adoption rates. However, lower usage rates are reported for registries (22.2%), vocabulary and ontology editors (19.4%), and data management plan (DMP) software (14.6%).
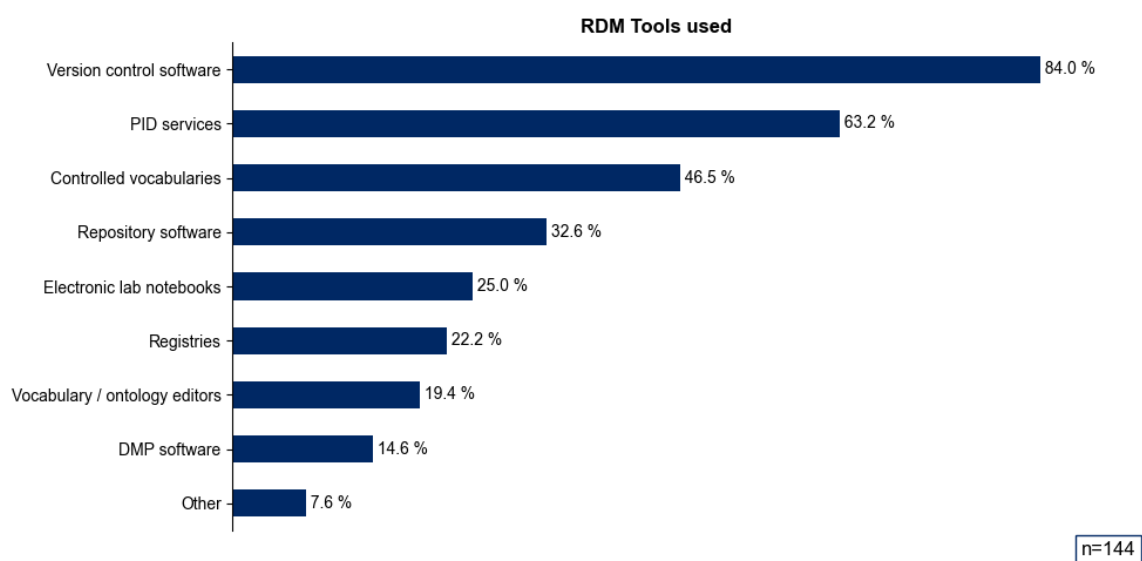

Figure 14: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 144, relative amounts refer to n)

Names of the RDM-related tools most reported to be used by the survey respondents are illustrated in Figure 15.

Figure 15: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Wordcloud-illustration of clustered free-text comments. See the Appendix for concrete survey data.)

Active (co-)development of RDM-related tools is reported for various tool categories (see Figure 16). Electronic lab notebooks, repository software, and controlled vocabularies emerge as the most frequently developed tools, each reported by 34.6% of respondents, revealing a focus on data documentation, data storage, and metadata standardization. Lower rates of active (co-)development are reported for version control software (15.4%), vocabulary and ontology editors (14.1%), PID services (12.8%), DMP software (11.5%), and registries (6.4%).
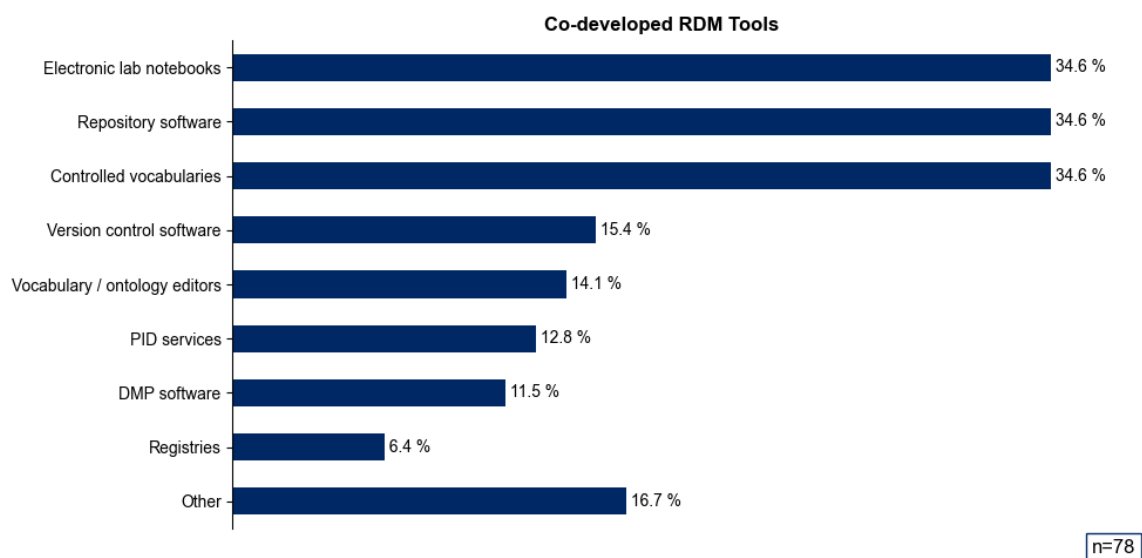

Figure 16: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 78, relative amounts refer to n)

For each category, the tool names most mentioned are listed below. For further detail, please refer to the respective figures in the Appendix.

- Version control: The version control-related software mentioned to be used by most survey respondents is Git (96 out of 156 respondents), followed by Subversion. Respondents report Git and HERMES to be version control software that is (co-)developed by Helmholtz staff.

- PID services: The PID services most mentioned to be used are ORCID and DataCite. Other named by the participants were ROR, EPIC handles, B2INST, Zenodo, IGSN, PIDA, DOI, PID4INST and Crossref. For this category, respondents report that IGSN, PIDA, and DataCite are (co-)developed by Helmholtz staff.

- Controlled vocabularies: A variety of controlled vocabularies is reported to being in use. The vocabularies most mentioned are Dublin Core, schema.org, DataCite, NERC vocabularies, as well as ISO standards. Respondents report NeXus, the Thin-film solar cell ontology (TFSCO), two domain ontologies PolyMat and PolyLab, as well as the Open Geospatial Consortium (OGC) vocabulary to be (co-)developed by Helmholtz staff.

- Repository software: The repository software survey respondents most often reported using is Invenio, followed by Zenodo, Git, SciCat, ICAT, Dataverse, RADAR, GeoNetwork, and Kadi4Mat. Respondents report SciCat, ICAT, and Invenio to be (co-)developed within Helmholtz.

- Electronic lab notebooks: The electronic lab notebooks most mentioned to be used are ElabFTW, Herbie, Chemotion, BioVia, MediaWiki, Elog, Kadi4Mat and NOMAD. It may be noted that also Google Docs and MS Excel are mentioned in this category. Respondents report (co-)development of Herbie, NOMAD, Kadi4Mat, MediaWiki, ELabFTW, the European Synchrotron Radiation Facility (ESRF) logbook, ICAT+, ELog and BioVia.

- Registries: The registries most mentioned to be used are re3data, Zenodo, and fairsharing.org, while (co-)development by Helmholtz staff is reported for re3data and the German Human Genome-Phenome Archive (GHGA).

- Semantic development tools: The vocabulary or ontology editing software mentioned most predominantly is Protégé, but also VocPopuli, Karma, WoRMS, ChEBI, ITIS and VocBench are mentioned for this category. Respondents report (co-)development of VocPopuli, ViMi (Labs), GUIDE, graph4mat, EVOKS, as well as OEP tools.

- DMP software: The DMP software most frequently reported to be used by survey respondents is the Research Data Management Organiser (RDMO), followed by FAIR-Wizard, Moses DMP and Shepard. Respondents report (co-)development of RDMO, but also ViMi Labs and Twinstash within Helmholtz.

Training

Respondents were asked whether they have had any training for their tasks in research data management. The data shown in Figure 17 reveals that 25.5% of respondents report to have received formal training for these tasks, while 74.5% report to have received no training.

Participants who reported positively to have received training were successively asked whether the training helped them to accomplish their RDM-related tasks. A majority of 32 out of 35 (91.4%) respondents answering this question rated past training experiences as "usually" or "always" useful for at least one of their RDM-related tasks.

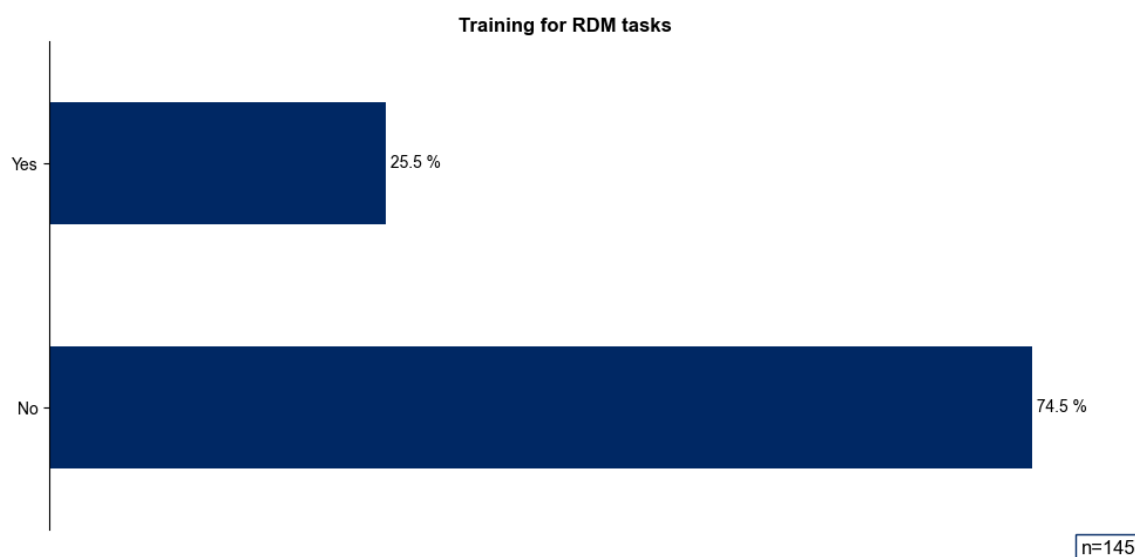**HELMHOLTZ**

Training for RDM tasks

Figure 17: "Have you had any training for your tasks in research data management?" (Question type: Single choice. Available to all respondents. Number of respondents who answered this question: n = 145, relative amounts refer to n)

This is further detailed in Figure 18, where respondents rated how useful past training was experienced with respect to RDM-related tasks that they previously reported to be working on. Most respondents "usually" or "always" experienced training positively with respect to its usefulness for "networking with other RDM-related personnel and other relevant organizations" (88.2%), followed by "publication of data and/or software in a FAIR compliant way" (76.2%), "provision of research data infrastructure" (71.3%) and the "creation of data management plans" (69.5%). Relative to these, respondents reported Moderate usefulness of past trainings with respect to their tasks in "consultation on RDM topics" (64.7%) and "training employees on RDM topics" (64.1%), as well as for "checking compliance with policies and guidelines" (55.5%), "technical support for research data infrastructure" (53.9%) and for the "enrichment of data and/or software with metadata" (53.8%).

According to the survey data, training was reported to have the least impact (the lowest rates of being "always" or "usually useful") with respect to the tasks "identification of data management needs" (47.4%), "RDM strategy on the organizational level" (46.2%), "provision of" and "technical support for long-term data archiving solutions" (both 33.3%), and regarding the "development of software for (meta)data collection and/or processing" (30.8%). Please note that percentages in this paragraph are counted relative to the number of training-related votes for each task.
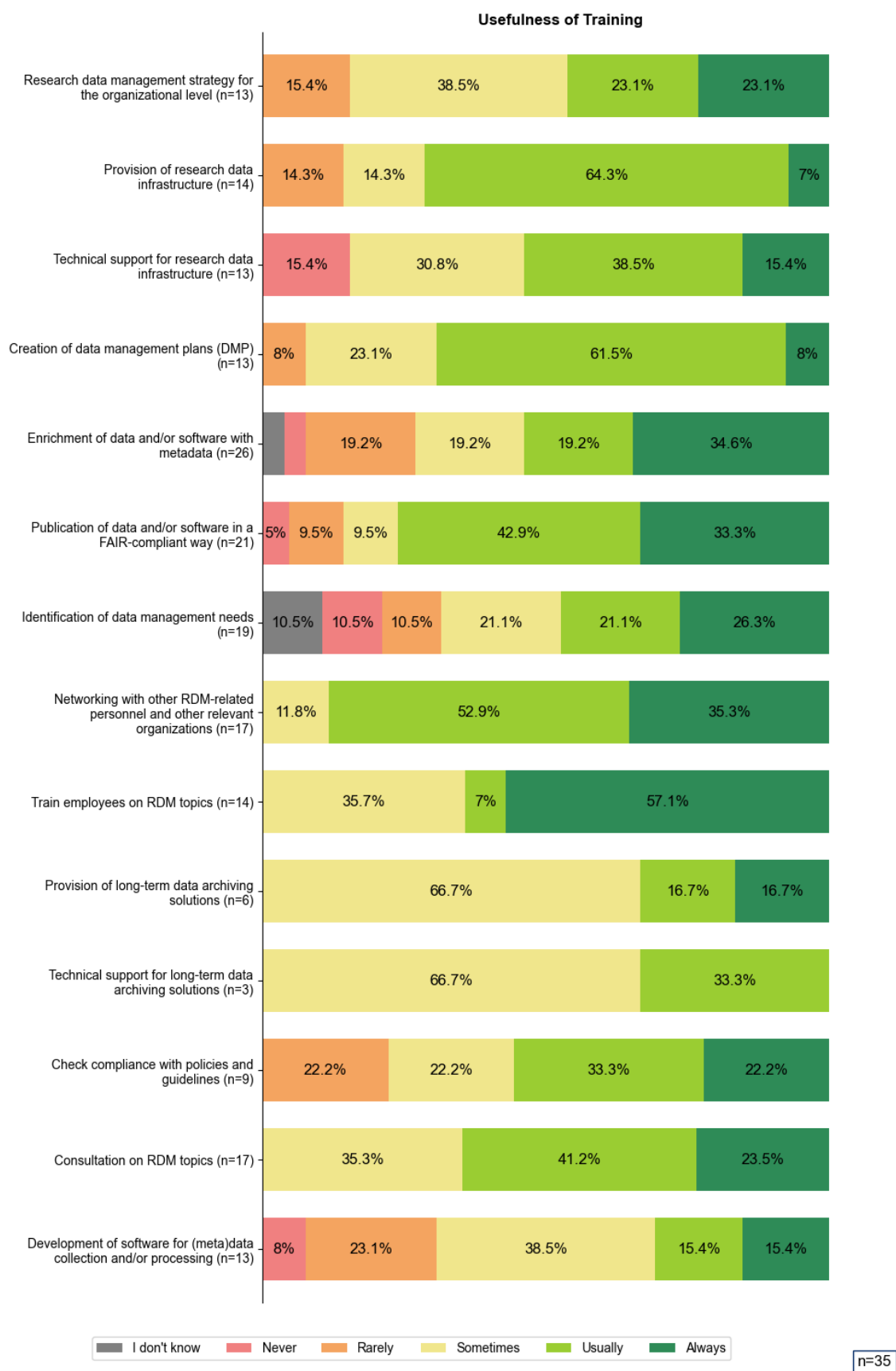
Figure 18: "Did the training help you in accomplishing the following tasks?" (Question type: array. Available to all respondents who said they have had a training for their RDM tasks. Relative amounts refer to the number of answers n for each subquestion, as shown next to each subquestion title.)

Identification of gaps and needs

The data shown in Figure 19 indicates various barriers that survey respondents report to hinder RDM-related work. With over 90%, the majority of all survey participants report difficulties in their RDM-related work. The most prominent challenge, mentioned by those 152 respondents who answered this question, is a lack of resources, 81.6%, including personnel, time and funding. In addition, 46.1% report insufficient technical knowledge, while 43.4% experience communication problems including collaborative difficulties. Moreover, 38.2% mention a lack of technical infrastructure and 34.2% a lack of technical solutions (e.g., for data discovery and standardization) and 21.7% express concerns about legal or ethical uncertainties (e.g., a lack of legal knowledge and data protection). The lowest number of respondents, 3.9%, reports to experience no obstacles in their RDM-related work. Further detail is found in the Appendix.

Obstacles

- Lack of resources — 81.6 %
- Missing technical knowledge — 46.1 %
- Communication problems — 43.4 %
- Missing technical infrastructure — 38.2 %
- Lack of technical solutions — 34.2 %
- Legal or ethical insecurities — 21.7 %
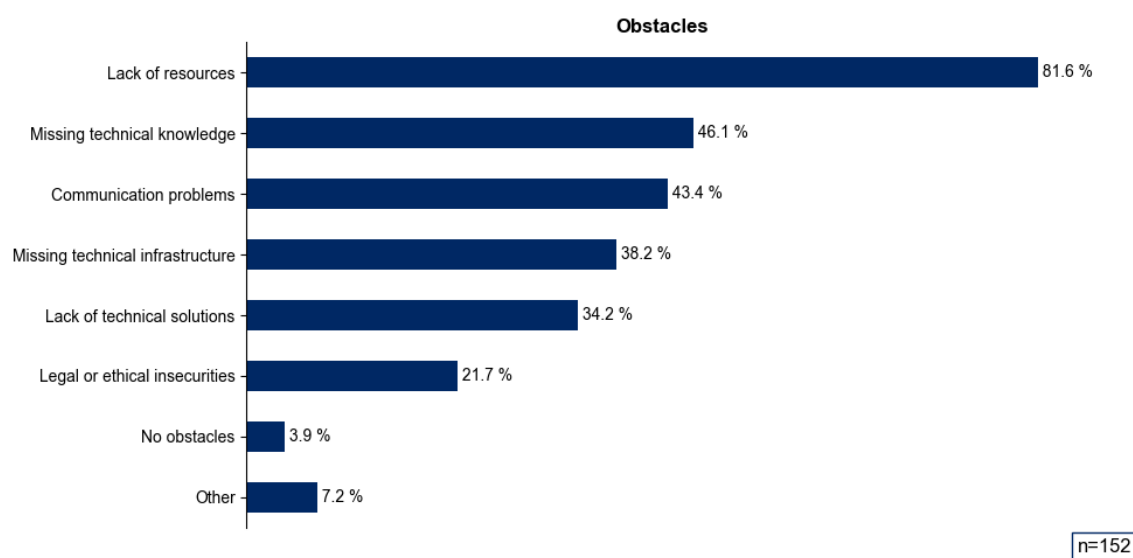- No obstacles — 3.9 %
- Other — 7.2 %

n=152

Figure 19: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 152, relative amounts refer to n)

Figure 20 reveals the need for RDM-related service formats, as reported by 85.3% of all survey respondents. Trainings are the most requested service format, mentioned by 58.6% of respondents who answered this question. The most requested contents for such trainings are RDM tools and workflows, best practices, as well as metadata schema development and metadata handling. This is followed by the need for RDM-related tools, 39.8%. Specifically, a need for tools to set up and manage RDM infrastructure, for Git-related tools, and for data storage and ontology-related tools was most mentioned. A need for technical support is expressed by 39.1%. Specifically, a need for support in setting up and managing RDM infrastructure and with RDM-related tools and workflows is mentioned. Additionally, participants express an interest in federated Helmholtz infrastructure. 35.3% express a need for recommendations, specifically on best practices, ontologies, and metadata standards, but also a need for recommendations on metadata development, data provenance and licensing is mentioned.

