HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HMC REPORT | 1 | HMC Community Survey 2021

A survey on research data management practices among researchers in the Helmholtz Association

December 2022

HMC HELMHOLTZ METADATA COLLABORATION

Keywords: Survey, Helmholtz Association, Research data management, Metadata, Researcher, FAIR

DOI: 10.3289/HMC_publ_05

Citation: Helmholtz Metadata Collaboration; A survey on research data management practices among researchers in the Helmholtz Association, 2022.

Acknowledgement
This publication was supported by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

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)**:** Witold Arndt (0000-0002-7713-9647), Silke Christine Gerlich (0000-0003-3043-5657), Volker Hofmann (0000-0002-5149-603X), Markus Kubin (0000-0002-2209-9385), Lucas Kulla (0000-0002-2484-2742), Christine Lemster (0000-0001-7764-1517), Oonagh Mannix (0000-0003-0575-2853), Katharina Rink (0000-0003-4874-3973), Marco Nolden (0000-0001-9629-0564), Jan Schweikert (0000-0003-4774-2717), Sangeetha Shankar (0000-0003-0387-7740), Emanuel Söding (0000-0002-4467-642X), Leon Steinmeier (0000-0001-9040-636X), Wolfgang Süß (0000-0003-2785-7736)

HMC group: Working group “Taskforce Survey”

Editors: Sören Lorenz, Ants Finke, Christian Langenbach, Klaus Maier-Hein, Stefan Sandfeld, Rainer Stotzka

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

Content

Abstract

Annotation of research data with rich metadata is important to make that data findable, accessible, interoperable, and reusable (Wilkinson et al. [2016]). This ensures the conducted research data is durable. Within the Helmholtz Association, the Helmholtz Metadata Collaboration (HMC) coordinates the mission to enrich Helmholtz-based research data with metadata by providing (information about) technical solutions, advice and ensuring uniform scientific standards for the use of metadata.

In 2021, HMC conducted its first community survey to align its services with the needs of Helmholtz researchers. A question catalogue with 49 (sub-)questions was designed and disseminated among researchers in all six Helmholtz research fields. The conditional succession of the questions was aligned with predetermined expertise levels ("no prior knowledge", "intermediate prior knowledge", "high level of prior knowledge"). 631 completed survey replies were obtained for analysis.

The HMC Community Survey 2021 provides insight into the management of research data as well as the data publication practices of researchers in the Helmholtz Association. The characterization of research-field-dependent communities will enable HMC to further develop targeted, community-directed support for the documentation of research data with metadata.

Zusammenfassung

Um die nachhaltige Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit von Forschungsdaten zu gewährleisten, ist die Annotation wissenschaftlichen Forschungsdaten mit reichhaltigen Metadaten Voraussetzung. Innerhalb der Helmholtz-Gemeinschaft koordiniert die Helmholtz Metadata Collaboration (HMC) die Bestrebungen, Helmholtz-Forschungsdaten mit Metadaten anzureichern, indem sie zu technischen Lösungen informiert und berät und für einheitliche wissenschaftliche Standards bei der Nutzung von Metadaten sorgt.

Mit dem Ziel, unser Portfolio an entwickelten Diensten, Kursen und Beratungen an den Bedürfnissen der Helmholtz-Forscherinnen und -Forscher auszurichten, führte HMC im Jahr 2021 eine Community-orientierte Umfrage durch. Gemäß einem kompetenzadaptiven Ansatz wurde ein Fragenkatalog mit insgesamt 49 (Unter-)Fragen entworfen und unter Forschenden aller sechs Helmholtz-Forschungsbereiche verteilt. 631 vollständig ausgefüllte Fragebögen wurden zur weiteren Auswertung berücksichtigt.

Die HMC Community-Umfrage 2021 bietet detaillierte Einblicke in den Umgang und das Management von Forschungsdaten sowie in die Datenveröffentlichungspraxis der Forschenden der Helmholtz-Gemeinschaft. Die Profilierung der Communities in den verschiedenen Forschungsbereichen ermöglicht es HMC, zielgerichtete und zielgruppenorientierte Unterstützungsangebote für die Dokumentation von Forschungsdaten mit Metadaten weiterzuentwickeln.

1 Introduction

Helmholtz Metadata Collaboration (HMC)

The Helmholtz Metadata Collaboration (HMC) platform enables researchers and data infrastructure providers to make their research data FAIR (Wilkinson et al. [2016]) - 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; Matter - to develop, share and consolidate community expertise in metadata together. HMC works towards the community-specific requirements that partially differ between research fields. Therefore, domain specific HMC Hubs were established in each HGF research field.

The HMC roadmap comprises measures for community building (including the provision of training), developing technical tools, generating FAIR policies and processes, gathering information on standards, and networking with national and international stakeholders. With its interdisciplinary mission, structure, and provided services; HMC is unique in the European science community.

THE HELMHOLTZ ASSOCIATION

The Helmholtz Association of German Research Centers (see 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, a think tank - The Helmholtz Information & Data Science Incubator - was established that aims at synergizing Helmholtz expertise and creating a future-oriented research environment. HMC represents one of its platforms.

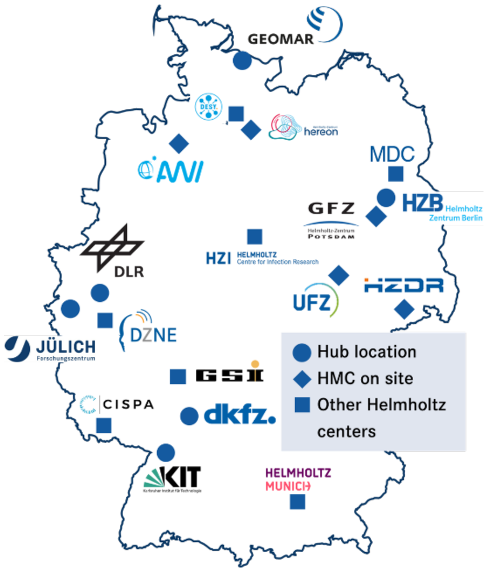


Figure 1: Helmholtz centers in Germany

HMC Community Survey 2021

To inquire and quantify the status and needs of the scientific community in each of the research fields, HMC conducted a survey in 2021. The goals of this survey were to investigate research data management practices, with a focus on metadata handling and data publication, as well as to identify demands within the different communities.

The survey data presented here informs the strategic planning of HMC by identifying service-and-knowledge gaps (both general and community specific), thereby enabling the design and establishment of services to advance the Helmholtz FAIR data landscape.

In the current document, we present the results of the HMC Community Survey 2021.

2 Survey Design and Analysis

Target group

The survey addressed the entire scientific staff in all Helmholtz research centres in Germany. Department leaders, institute leaders, and other staff working in the area of management and coordination were not a part of the target group. Survey respondents were expected to have variable degrees of expertise with respect to the application of the FAIR data guidelines. Therefore, the survey questionnaire was designed to dynamically adapt to the respondents' variable expertise levels across subject topics.

Survey goals

The HMC Community Survey 2021 aimed to get a deep insight into the research data management practices of scientific staff within the Helmholtz Association. Subject topics covered by the survey are data management and data publication practices, with the overall aim to identify service demands in the different research fields.

The following sub-goals were formulated to address the overall aim in a holistic manner:

1. Understand the status quo of research data management practices in Helmholtz's research communities.
   - 1.1. Identify demands in these research data management practices towards realizing data management practices in a FAIR manner.
   - 1.2. Qualitative collection of repositories, tools, formats and standards that are currently being used by the individual research communities
2. Identify needs and demands of Helmholtz's research communities regarding data management in FAIR manner.
   - 2.1. Identify needs and concerns explicitly expressed by the research communities.
   - 2.2. Understand susceptibility of research communities regarding different service contents and formats.