Around one third of respondents express an interest in workshop formats, 32.3%, and networking events, 31.6%, specifically focusing on community engagement, best practices, RDM tools and workflows, ontologies and metadata handling. (See the Appendix for further detail). 25.6% express an interest in attending interest groups, 18.0% talks and 12.8% helpdesk formats. Charts with further details can be found in the Appendix of this report.

HELMHOLTZ

Interest in Service Formats

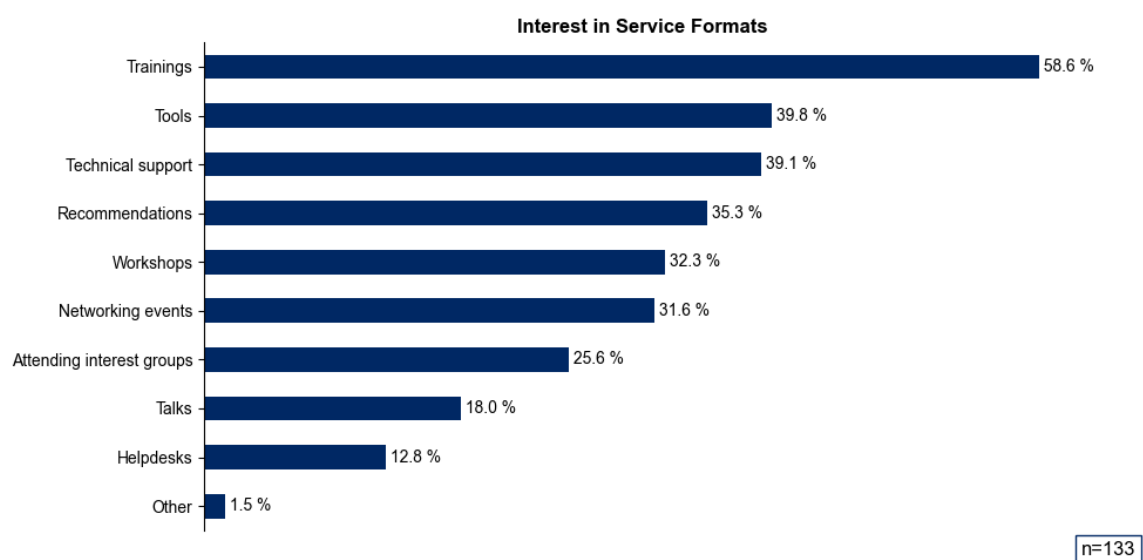| Service Format	Percentage
Trainings	58.6 %
Tools	39.8 %
Technical support	39.1 %
Recommendations	35.3 %
Workshops	32.3 %
Networking events	31.6 %
Attending interest groups	25.6 %
Talks	18.0 %
Helpdesks	12.8 %
Other	1.5 %

n=133

Figure 20: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 133, relative amounts refer to n)

Discussion

The following sections discuss the survey results presented in the preceding chapter. Where relevant, findings from the 2021 survey are included as a comparative baseline; target groups are highlighted in italics for clarity.

Statistical robustness

With 156 complete survey replies, the data provides a relatively robust measure for analyzing trends in the current state of research data management in the Helmholtz Association: for a tentative estimate of up to 5%[10] (2150) of all Helmholtz staff being involved in research data management-related tasks, at 95% confidence level and conservatively assuming epistemic uncertainty, we assume sample size-related error margins of less than 8%. This estimation enhances the validity and generalizability of the survey findings, allowing for a relatively robust assessment of the organization-wide landscape of data management practices, challenges, and needs from the perspective of data professionals.

Data professionals in Helmholtz

The survey data reveals a relatively high degree of multi and cross-level engagement in RDM-related work in the Helmholtz Association (Hypothesis 1). Around 65% of survey respondents indicate engagement with multiple organizational levels. Moreover, the heterogeneity in professions involved in RDM-related tasks emphasizes the multifaceted nature and heterogeneous distribution of data management work in the Helmholtz Association. This suggests that strategies aiming at enhancing RDM and FAIR data practices across the organization require a coordinated approach to address these multifaceted stakeholders on all organizational levels within and across centers.

We observe that staff from most participating centers report engagement in each of the probed RDM-related work packages. This suggests a notable alignment of RDM-related tasks across the participating Helmholtz centers (Hypothesis 2). This observation may indicate an increasingly coordinated commitment to RDM-related activities throughout the Helmholtz Association. The consistency of RDM-related work across centers suggests future enhancement through improved coordination and collaboration of these efforts across organizational levels and centers.

Data policies and guidelines

The relatively high awareness of 75.4% with respect to data management policies and guidelines is complemented by the observation that nearly a quarter of respondents (24.6%) are unaware of these policies. This suggests an opportunity for improvement in policy dissemination and education to reach the remaining 24.6%. This could be addressed by including policy-related content in role-specific

---

[10] Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. Nature, 578(7796), 491-491. https://doi.org/10.1038/d41586-020-00505-7 .

training plans. In addition, overarching review of data policies across centers may be continued to ensure alignment with the FAIR data principles in the centers.

These considerations may be complemented with a quote by a participant of a post-survey focus group on this topic who suggests:

> *"Policies are not a checklist for researchers, they are a strategic tool to help management decide what (…) where/how to implement. Policy is very much a proxy for how aware scientists are of data management strategy in their center."*

While most respondents are aware of their center's data management policy or guidelines, the survey data also suggests a gap between policy awareness and practical implementation: 70.4% of those reporting to be aware of their center's policies or guidelines do "rather" or "fully agree" to aligning their work accordingly. Multiplication of the respective fractions leads to an average amount of slightly more than half of all survey respondents reporting to align their work to their center's policy (Hypothesis 5).

To address this gap, the implementation of policy-related auditing processes could be discussed, as suggested by the survey data. For auditing processes to remain scalable, these might profit from standardized, machine-actionable data management plans.

The data also show a significant adherence to external policies and guidelines, pointing out to the multifaceted nature of research data management in a multi-disciplinary research environment like the Helmholtz Association. A coordinated approach to harmonized integration of disciplinary aspects in guidelines and policies across centers could leverage this trend.

Research data management and FAIR practices

The survey data provides valuable insight into the heterogeneous alignment of RDM-related work in the Helmholtz Association with respect to the FAIR data principles (Hypothesis 4). Data collected with self-assessment questions, that were derived from the FAIR-Aware questionnaire, reveal varying adoption rates with respect to selected FAIR principles among *RDM-related staff* in the Helmholtz Association. Findability practices, with over 50% of respondents consistently applying descriptive metadata and persistent identifiers, show the strongest adoption rates. Although over 50% report that access to data is always or usually controlled, on average, accessibility practices show lower adoption rates than those reported for the findability of data. Relative to these, interoperability and reusability practices show lowest adoption rates, which is in line with the survey data collected in 2021 where few *researchers* reported "[using] … international templates, schemas or standards".[11]

This data clearly suggests areas for improvement, particularly regarding the implementation of controlled vocabularies and community-endorsed metadata standards. These aspects could be leveraged by role and discipline-specific training modules and recommendations focusing on interoperability and reusability. A coordinated approach to guidelines, trainings, and tools for improved (meta)data curation could further enhance FAIR data management practices across the organization.

Feedback loops based on regular monitoring and assessment of these FAIR data practices can support the continuous enhancement of the FAIR data landscape across the Helmholtz Association.

---

[11] Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022). A survey on research data management practices among researchers in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_05

HELMHOLTZ

Ecosystem of RDM-related tools

The survey data reveals a diverse ecosystem of research data management-related tools used across the Helmholtz Association, with varying levels of adoption and in-house (co-)development (Hypothesis 6). This diversity reflects the complex and multifaceted nature of RDM practices within Helmholtz.

In addition, a post-survey focus group on this topic revealed the need for resources to bring (RDM-related) software to a reusable state:

> *"Science needs the opportunity to move software projects smoothly from development to maintenance state. This is missing now."*

Data management and FAIR data practices across the organization could benefit from coordinated efforts and resources to develop and reuse RDM-related tools within and across centers.

Training

The survey data informs about low training rates of only one fourth of Helmholtz staff working on RDM-related tasks (Hypothesis 3). This may reflect the fact that research data management is a rapidly evolving, relatively new working field where lots of training offers are potentially yet to be developed.

To address the low training rates observed in the survey data, HMC could further enhance their training programs by providing structured curricula and ensure that all Helmholtz staff members have access to relevant and adequate training.

The survey data also reveals that training can enhance the effectiveness of RDM-tasks, since the vast majority of trained respondents reported to find their training useful for at least one aspect of their work. While training was perceived most beneficial for networking with other RDM personnel, lower perceived usefulness was reported for other tasks, suggesting that available training may not fully address all RDM-related work areas.

These considerations are complemented with participants' quotes in a post-survey focus group on this topic who suggest

> *"An organization should take the responsibility [for RDM-related trainings]. The problem is not the lack of training courses."*

and

> *"We don't know who offers the trainings, where people can find it and register for it."*

These observations suggest a need for enhanced communication of available training programs on RDM-related topics as well as their organizational implementation into data professionals' curricula.

Identification of gaps and needs

The data reveals significant barriers that can hinder RDM-related work towards coherent FAIR data practices within the Helmholtz Association (Hypothesis 8). As over 90% of the *RDM-related staff* report respective difficulties, the challenges most reported are a lack of resources, a lack of technical knowledge and communication issues. A lack of technical infrastructure and technical solutions are mentioned by 38.2% and 34.2% of the survey respondents, respectively (Hypothesis 7). Mirroring

these survey data, the challenge most reported by *researchers* in 2021 was also a lack of resources, followed by a lack of technical solutions.[12]

A possible solution to these challenges could be the coordinated development of technical infrastructure and tools for improved RDM and FAIR data practices across the Helmholtz Association.

The survey results underscore a clear demand for RDM-related service formats, with 85.3% of respondents expressing concrete needs for support (Hypothesis 9). As suggested by the low training rates observed in the data, RDM-related training is the most frequently requested support format. This data shows a high willingness of *RDM-related Helmholtz staff* to enhance their skillset. This data is consistent with the data collected in the 2021 survey, where over 90% of *researchers* responded that they were very interested or moderately interested in training and workshop formats.

Moreover, in the current survey, *RDM-related staff* expresses an interest in RDM-related tools, infrastructure and solutions as well as the need for technical support. This aligns with the needs reported by *researchers* in the 2021 survey, where most respondents expressed a need for RDM software and tools and reported their interest in technical support.

These results suggests possible enhancement of RDM- and FAIR data practices in the Helmholtz Association through improved coordination and collaboration of multiple efforts across organizational levels and centers. Possible areas of improvement may be coordinated approaches to foster collaboration, exchange and reuse of RDM-related training, knowledge, tools and infrastructure across centers.

Assessment of hypotheses

Table 1 shows an evaluation of the initial hypotheses. Hypotheses 1, 2, 6-9 are supported, hypotheses 4-5 are partially supported, and hypothesis 3 is not supported by the survey data.

Table 1: Evaluation of initial hypotheses based on the analysis of the survey data.

HYPOTHESIS	SUPPORTED BY THE SURVEY DATA?
1. RDM-related work is done on different institutional levels (center, institute(s), departments, research groups) within Helmholtz.	Supported
2. Work packages among RDM-related personnel are streamlined cross the centers.	Supported
3. At least 50% of staff with RDM-related tasks have had a training for these tasks.	Not supported
4. At least 50% of staff with RDM-related tasks in Helmholtz align their work with the FAIR data principles.	Partially supported
5. At least 50% of staff with RDM-related tasks align their work based on the policies and guidelines in their center.	Partially supported
6. Tools commonly used by RDM-related personnel are also used by others in the respective communities.	Supported

---

[12] Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022). A survey on research data management practices among researchers in the Helmholtz Association (HMC Report). https://doi.org/10.3289/HMC_publ_05

HELMHOLTZ

HYPOTHESIS	SUPPORTED BY THE SURVEY DATA?
7. At least 50% of staff with RDM-related tasks report that (technical) infrastructure for implementing the FAIR data principles is available.	Supported (indirectly)
8. RDM-related personnel have gaps and needs with respect to the FAIR compliance of data.	Supported
9. RDM-related personnel are interested in the services offered by HMC.	Supported

Conclusions

The HMC Data Professionals Survey 2024, targeted at Helmholtz staff regularly working on tasks related to research data management (RDM), was answered by 156 respondents from 16 Helmholtz centers and all six Helmholtz research fields. This number allows for a relatively robust analysis of general trends in the current state of research data management and FAIR data practices, and respective gaps and needs in the Helmholtz Association.

The survey data reveals a high degree of alignment in RDM-related work packages, alongside a relatively high degree of multi-professional, and cross-level engagement in RDM-related work across the entire organization.

Most respondents report to be aware of their center's data policies and mostly align their work to these policies. Nevertheless, the data reveals potential gaps between policy awareness and practical implementation as well as opportunities for improved policy dissemination and education.

According to their self-assessment, Helmholtz staff working on RDM-related topics often align their work with selected FAIR data principles. Adoption rates with respect to selected FAIR data principles vary between one-third and two-thirds, which suggests clear areas for improvement, in particular with respect to the interoperability and reusability of research data.

FAIR data practices are supported by an ecosystem of RDM-related tools being used and, in parts, (co-)developed in various Helmholtz centers. Qualitative data from focus groups suggests a need for resources and support to further enhance the reusability of these tools across centers.

Most survey respondents report gaps and needs that can hinder progress in their RDM-related work. The survey results reveal a low level of formal training among Helmholtz staff with respect to their RDM-related tasks. Despite most respondents expressing interest in such training opportunities, this suggests a need for enhanced development and communication of such training programmes, as well as their regular inclusion in the curricula of data professionals. Morover, respondents report a lack of essential resources, followed by technical knowledge, infrastructure and technical solutions, while expressing a need for RDM-related tools, technical support, and recommendations for further enhancement of RDM and FAIR data practices across the organization.

In summary, the presented survey results reveal current trends, gaps and needs with respect to research data management and FAIR data practices in the Helmholtz Association and will guide HMC's strategic initiatives to improve the FAIR data ecosystem within the Helmholtz Association.

Credits and Acknowledgements

Author contributions

In the following paragraph, author contributions are summarized following the CRediT[13] taxonomy.

S.C.G., M.K., L.K., C.L., J.S., and S.S. conceptualized the study. M.K., L.K., S.S. and K.U.S. curated the data. M.K., L.K. and S.S. formally analyzed the data. M.K., L.K., and S.S. conducted the investigation. S.C.G., M.K., L.K., C.L., J.S., and S.S. developed the methodology. L.K. and S.S. worked on the software. M.K., L.K., A.S. and S.S. validated the data. M.K., L.K., and S.S. visualized the data. M.K., L.K. and S.S. wrote the original draft of this report. M.K., C.L., S.S., K.U.S. and the editors of this report reviewed and edited the report. M.K. and L.K. administrated the project. The editors of this report supervised the project.

Disclosure statement

In the preparation of this manuscript, we utilized large language models for the following purposes. Llama 3.3 was used to precategorize common categories in free text answers. These mappings were manually revised by the project team. Where all substantive content of this study including data analysis was generated or performed by the authors of this study, DeepL and Perplexity AI (based on OpenAI's GPT models) were used to assist in improving clarity of expression and stimulating ideas for broader implications. All output of these large language models was carefully reviewed, selected, and edited by the authors to ensure accuracy and relevance to the study.

Acknowledgements

---

[13] CRediT (Contributor Roles Taxonomy): https://credit.niso.org

Appendix: Supplementary Figures

Figure A1: "Please rate the following statements concerning the data(sets) you process or make publicly accessible." (Question type: array. Available to all respondents. Relative amounts refer to the number of answers n for each subquestion, as shown next to each subquestion title.)