3. Establish a relationship with Helmholtz's research communities.
    3.1. Understand researchers' motivations for engaging in FAIR data practices.
    3.2. Understand what topics researchers in Helmholtz's research communities are working on or interested in.
    3.3. Understand how HMC can establish an open, effective and friendly communication with Helmholtz's research communities.

Survey questionnaire

Numerous surveys on the topic of research data management and data sharing practices have been conducted by third parties in the past [Herres-Pawlis et al. [2020], Fane et al. [2019], Radosavljevic et al. [2019], Brenger et al. [2019], Weng et al. [2018], Arndt et al. [2018], Hesse et al. [2017], Berghmans et al. [2017], Treadway et al. [2016], Paul-Stüve et al. [2015], Simukovic et al. [2013], Hauck et al. [2016], Feldsien-Sudhaus and Rajski [2016], Lemaire et al. [2016], Bauer et al. [2015], Krähwinkel [2015], Tenopir et al. [2015], Heinrich et al. [2015]]. These were used as references when preparing the present survey.

The survey questionnaire for the present survey was prepared by a dedicated task force of HMC ensuring representation from all Helmholtz research fields.

Target-group-oriented language was chosen for designing the questionnaire, allowing to be easily understood by researchers from all expertise levels. For example, the term "documentation" was used as a target-group-oriented term for the abstract and often ambiguous term "metadata". Questions, answer options, answer types, and survey logics were documented in YAML files.

The following types of questions were used for the survey.

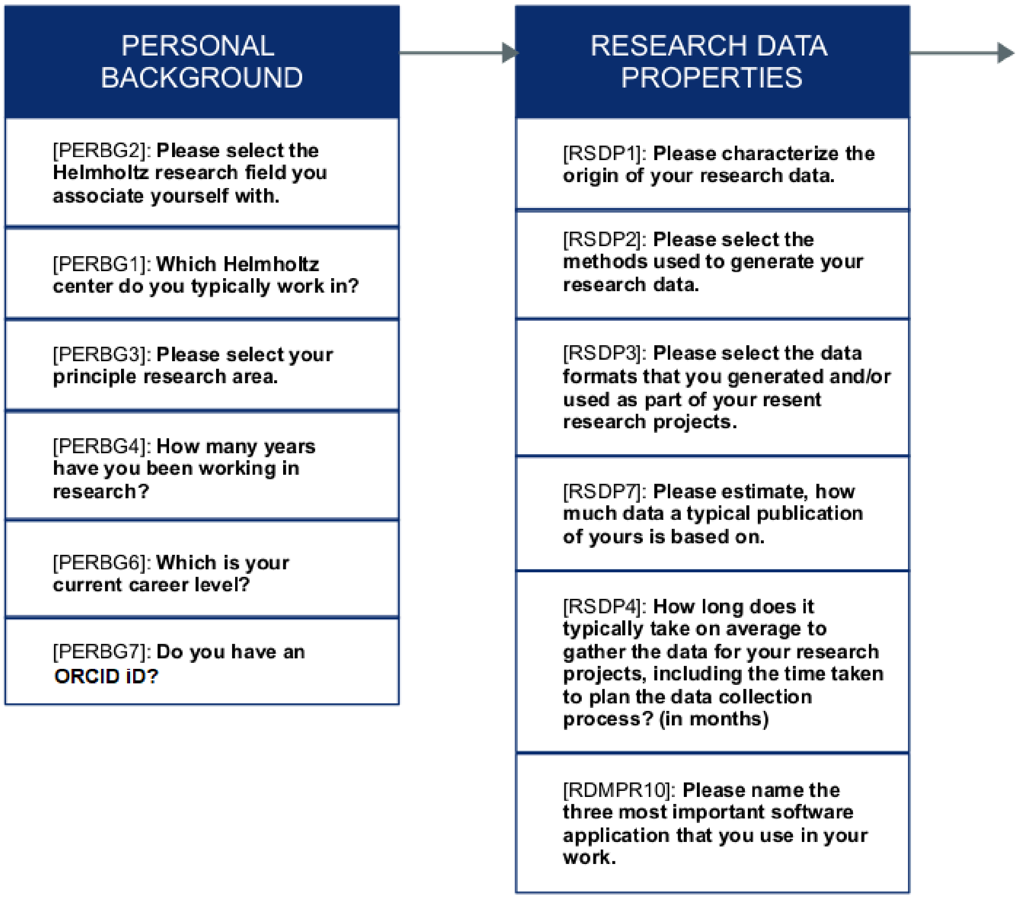
- In "single choice" questions, respondents were allowed to choose up to one answer option from a predefined list of answers.
- In "multiple choice" questions, respondents were allowed to choose a set of answers from a predefined list, ranging from none to all answer options. In the exceptional case that the maximum number of answers was limited, this was clearly indicated.
- In "matrix" questions, respondents could individually rate on categories (sub-questions) by choosing a single answer option out of a pre-defined list for each category.
- 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 an arbitrary text into one or more input field(s).