RDM Tools used
Electronic lab notebooks

ELabFTW	17.6 %
Herbie	14.7 %
Chemotion	11.8 %
Biovia ELN	11.8 %
MediaWiki	8.8 %
ELog	8.8 %
Kadi4Mat	8.8 %
NOMAD	8.8 %
Other	5.9 %
Google Docs	2.9 %
Confluence	2.9 %
Joplin App	2.9 %
MS Excel	2.9 %
RSpace	2.9 %
ESRF Electronic logbook	2.9 %
ICAT+	2.9 %
LabCollector	2.9 %
Google/Sync	2.9 %
Scilog	2.9 %
API	2.9 %
Snip	2.9 %
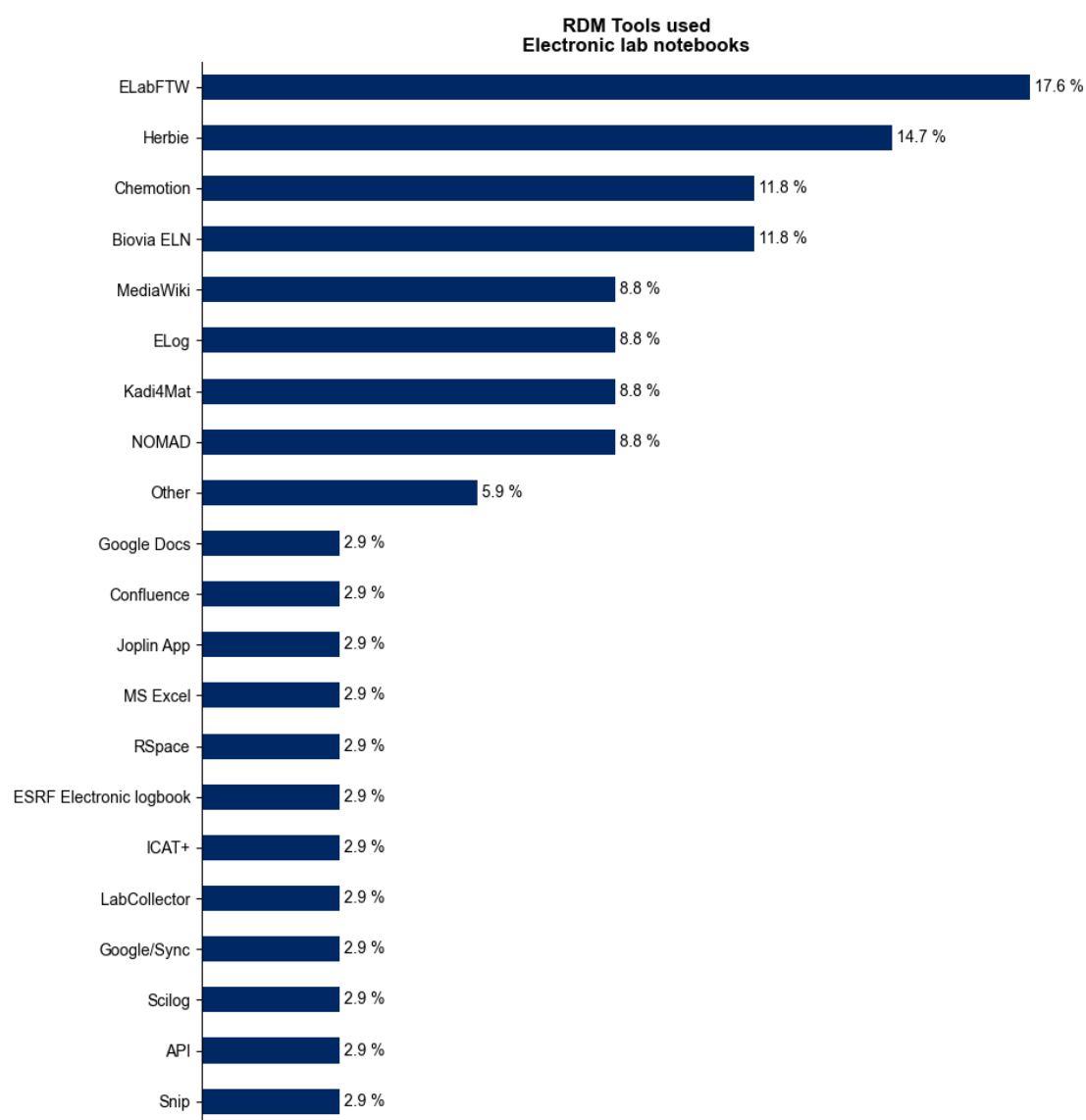
n=34

Figure A2: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 34, relative amounts refer to n)

HELMHOLTZ

RDM Tools used
Controlled vocabularies

Dublin Core	37.5 %
schema.org	30.4 %
DataCite	14.3 %
NERC Vocabularies	10.7 %
ISO	8.9 %
CF Conventions	7.1 %
ODM2	7.1 %
BFO	5.4 %
GCMD	3.6 %
NeXus	3.6 %
PROV-O	3.6 %
Other	3.6 %
voc4cat	3.6 %
Inspire	3.6 %
stratigraphy.org	1.8 %
OneGeochemistry	1.8 %
FABIO	1.8 %
FOAF	1.8 %
FRBR	1.8 %
PaNET	1.8 %
re3data	1.8 %
TEI	1.8 %
Wikidata	1.8 %
LinkML	1.8 %
PolyMat	1.8 %
ERC Vocabularies	1.8 %
QUDT	1.8 %
AMS	1.8 %
EMMO	1.8 %
DC	1.8 %
OBO Foundry	1.8 %
American Meteorological Society	1.8 %
DCAT	1.8 %
codemeta	1.8 %
MIxS	1.8 %
gensc.org	1.8 %
OLS	1.8 %
TFSCO	1.8 %
NFDI4Chem	1.8 %
Relation Ontology	1.8 %
CHMO	1.8 %
PANGAEA schema	1.8 %
Eprints	1.8 %
Open Energy Ontology	1.8 %
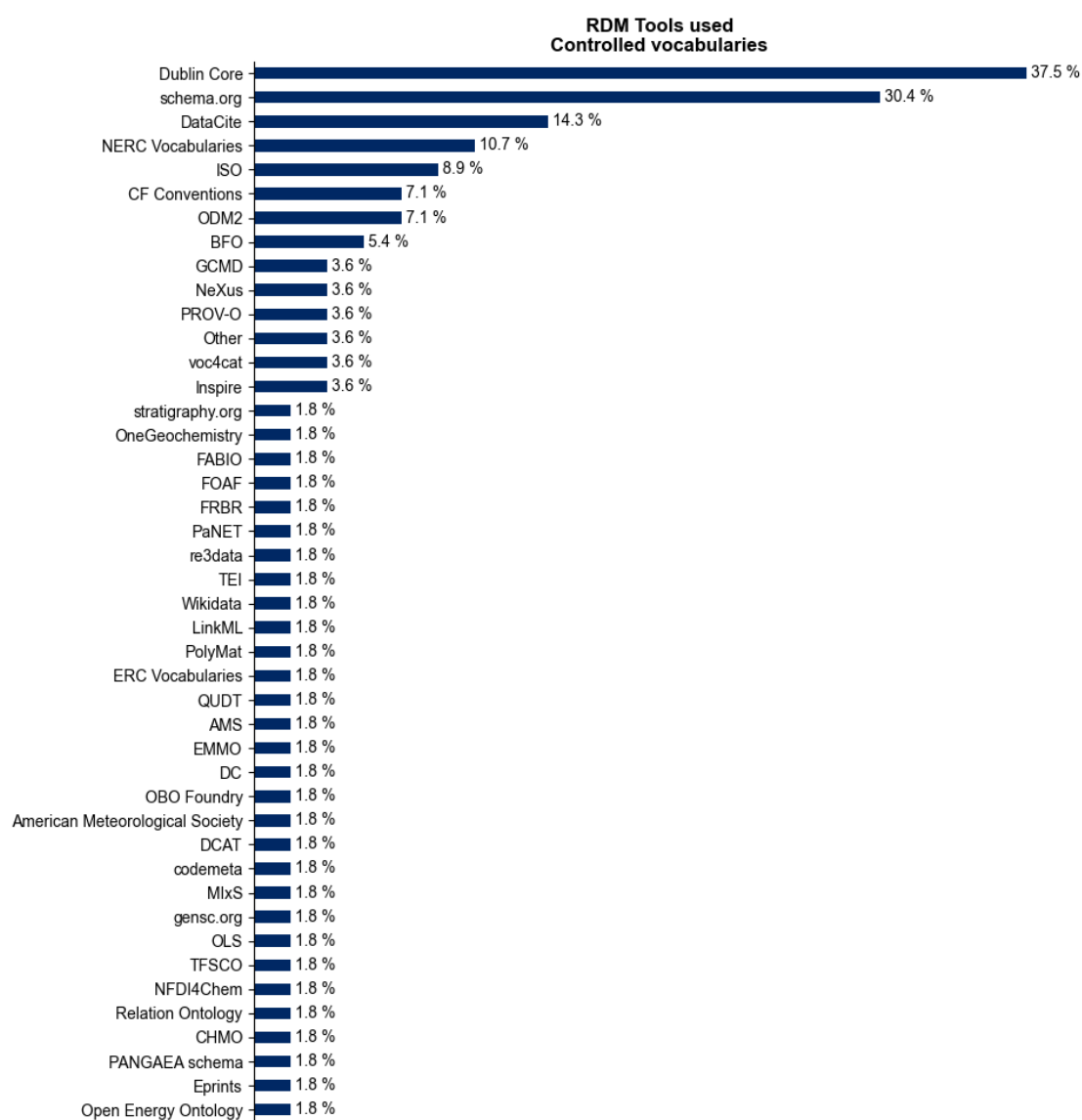
n=56

Figure A3: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 56, relative amounts refer to n)

Figure A4: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 38, relative amounts refer to n)

RDM Tools used
Registries

Tool	Percentage
re3data	48.1 %
Zenodo	25.9 %
fairsharing.org	11.1 %
b2inst	3.7 %
EGA	3.7 %
Pangaea	3.7 %
RADAR	3.7 %
Other	3.7 %
Fabrica	3.7 %

n=27

Figure A5: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 27, relative amounts refer to n)

RDM Tools used
PID services

Tool	Percentage
ORCID	76.3 %
DataCite	42.1 %
ROR	11.8 %
EPIC	6.6 %
handle.net	5.3 %
B2INST	3.9 %
Zenodo	3.9 %
IGSN	2.6 %
PIDA	2.6 %
DOI	2.6 %
PID4INST	1.3 %
Crossref	1.3 %

n=76

Figure A6: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 76, relative amounts refer to n)

HELMHOLTZ

RDM Tools used
Version control software

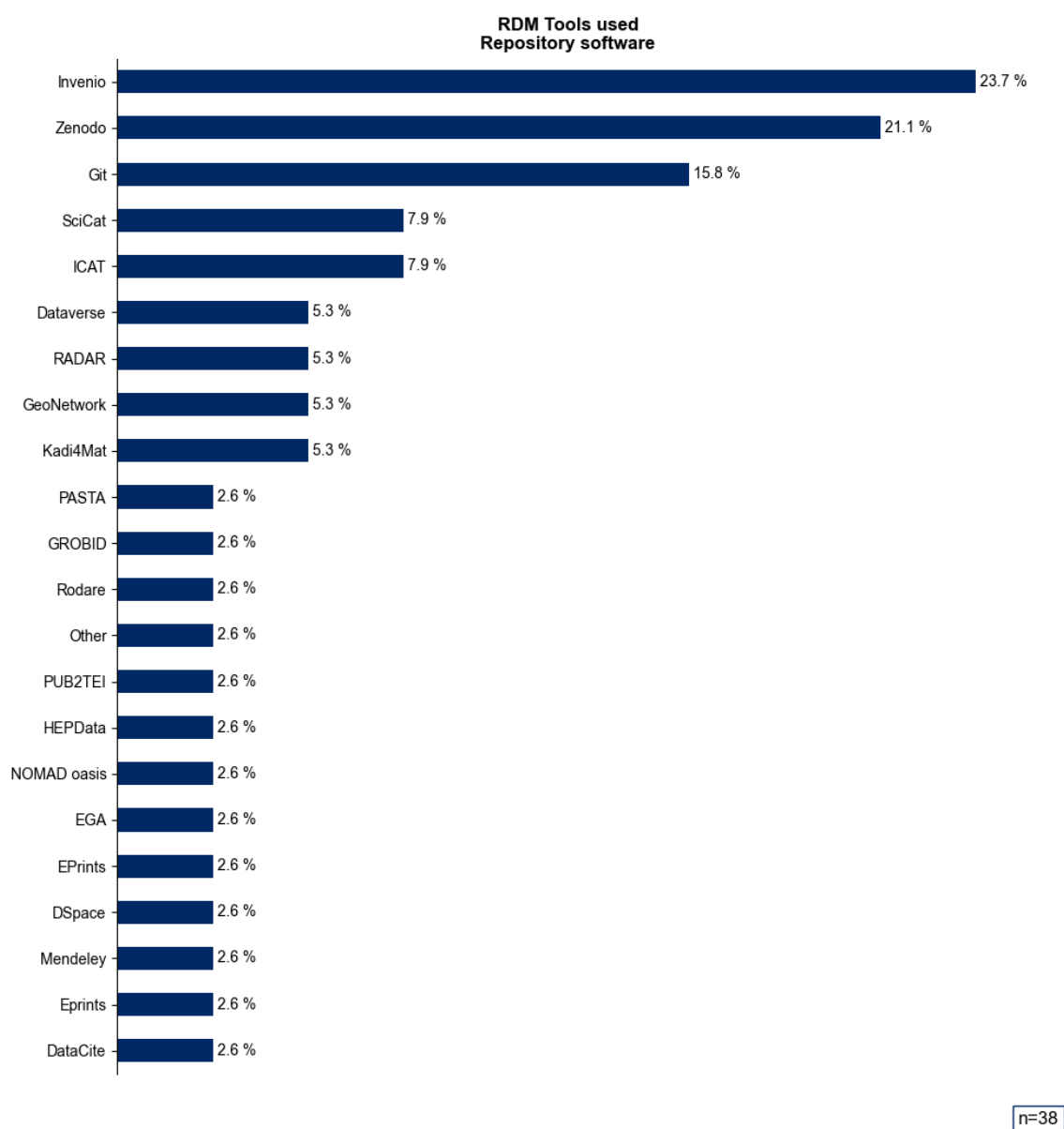Git — 98.0 %

Subversion — 5.0 %

n=101

Figure A7: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 101, relative amounts refer to n)

RDM Tools used
DMP software

RDMO — 89.5 %

Fair-wizard — 10.5 %

Moses DMP — 5.3 %
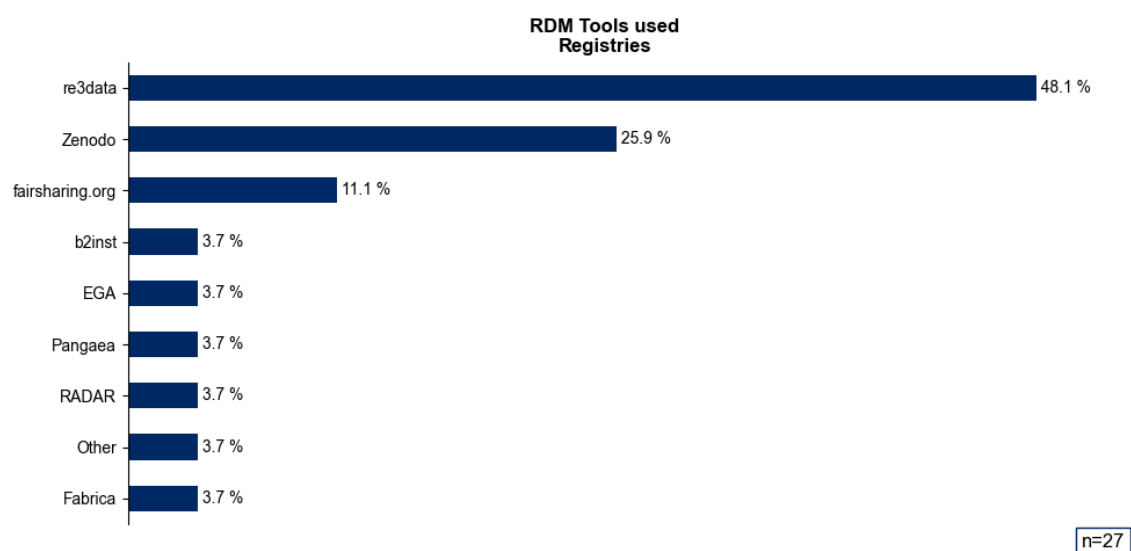
Shepard — 5.3 %

n=19

Figure A8: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 19, relative amounts refer to n)

HELMHOLTZ

RDM Tools used
Vocabulary / ontology editors

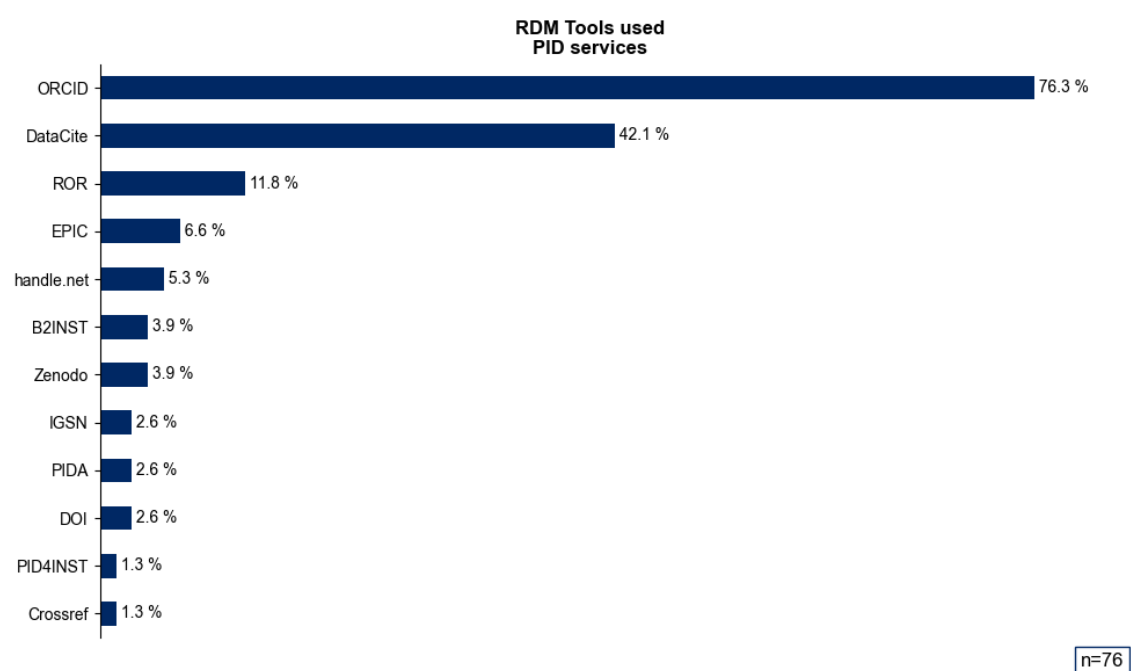| Tool	Percentage
Protégé	88.9 %
VocPopuli	5.6 %
Karma	5.6 %
WoRMS	5.6 %
ChEBI	5.6 %
ITIS	5.6 %
VocBench	5.6 %

n=18

Figure A9: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 18, relative amounts refer to n)

RDM Tools used
Other

Tool	Percentage
LinkML	10.0 %
Json schema	10.0 %
Sensor Management System	10.0 %
Geo.X	10.0 %
Other	10.0 %
OMERO	10.0 %
SciCat	10.0 %
BBEdit	10.0 %
MS Excel	10.0 %
VS Codium	10.0 %
CaRMeN	10.0 %
Adacta	10.0 %
KadiStudio	10.0 %
KadiAI	10.0 %
Shepard	10.0 %

n=10

Figure A10: "Which tools do you use for your RDM-related work? Please list the full names of the most relevant ones for your work." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 10, relative amounts refer to n)
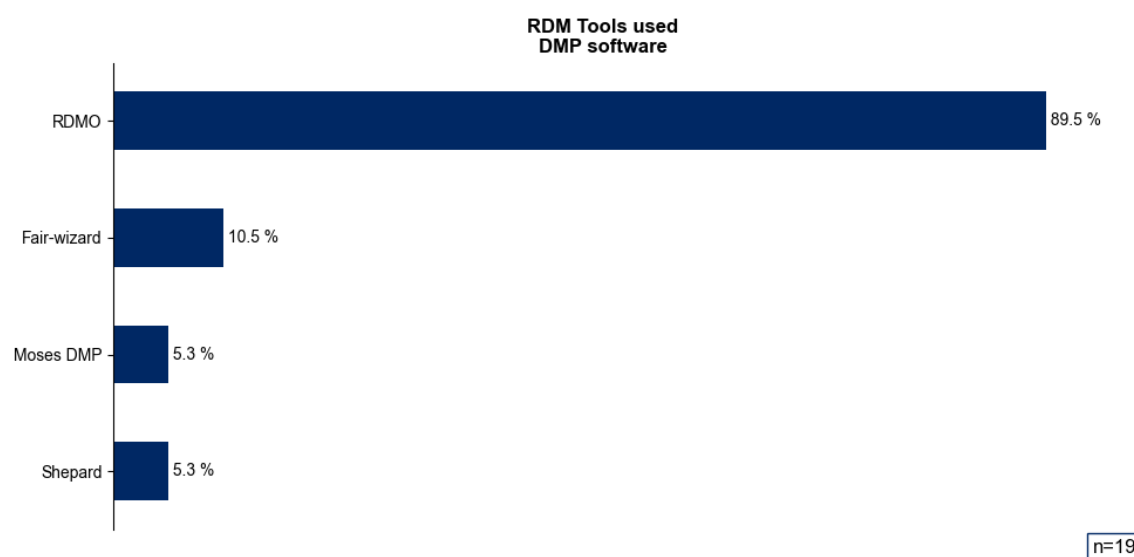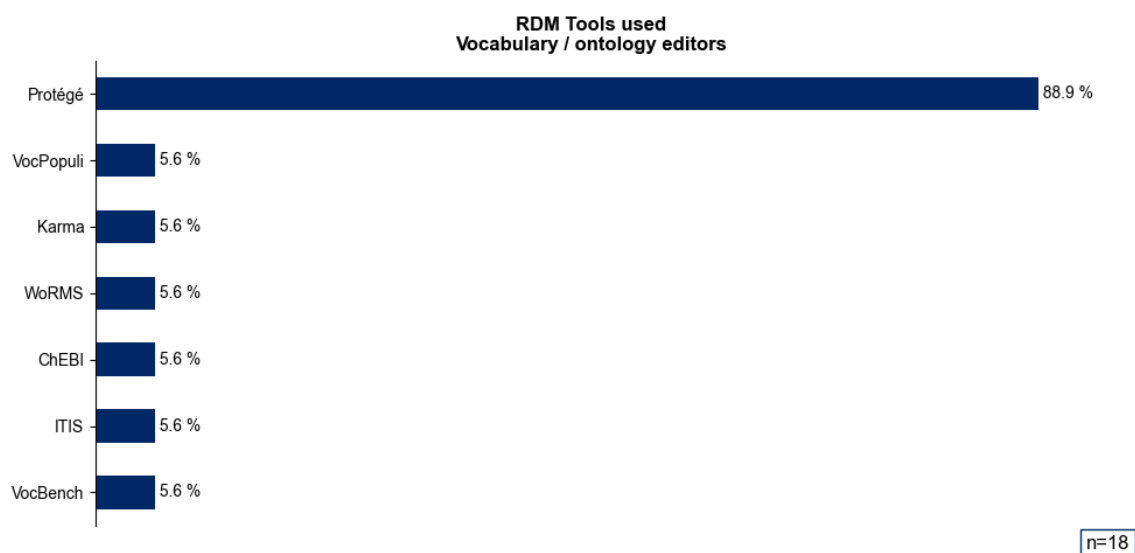
HELMHOLTZ
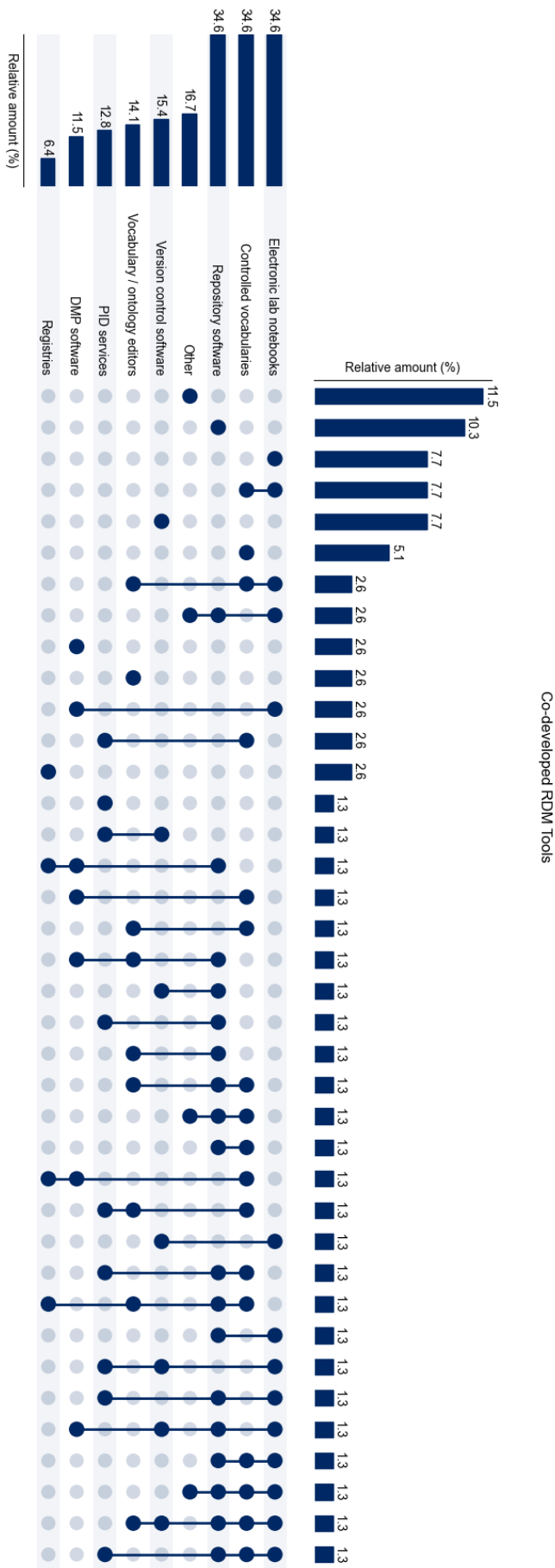
Figure A11: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice. Available to all respondents. Number of respondents who answered this question: n = 78, relative amounts refer to n)

Co-developed RDM Tools
Electronic lab notebooks

Tool	Percentage
Herbie	22.2 %
NOMAD	16.7 %
Kadi4Mat	11.1 %
MediaWiki	11.1 %
ELabFTW	11.1 %
Other	5.6 %
ESRF	5.6 %
ICAT+	5.6 %
ELog	5.6 %
BioVia	5.6 %

n=18

Figure A12: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 18, relative amounts refer to n)
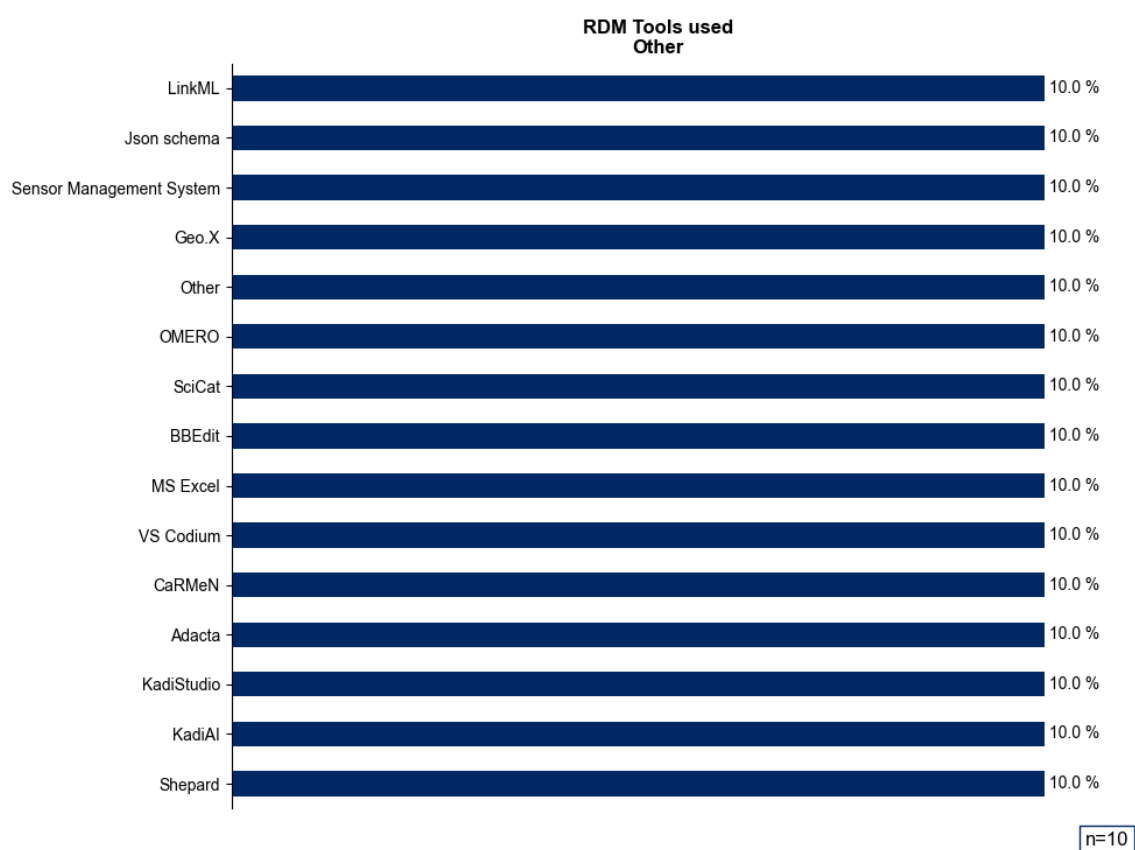
Co-developed RDM Tools
Controlled vocabularies

Tool	Percentage
NeXus	16.7 %
Other	16.7 %
TFSCO	11.1 %
PolyMat	11.1 %
PolyLab	11.1 %
OGC	11.1 %
EM glossary	5.6 %
openNP catalog	5.6 %
OTP	5.6 %
EMMO	5.6 %
GUIDE	5.6 %
NEP Schema	5.6 %
ILDG Metadata Schema	5.6 %
openPMD	5.6 %
voc4cat	5.6 %
I-Adopt	5.6 %
Open Energy Ontology	5.6 %

n=18

Figure A13: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 18, relative amounts refer to n)

HELMHOLTZ

Co-developed RDM Tools
Repository software

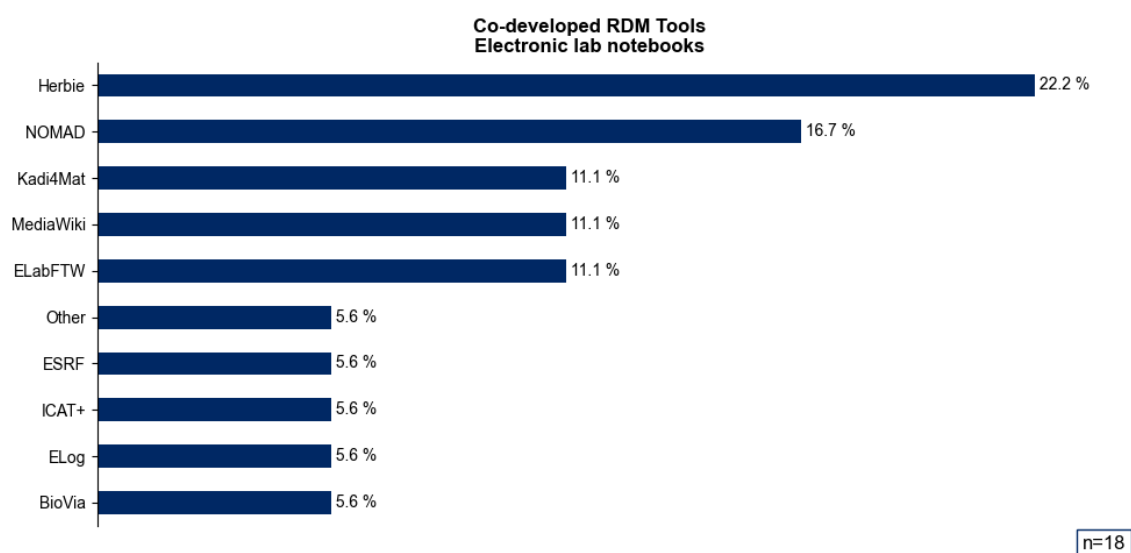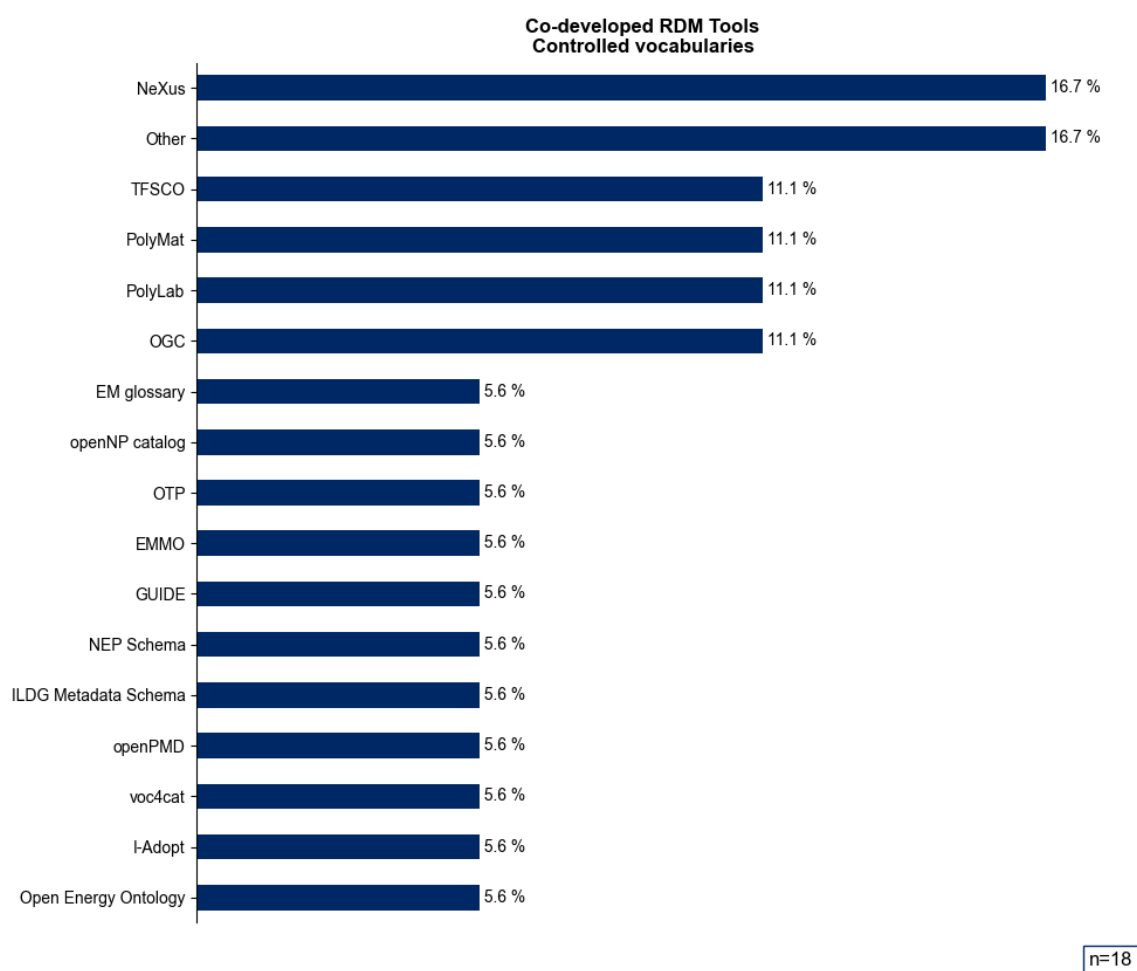| Tool	Percentage
SciCat	9.5 %
ICAT	9.5 %
Invenio	9.5 %
Metador	4.8 %
GHGA	4.8 %
Rodare	4.8 %
Nefeli	4.8 %
vimi labs	4.8 %
GBD	4.8 %
EPrints	4.8 %
Other	4.8 %
RADAR	4.8 %
Punch MDC	4.8 %
MetaStore	4.8 %
Kadi4Mat	4.8 %
Dataverse	4.8 %
GeoNetwork	4.8 %
Databus	4.8 %

n=21

Figure A14: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 21, relative amounts refer to n)