The final question catalogue[1] contains 49 (sub-)questions, grouped into five subject areas (question groups) which aim at addressing the following scopes, where survey logics were set to adapt questions dynamically to the respondents' experiences (see Figures 2 and 3).

Personal background

Question group “Personal background” (PERBG) aims at characterizing the survey respondents with respect to their Helmholtz centre and research field, in addition to their scientific discipline[2], career level, and research experience. This information is particularly valuable for the internal characterization of target-group-specific service needs.

In this question group, question PERBG4 was adopted from references [Herres-Pawlis et al. [2020], Hauck et al. [2016]], and PERBG8 from reference [Fane et al. [2019]].



[PERBG2]: Please select the Helmholtz research field you associate yourself with.

[PERBG1]: Which Helmholtz center do you typically work in?

[PERBG3]: Please select your principle research area.

[PERBG4]: How many years have you been working in research?

[PERBG6]: Which is your current career level?

[PERBG7]: Do you have an ORCID iD?

[RSDP1]: Please characterize the origin of your research data.

[RSDP2]: Please select the methods used to generate your research data.

[RSDP3]: Please select the data formats that you generated and/or used as part of your resent research projects.

[RSDP7]: Please estimate, how much data a typical publication of yours is based on.

[RSDP4]: How long does it typically take on average to gather the data for your research projects, including the time taken to plan the data collection process? (in months)

[RDMPR10]: Please name the three most important software application that you use in your work.

Figure 2: Sequence of questions in the survey (1/2). The boxes indicate the question group (blue solid boxes) and the questions in each question group (white boxes). Sequence of question was from top to bottom. Solid arrows indicate the order of question groups. Continuation in Figure 3. Abbreviations: PERBG - Personal Background; RSDP - Research data properties.

[1] Questions and answers are publicly available here: https://codebase.helmholtz.cloud/hmc/hmc-public/surveys/hmc-community-survey-2021/limesurvey.

[2] For the question “Please select your principle research area.” (PERBG3) and the follow-up questions “Please specify.”, answer options were adopted from the list of research disciplines in the FAIRsharing Subject Ontology (SRAO), https://fairsharing.org/ontology/subject/SRAO.owl.

Research data properties

Question group “Research data properties” (RSDP) aims at characterizing the research data generated or used by the respondents, as well as the data sources, methods, tools and data formats used. This information is particularly useful to understand the nature and the origin of the data, Helmholtz’s research communities are working with and to correlate these with specific service needs.

In this question group, question RSDP1 was adopted from references [Weng et al. [2018], Simukovic et al. [2013], Lemaire et al. [2016]], RSDP2b from [Paul-Stüve et al. [2015], Weng et al. [2018], Simukovic et al. [2013], Lemaire et al. [2016]], RSDP3 from [Paul-Stüve et al. [2015], Weng et al. [2018], Simukovic et al. [2013], Bauer et al. [2015], Krähwinkel [2015], Hauck et al. [2016]], RSDP7 from [Berghmans et al. [2017]], RSDP4 from [Brenger et al. [2019]], and RSDP10 was adopted from references [Paul-Stüve et al. [2015], Weng et al. [2018], Simukovic et al. [2013]].