Co-developed RDM Tools
Registries

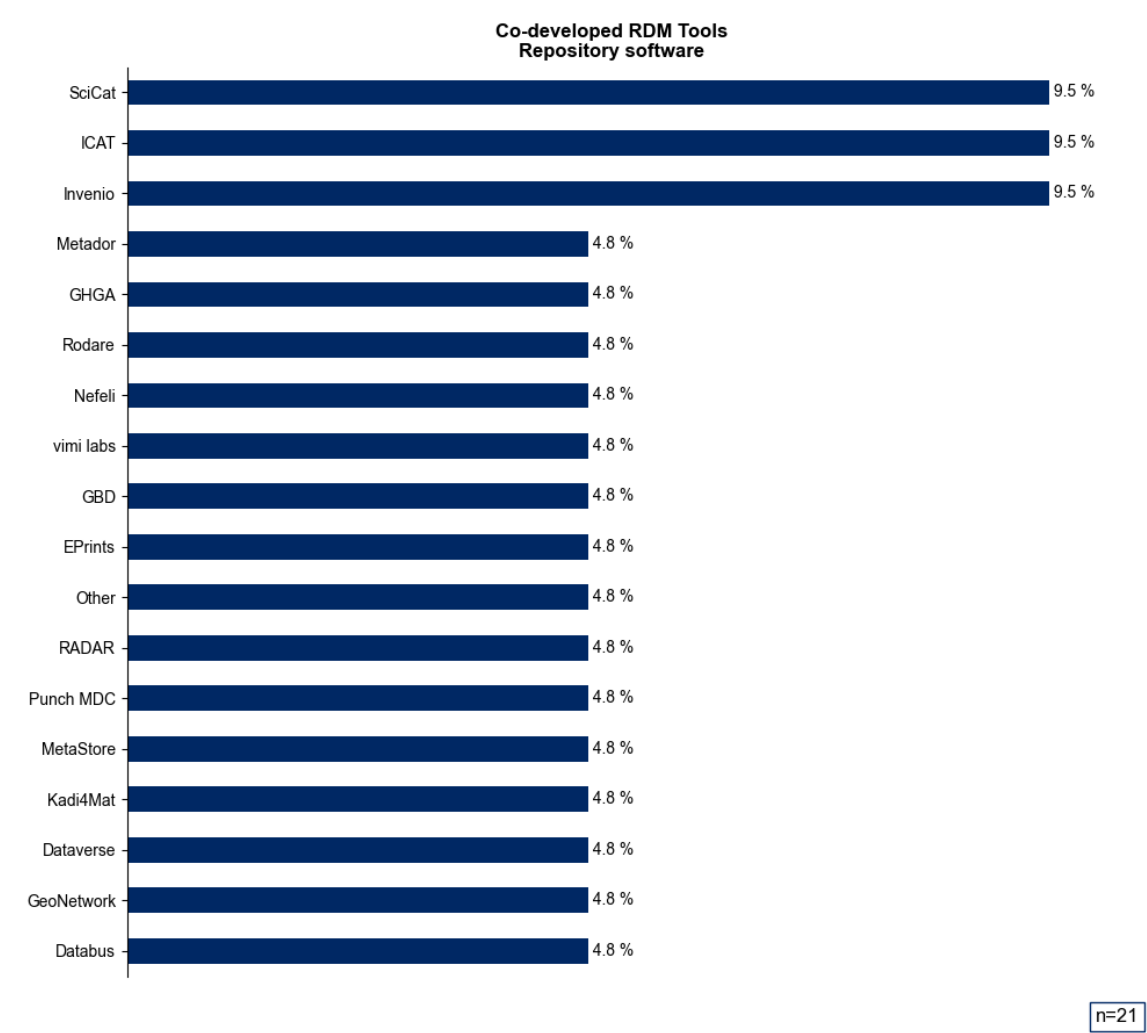| Tool	Percentage
re3data	50.0 %
GHGA	25.0 %
Other	25.0 %

n=4

Figure A15: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 4, relative amounts refer to n)

HELMHOLTZ

**Co-developed RDM Tools
PID services**

IGSN ▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬ 40.0 %

PIDA ▬▬▬▬▬▬▬▬ 20.0 %

DataCite ▬▬▬▬▬▬▬▬ 20.0 %

Other ▬▬▬▬▬▬▬▬ 20.0 %

n=5

Figure A16: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 5, relative amounts refer to n)

**Co-developed RDM Tools
Version control software**

Git ▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬ 90.0 %
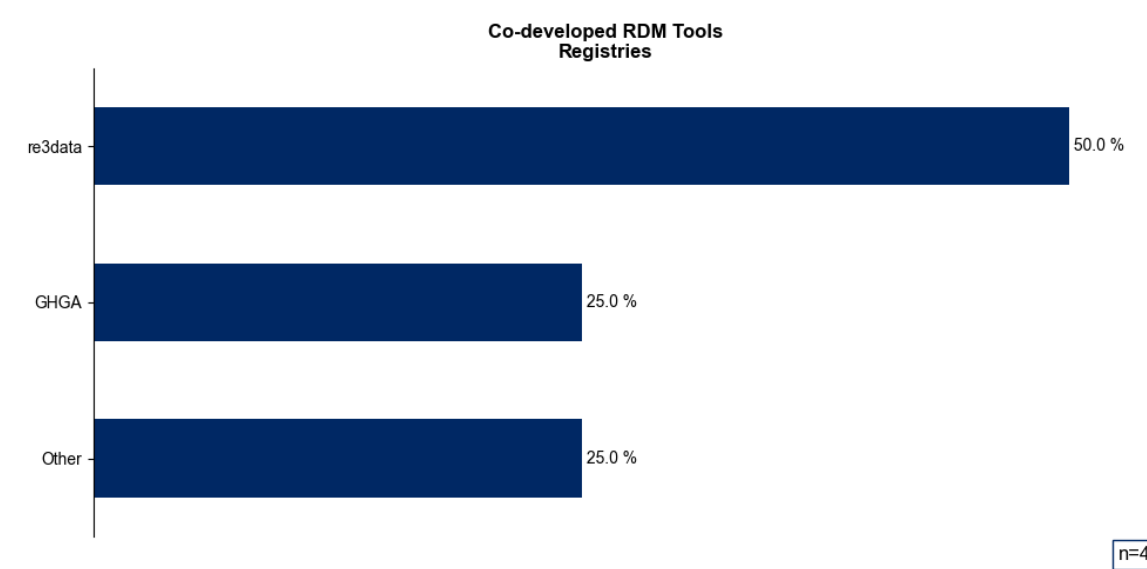
HERMES ▬▬▬ 10.0 %

n=10

Figure A17: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 10, relative amounts refer to n)

Co-developed RDM Tools
DMP software

RDMO — 57.1 %
vimi labs — 14.3 %
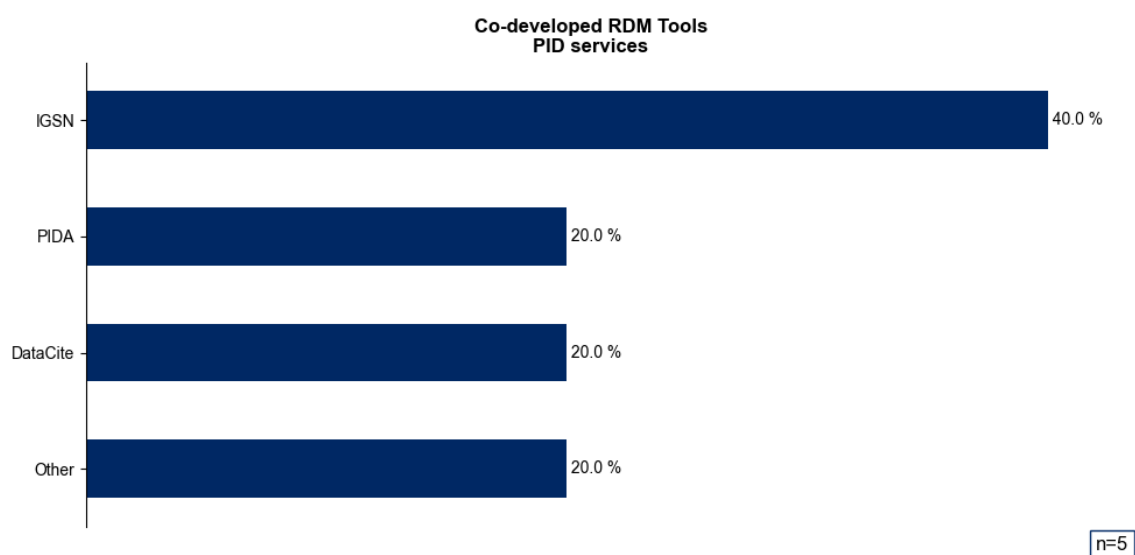Other — 14.3 %
Twinstash — 14.3 %

n=7
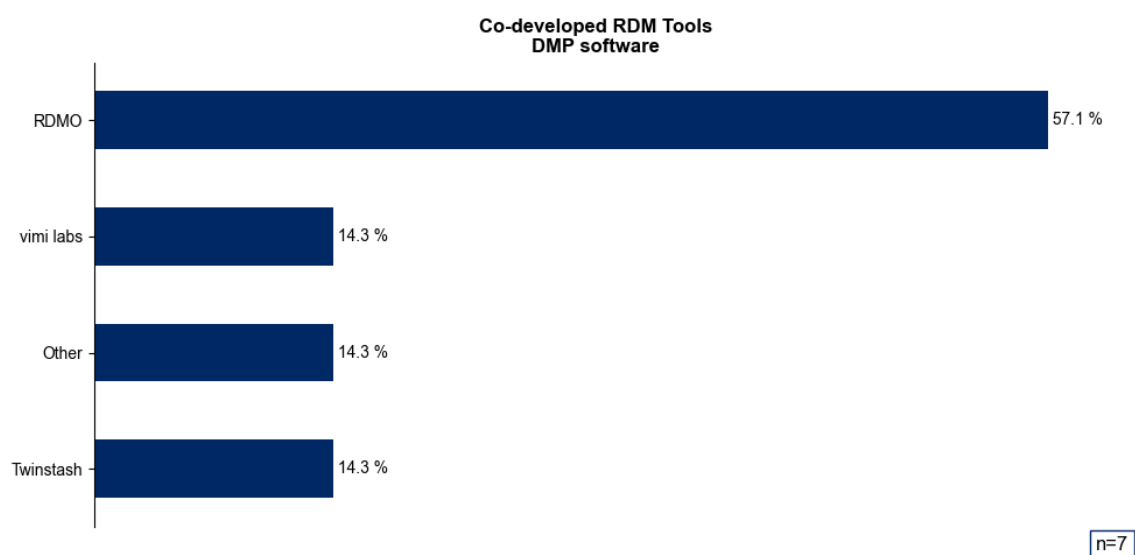
Figure A18: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 7, relative amounts refer to n)

Co-developed RDM Tools
Vocabulary / ontology editors

VocPopuli — 28.6 %
graph4mat — 14.3 %
GUIDE — 14.3 %
vimi — 14.3 %
EVOKS — 14.3 %
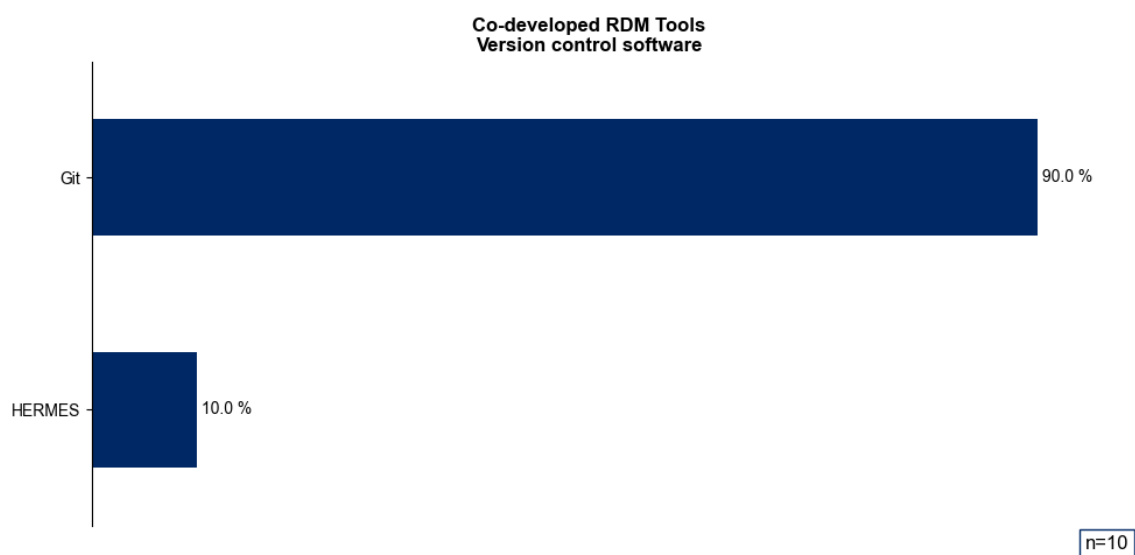OEP Tools — 14.3 %
Other — 14.3 %

n=7

Figure A19: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 7, relative amounts refer to n)

Figure A20: "Which of the following tool categories are (co-)developed in your organizational level/unit?" (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 13, relative amounts refer to n)
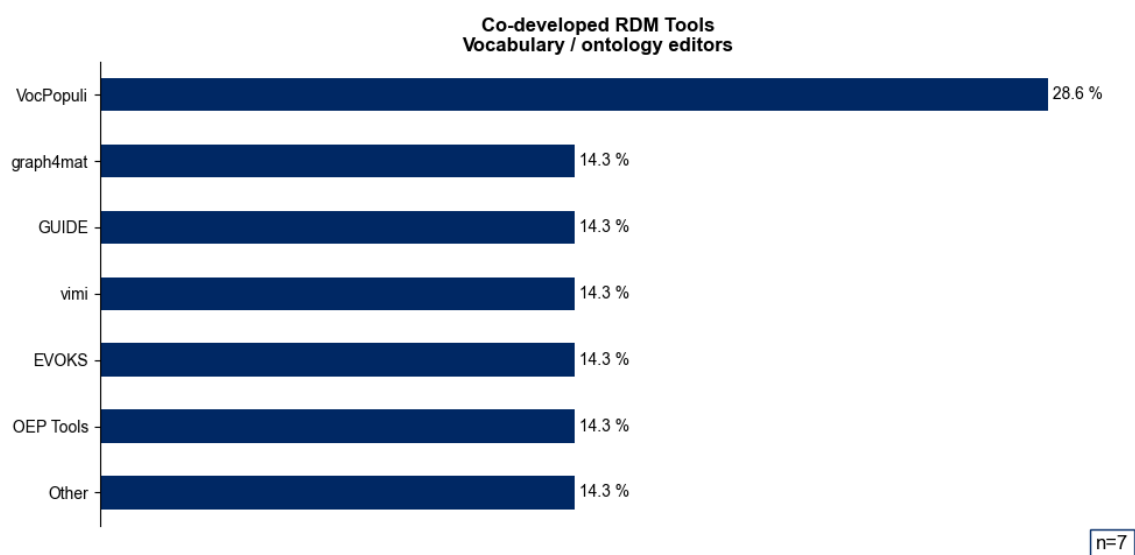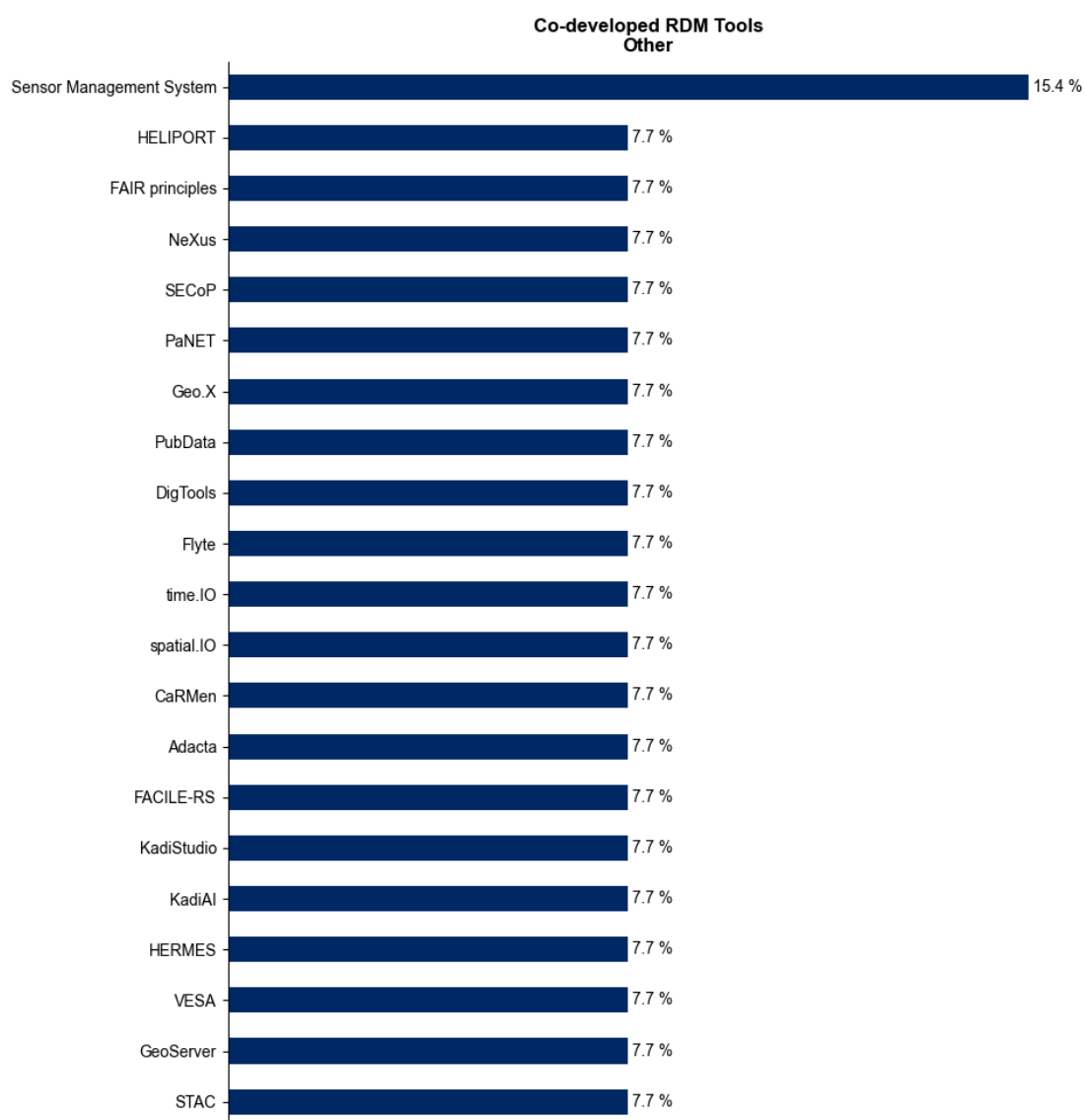
Figure A21: "What obstacles or difficulties do you encounter when working on research data management-related tasks?" (Question type: multiple choice. Available to all respondents. Number of respondents who answered this question: n = 152, relative amounts refer to n)
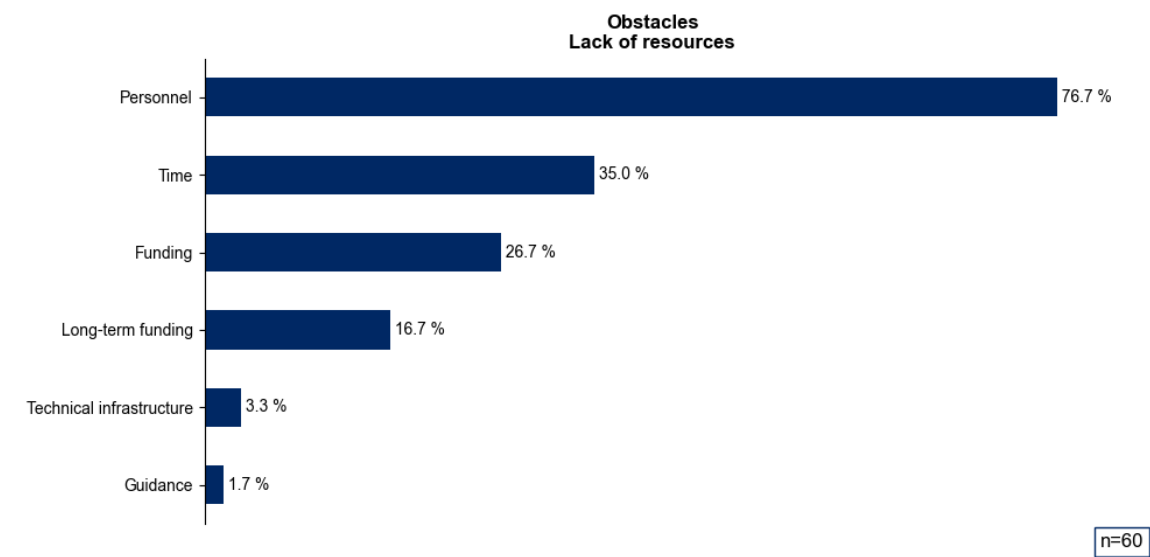
Figure A22: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 60, relative amounts refer to n)