Data publishing

Question group “Data publishing” (DTPUB) addresses the respondents’ experience in making some of their research data publicly available. A particular focus is on the motivations and challenges experienced by the respondents, which helps illuminating the experience and perception of the topic in the various research communities throughout the Helmholtz Association.

Information from this question group will help to understand the research communities’ habits and experiences in publishing research data as one of many first steps towards FAIR data, as well as to identify researchers’ motivations and obstacles towards (not) publishing research data.

In this question group, question DTPUB1b was adopted from references [Fane et al. [2019], Berghmans et al. [2017], Herres-Pawlis et al. [2020]], DTPUB3 from [Fane et al. [2019], Bauer et al. [2015]], DTPUB4a from [Fane et al. [2019]], and DTPUB4b was adopted from reference [Bauer et al. [2015]].

Research data management practices

Question group “Research data management practices” (RDMPR) aims at the respondents’ practices regarding their research data management as well as their related motivations and difficulties experienced herein. A particular focus of this question group is how the respondents’ research data is documented: Respondents are asked whether this is done in an analogue or digital way, in an unstructured or structured way and which metadata categories are being considered for describing research data. Consequently, respondents with increasingly advanced research data management practices are asked questions that require increasing levels of expertise. Examples of questions that are shown to respondents with presumably advanced research data management practices are, for example, whether and which international standards are used to describe the data.

Information from this question group will help to characterize how advanced data management practices are in Helmholtz's various research communities and to develop and address related needs for services and support.

In this question group, question RDMPR1 was adopted from references [Paul-Stüve et al. [2015], Weng et al. [2018], Simukovic et al. [2013], Bauer et al. [2015], Lemaire et al. [2016], Heinrich et al. [2015], Hauck et al. [2016], Feldsien-Sudhaus and Rajski [2016]].