**Obstacles
Missing technical knowledge**

Skill gap	50.0 %
Other	10.7 %
Lack of Training and Support	7.1 %
Lack of tools	7.1 %
Complexity and Choice	7.1 %
Personnel	7.1 %
Lack of proper data management	3.6 %
Lack of standards	3.6 %
Time	3.6 %
Lack of Domain Expertise	3.6 %
Lack of motivation	3.6 %

n=28

Figure A23: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 28, relative amounts refer to n)
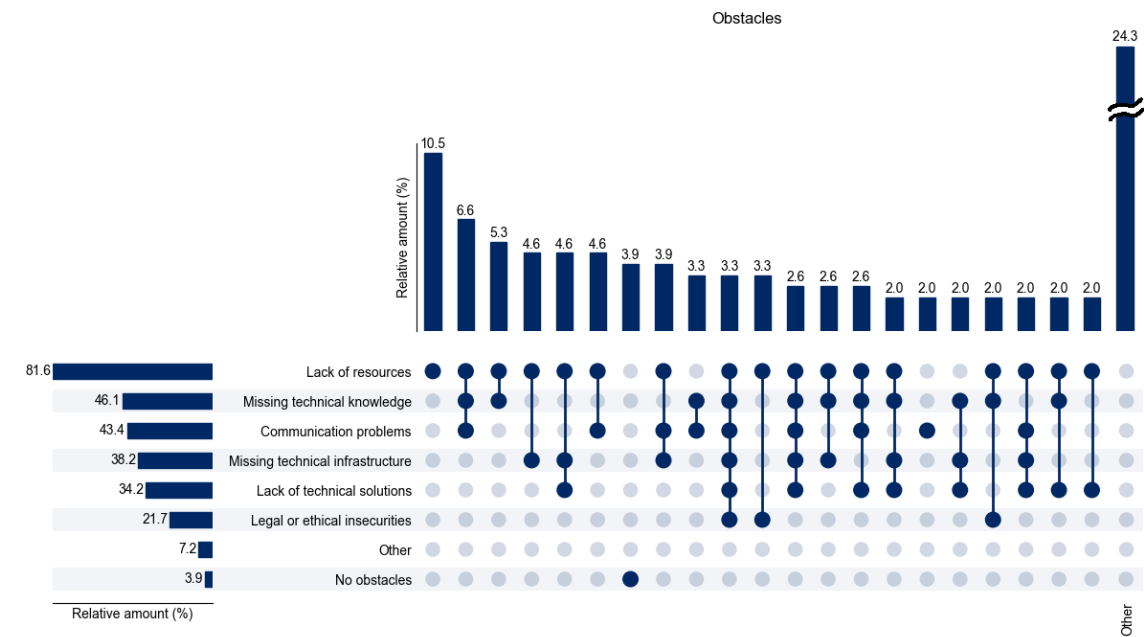
**Obstacles
Lack of technical solutions**

|	
Other	22.7 %
Data Catalogs and Discovery	22.7 %
Data Standards	13.6 %
Access Control	9.1 %
Data Provenance and Integrity	4.5 %
Collaboration Tools	4.5 %
Data Citation and Identifiers	4.5 %
Personnel	4.5 %
Versioning	4.5 %
Offline Data Storage	4.5 %
User Interface and Experience	4.5 %
Interoperability	4.5 %

n=22

Figure A24: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 22, relative amounts refer to n)

HELMHOLTZ

Obstacles
Legal or ethical insecurities

Lack of Legal Knowledge	25.0 %
Other	18.8 %
Time consuming	18.8 %
GDPR Implementation Challenges	18.8 %
Complicated	6.2 %
Lack of Legal Collaboration	6.2 %
Administrative Burden	6.2 %
Data Authorship	6.2 %
Lack of Clear Guidelines	6.2 %

n=16

Figure A25: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 16, relative amounts refer to n)
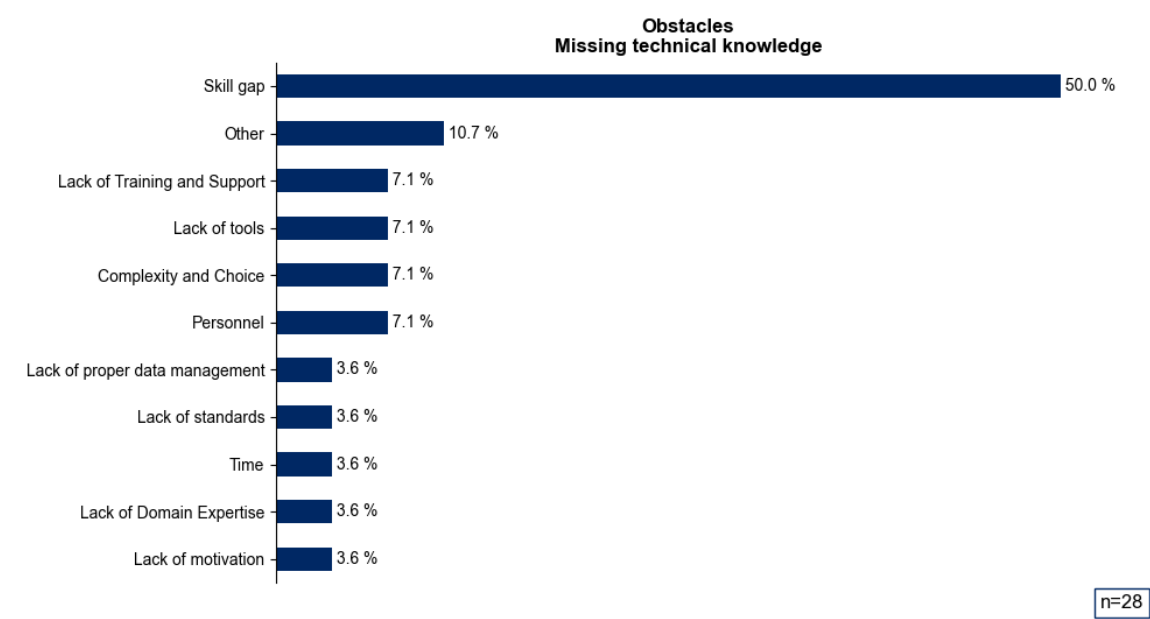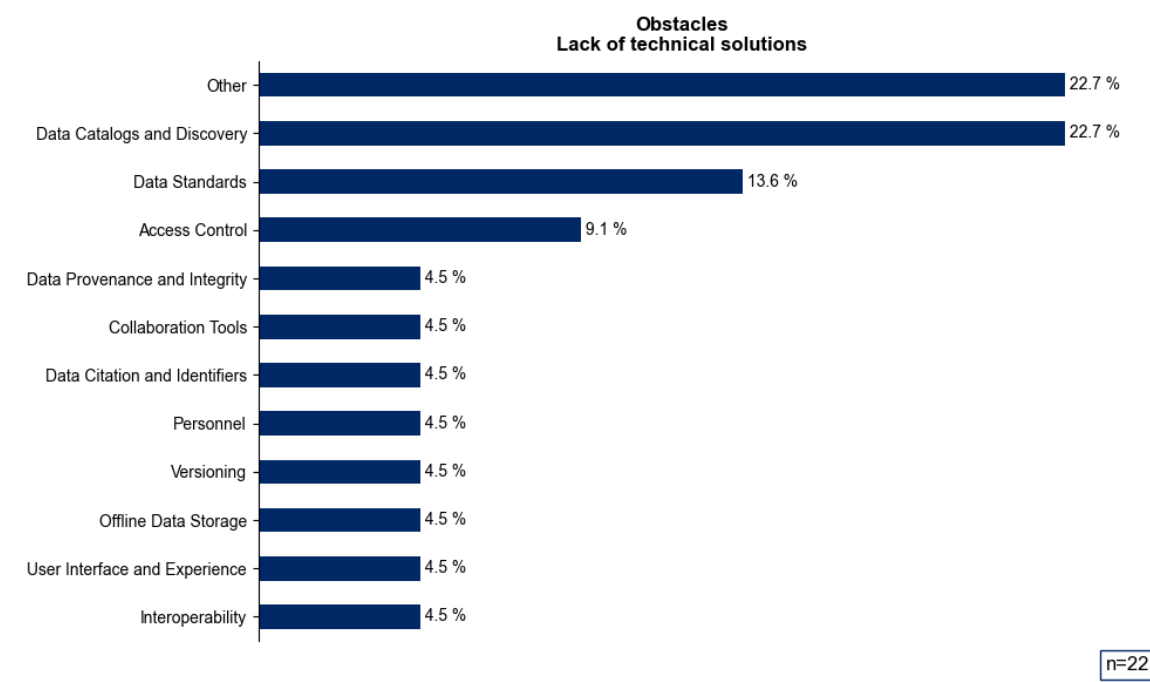

Obstacles
Missing technical infrastructure

Data management solutions	46.4 %
Other	21.4 %
Data Storage	10.7 %
Security Breach	7.1 %
Integration of standards into existing tools	3.6 %
Data Catalog	3.6 %
Laboratory Workflow Integration	3.6 %
Access Control	3.6 %
Persistent Identifiers	3.6 %
Time Constraints	3.6 %
Personnel	3.6 %
Automated Workflows	3.6 %

n=28

Figure A26: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 28, relative amounts refer to n)
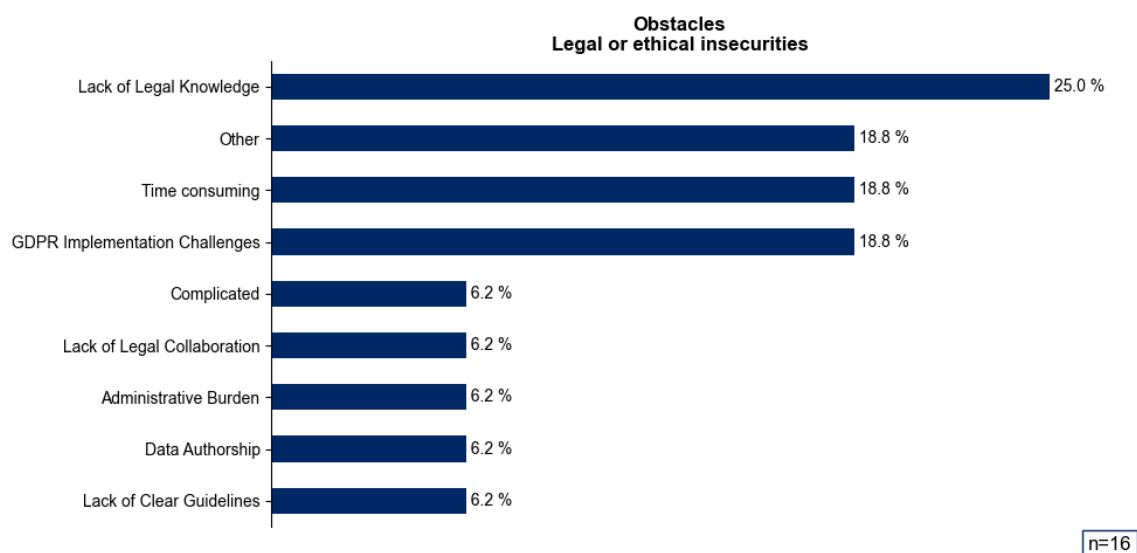
HELMHOLTZ

**Obstacles
Communication problems**

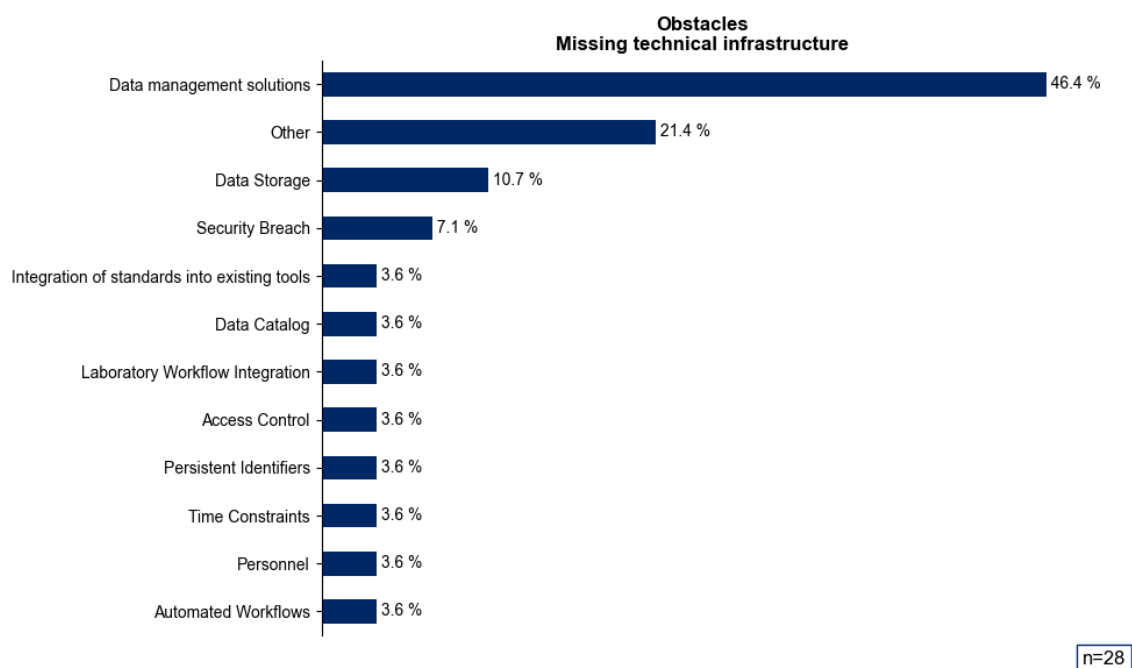| Obstacle	Percentage
Collaboration Difficulties	25.8 %
Lack of Knowledge	19.4 %
Data Sharing Reluctance	9.7 %
Standards Implementation Challenges	6.5 %
Communication Delays	6.5 %
Promoting Open Science	6.5 %
Other	6.5 %
Data Discoverability Issues	3.2 %
Lack of knowledge	3.2 %
Documentation and Guidance Needs	3.2 %
Lack of Personnel	3.2 %
Project Management Issues	3.2 %
Policy Implementation Challenges	3.2 %
Lack of Time	3.2 %
Data Management Culture and Adoption Challenges	3.2 %
Expectation Management	3.2 %
Unclear responsibilities	3.2 %

n=31

Figure A27: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 31, relative amounts refer to n)
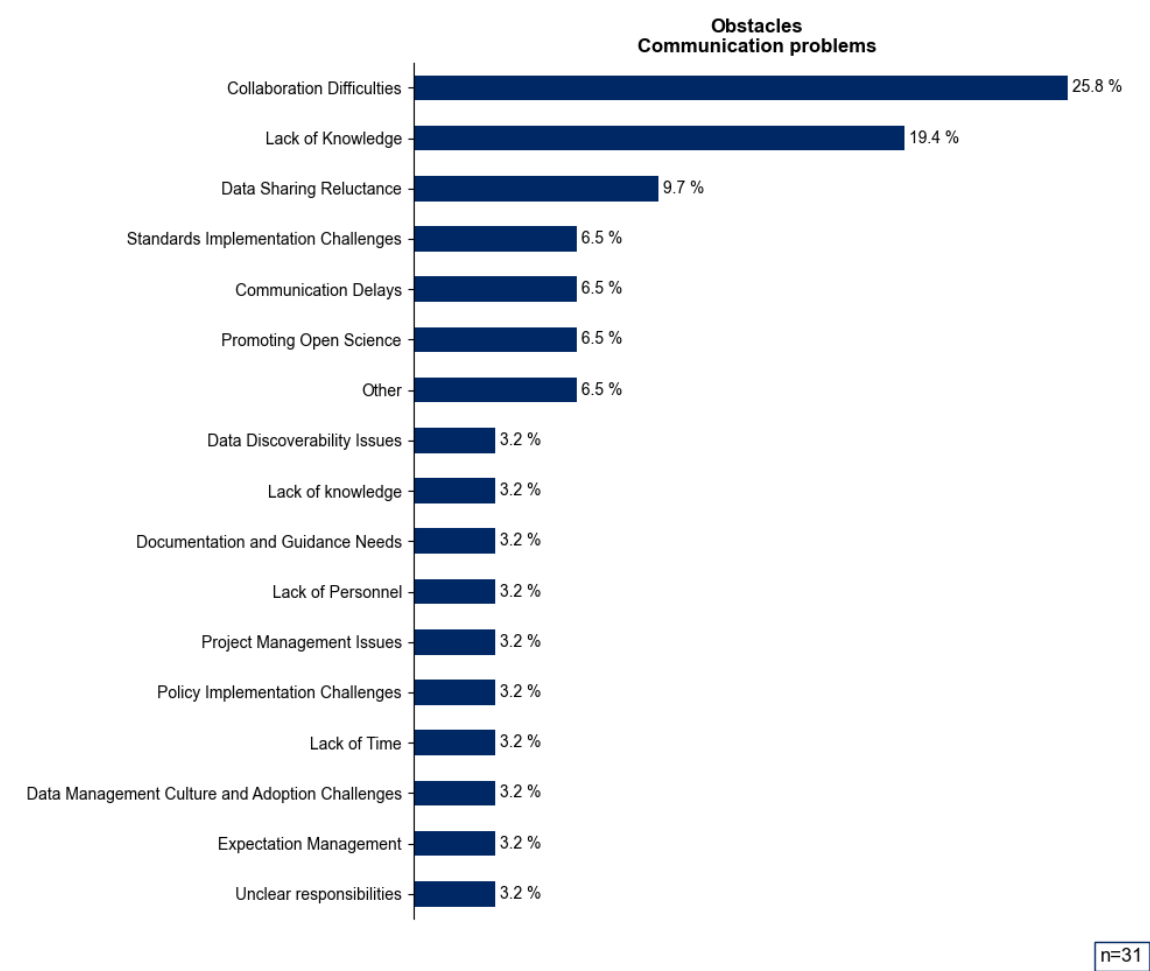
**Obstacles
Other**

Obstacle	Percentage
Lack of data exchange across tools	28.6 %
Lack of best practices	28.6 %
Lack of interconnecting standards	14.3 %
Lack of quality control for data publications	14.3 %
Lack of long-term funding	14.3 %

n=7

Figure A28: "What obstacles or difficulties do you encounter when working on research data management-related tasks? Please specify." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 7, relative amounts refer to n)
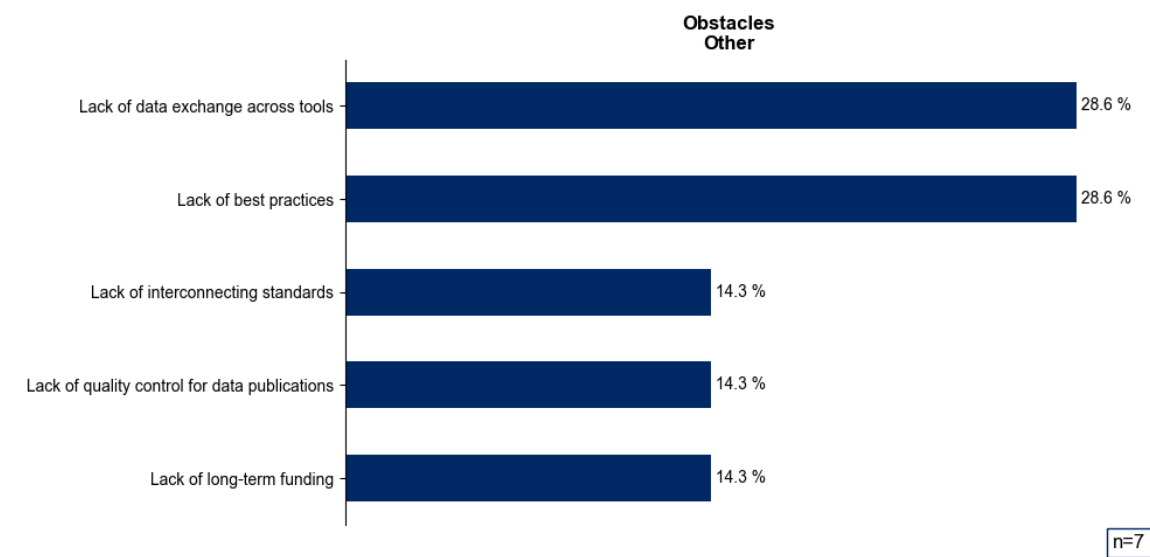
HELMHOLTZ

Figure A29: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 133, relative amounts refer to n)
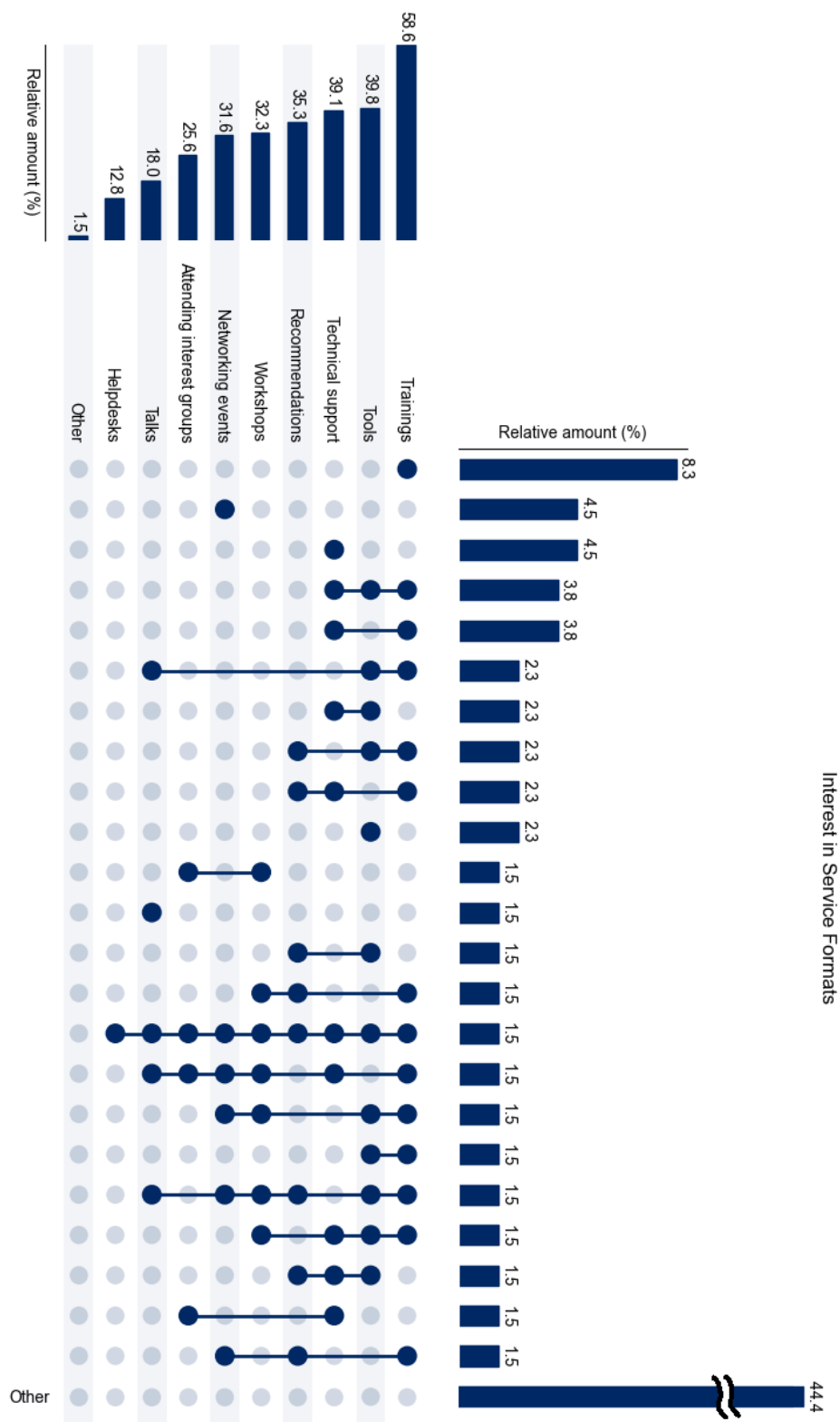
Figure A30: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 34, relative amounts refer to n)
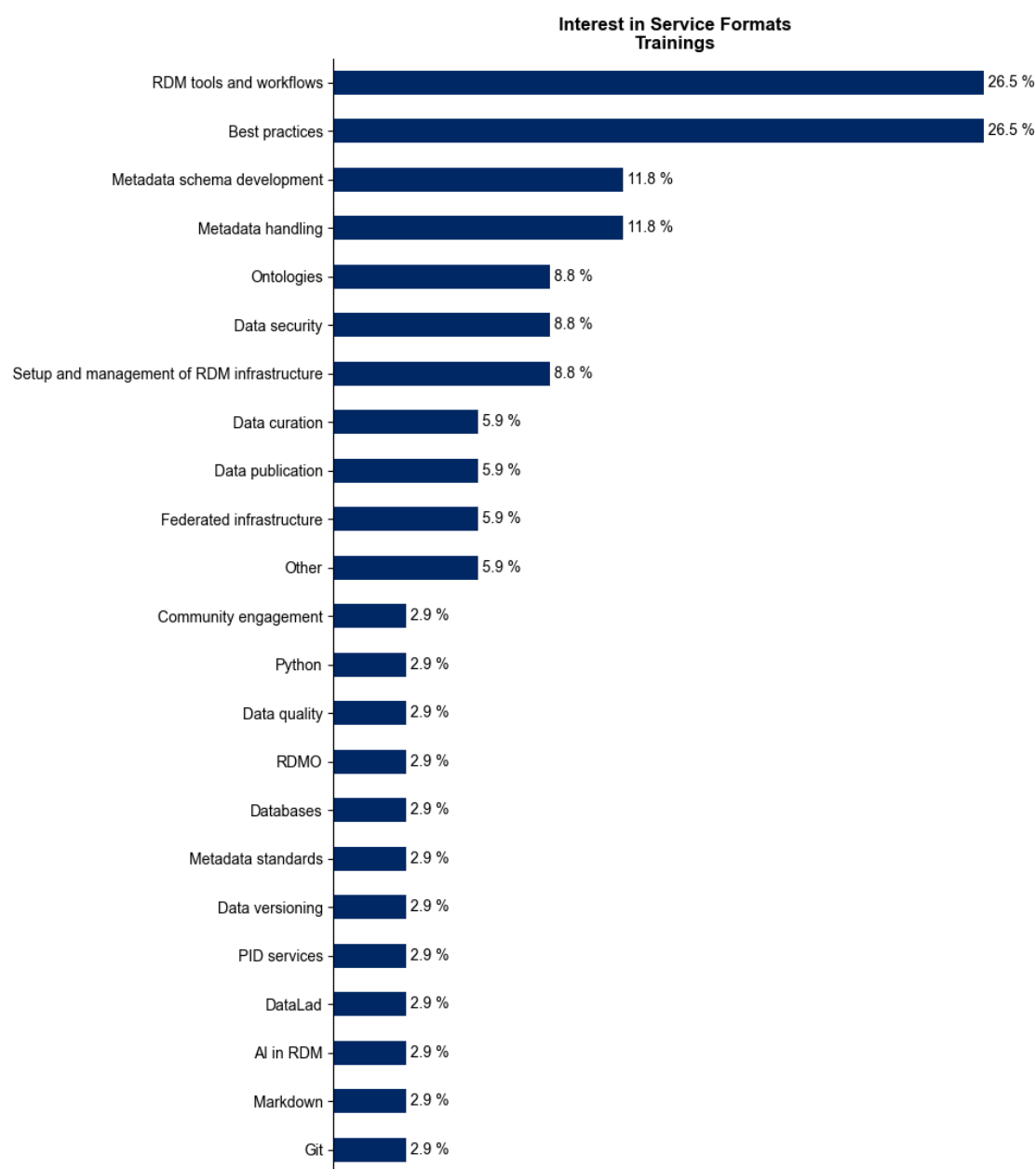
**Interest in Service Formats
Technical support**

Figure A31: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 21, relative amounts refer to n)
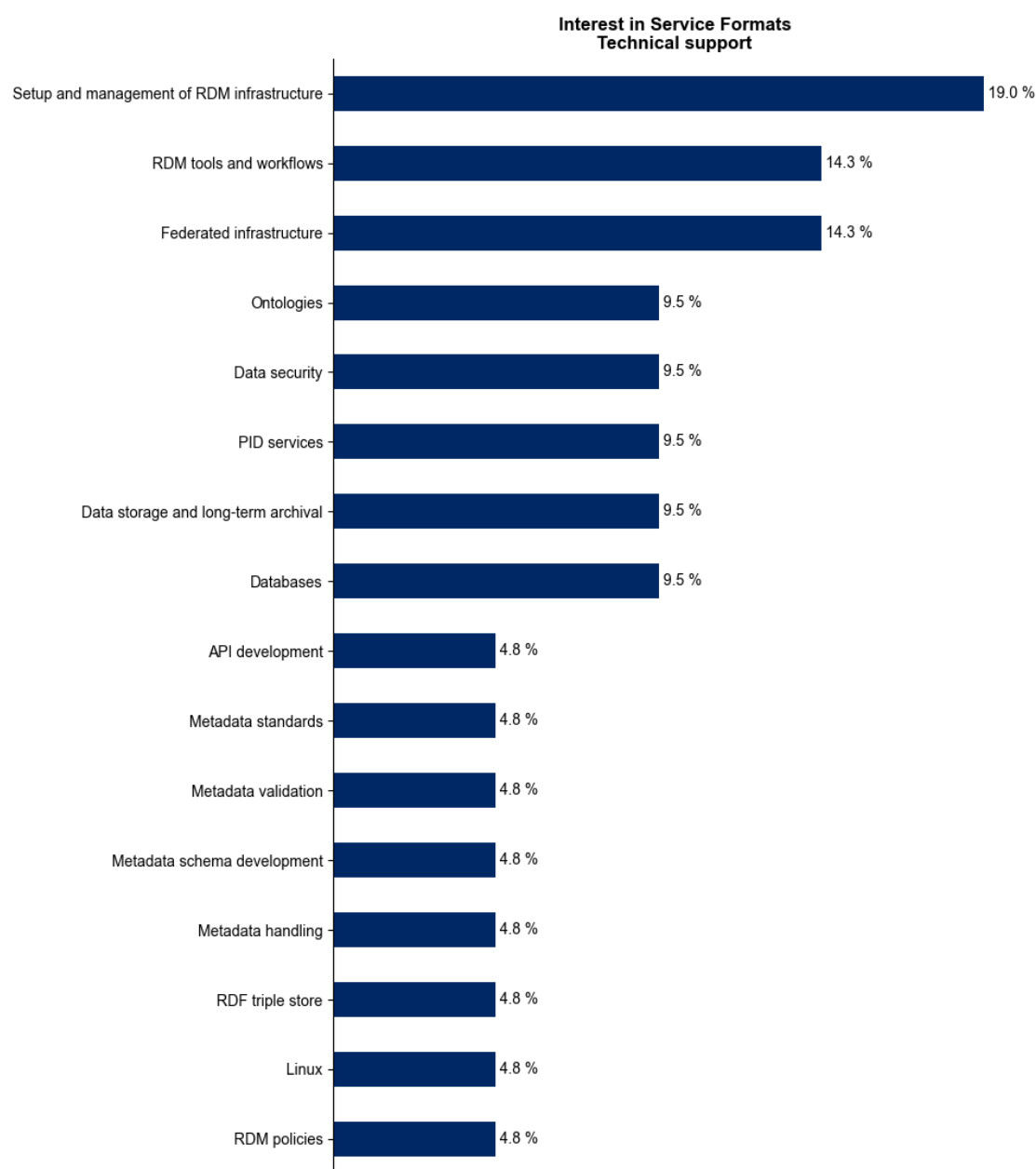
Figure A32: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 16, relative amounts refer to n)

Interest in Service Formats
Talks

Topic	Percentage
RDM tools and workflows	40.0 %
Best practices	30.0 %
Local implementations of RDM	20.0 %
Experiences with RDM implementations	10.0 %
AI in RDM	10.0 %
Findability of data	10.0 %
Synergies in RDM	10.0 %
Metadata standards	10.0 %

n=10

Figure A33: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 10, relative amounts refer to n)
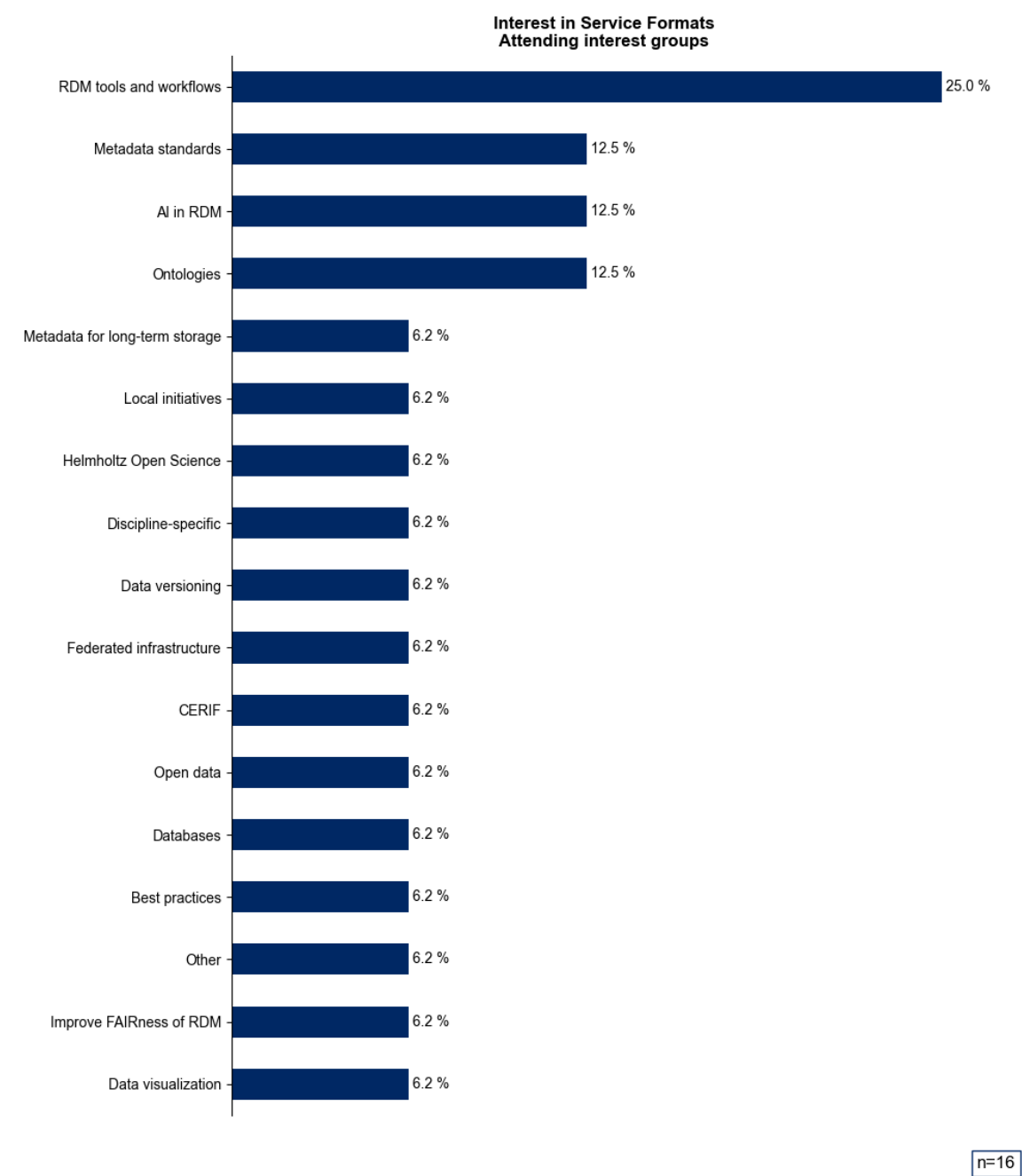
Interest in Service Formats
Tools

Topic	Percentage
Setup and management of RDM infrastructure	18.8 %
Git	12.5 %
Data storage and long-term archival	12.5 %
Ontologies	12.5 %
Federated infrastructure	6.2 %
RDMO	6.2 %
Data versioning	6.2 %
Data provenance	6.2 %
FAIR assessment tools	6.2 %
Collaboration Tools	6.2 %
Metadata standards	6.2 %
Community engagement	6.2 %
API development	6.2 %
Markdown	6.2 %
Metadata handling	6.2 %
Findability of data	6.2 %
Data visualization	6.2 %
Data management plan	6.2 %

n=16

Figure A34: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 16, relative amounts refer to n)
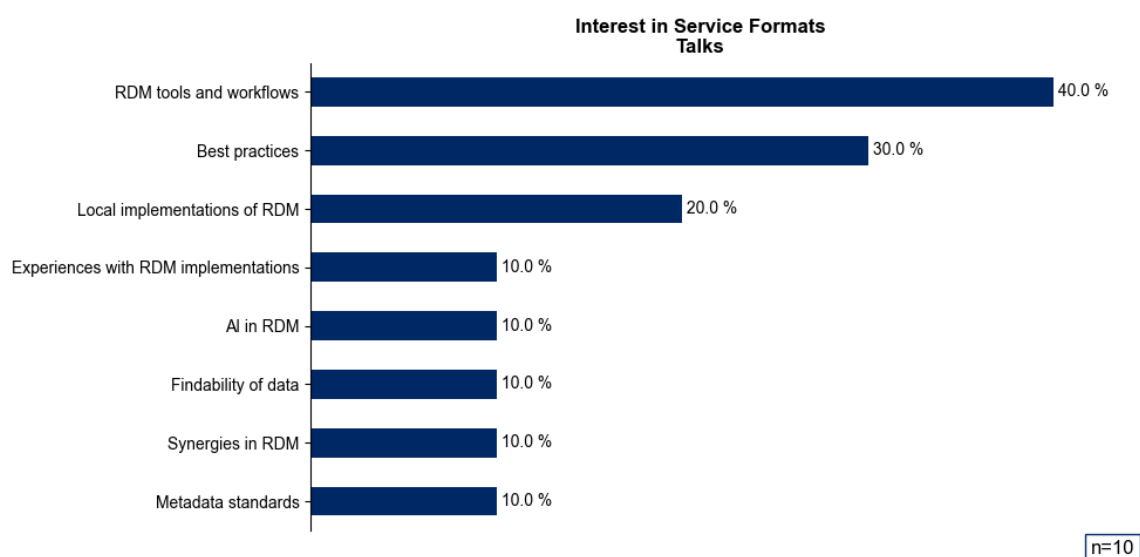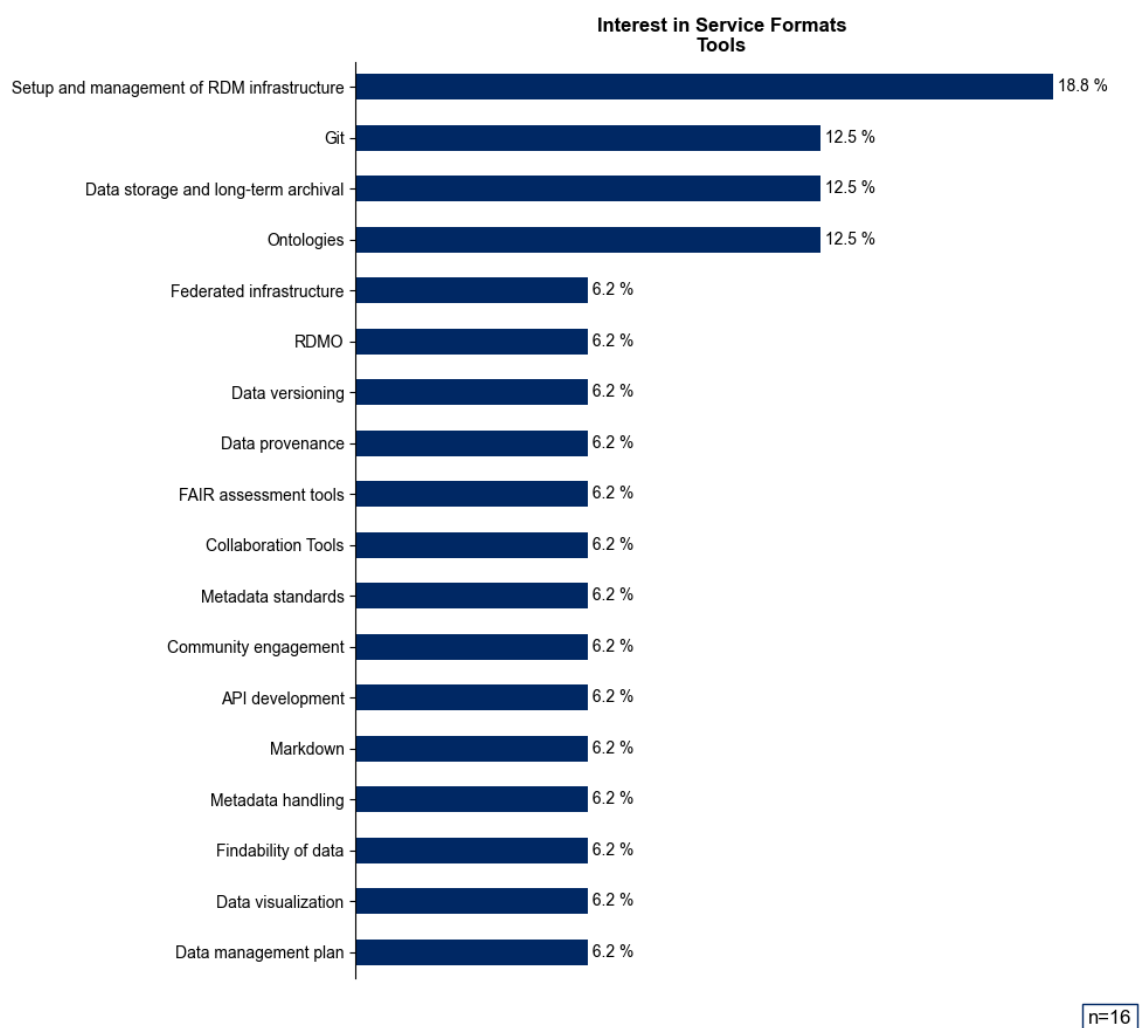
HELMHOLTZ

**Interest in Service Formats
Recommendations**

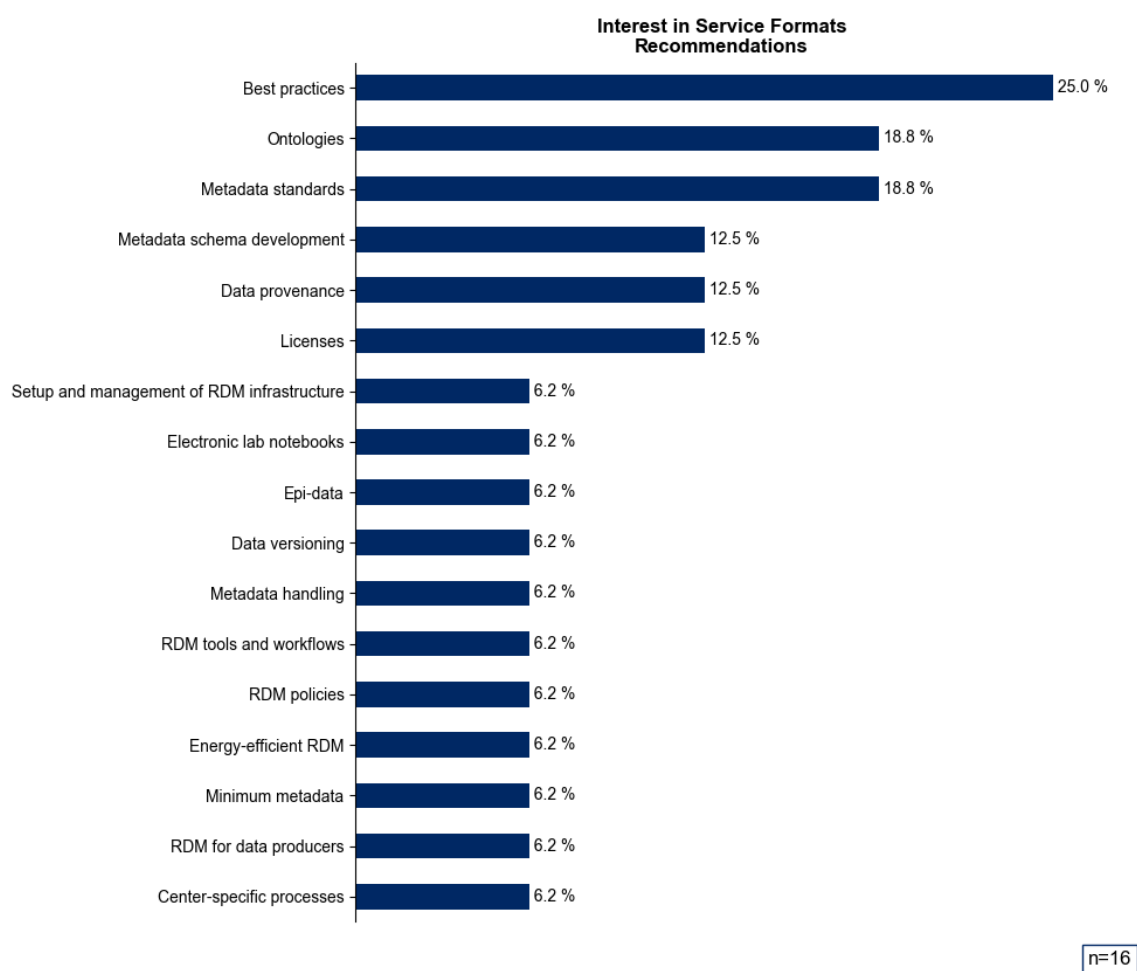| Topic	Percentage
Best practices	25.0 %
Ontologies	18.8 %
Metadata standards	18.8 %
Metadata schema development	12.5 %
Data provenance	12.5 %
Licenses	12.5 %
Setup and management of RDM infrastructure	6.2 %
Electronic lab notebooks	6.2 %
Epi-data	6.2 %
Data versioning	6.2 %
Metadata handling	6.2 %
RDM tools and workflows	6.2 %
RDM policies	6.2 %
Energy-efficient RDM	6.2 %
Minimum metadata	6.2 %
RDM for data producers	6.2 %
Center-specific processes	6.2 %

n=16

Figure A35: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 16, relative amounts refer to n)
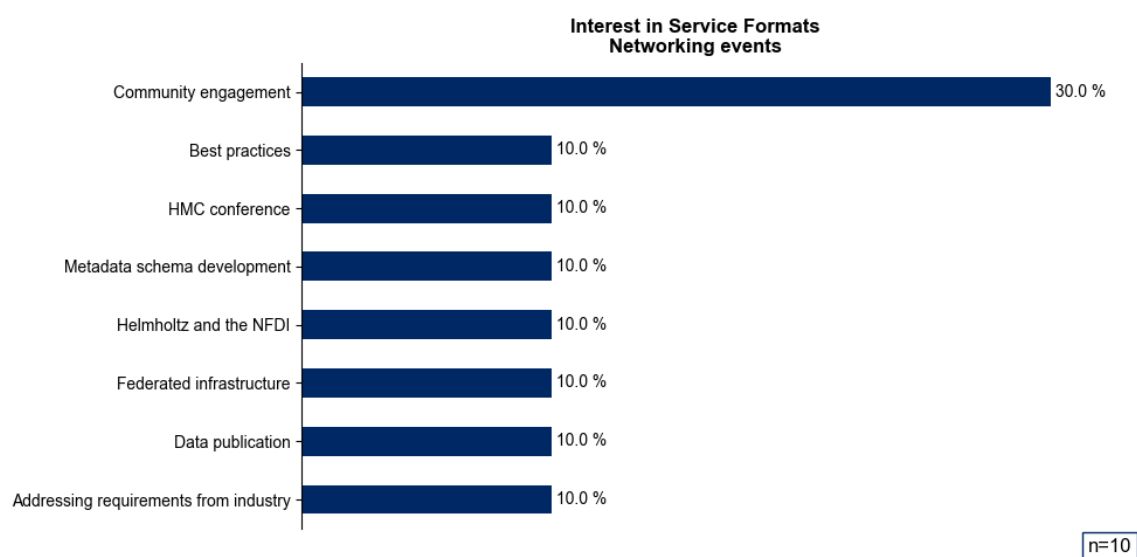
**Interest in Service Formats
Networking events**

Topic	Percentage
Community engagement	30.0 %
Best practices	10.0 %
HMC conference	10.0 %
Metadata schema development	10.0 %
Helmholtz and the NFDI	10.0 %
Federated infrastructure	10.0 %
Data publication	10.0 %
Addressing requirements from industry	10.0 %

n=10

Figure A36: "Which service formats would you be interested in? Please specify topics." (Question type: Multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 10, relative amounts refer to n)
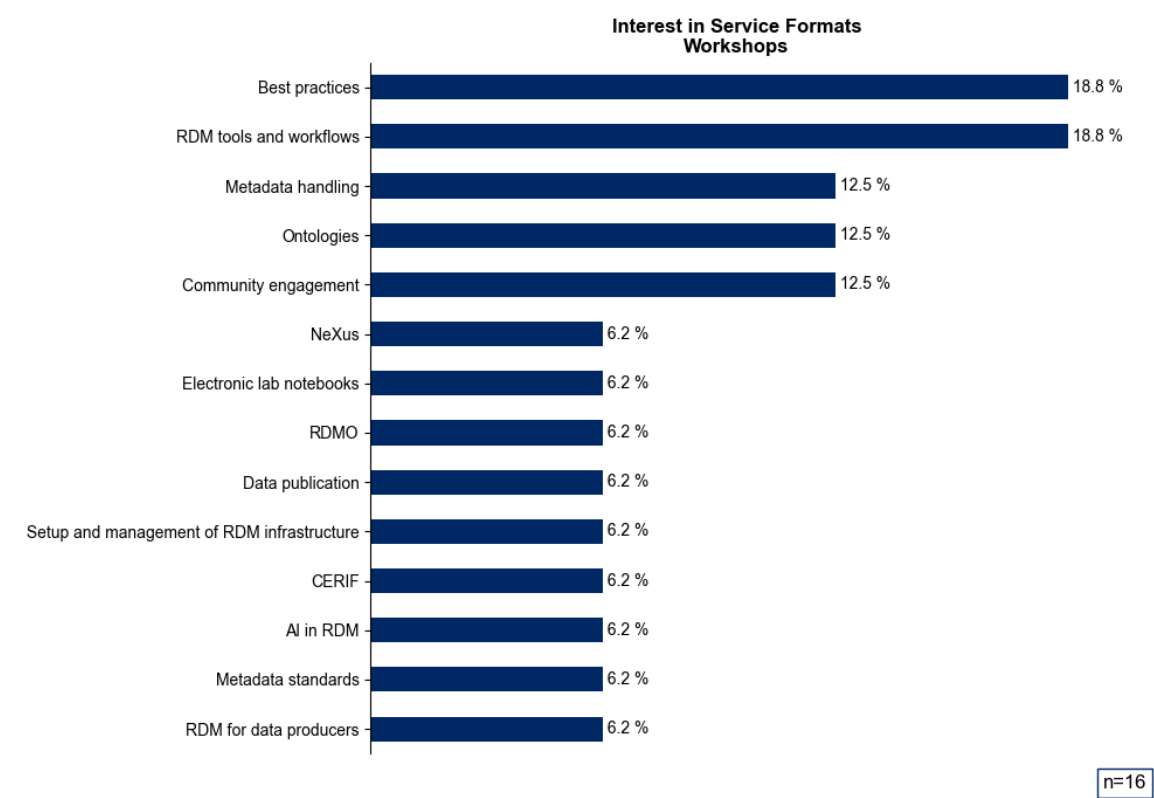
HELMHOLTZ

Figure A37: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 16, relative amounts refer to n)
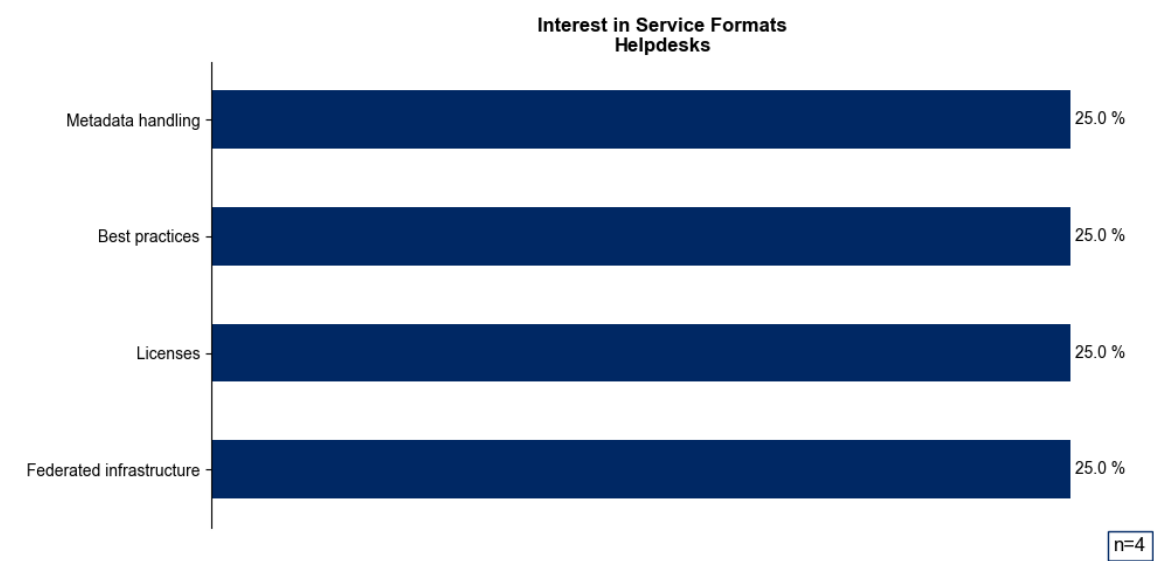

Figure A38: "Which service formats would you be interested in? Please specify topics." (Question type: multiple choice with comments. Available to all respondents. Number of respondents who answered this question: n = 4, relative amounts refer to n)