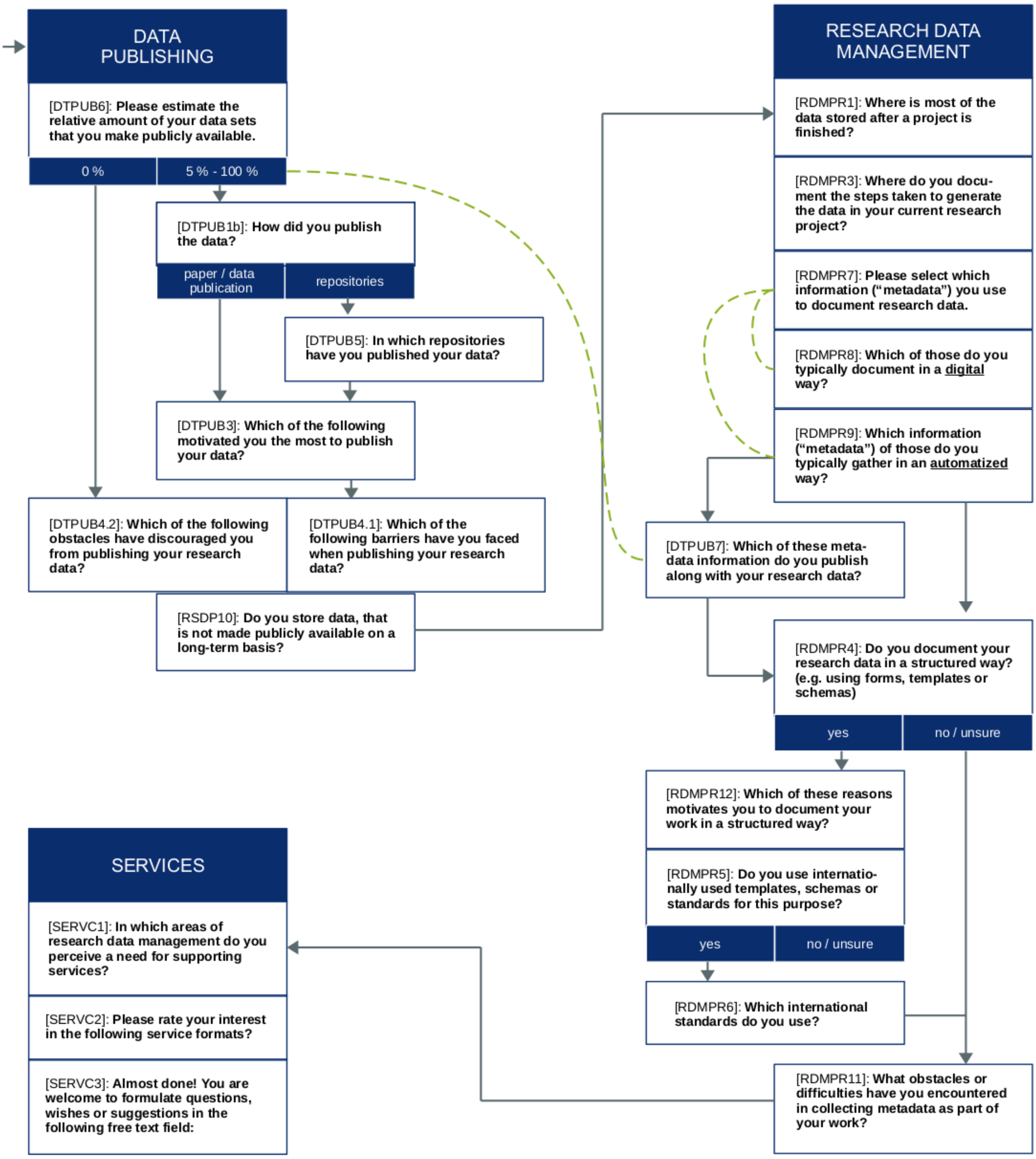

DATA PUBLISHING
[DTPUB6]: Please estimate the relative amount of your data sets that you make publicly available.
0 %
5 % - 100 %
[DTPUB1b]: How did you publish the data?
paper / data publication
repositories
[DTPUB5]: In which repositories have you published your data?
[DTPUB3]: Which of the following motivated you the most to publish your data?
[DTPUB4.2]: Which of the following obstacles have discouraged you from publishing your research data?
[DTPUB4.1]: Which of the following barriers have you faced when publishing your research data?
[RSDP10]: Do you store data, that is not made publicly available on a long-term basis?
RESEARCH DATA MANAGEMENT
[RDMPR1]: Where is most of the data stored after a project is finished?
[RDMPR3]: Where do you document the steps taken to generate the data in your current research project?
[RDMPR7]: Please select which information ("metadata") you use to document research data.
[RDMPR8]: Which of those do you typically document in a digital way?
[RDMPR9]: Which information ("metadata") of those do you typically gather in an automatized way?
[DTPUB7]: Which of these meta-data information do you publish along with your research data?
[RDMPR4]: Do you document your research data in a structured way? (e.g. using forms, templates or schemas)
yes
no / unsure
[RDMPR12]: Which of these reasons motivates you to document your work in a structured way?
[RDMPR5]: Do you use internationally used templates, schemas or standards for this purpose?
yes
no / unsure
[RDMPR6]: Which international standards do you use?
[RDMPR11]: What obstacles or difficulties have you encountered in collecting metadata as part of your work?
SERVICES
[SERVC1]: In which areas of research data management do you perceive a need for supporting services?
[SERVC2]: Please rate your interest in the following service formats?
[SERVC3]: Almost done! You are welcome to formulate questions, wishes or suggestions in the following free text field:

Figure 3: Sequence of questions in the survey (2/2). Continuation of Figure 2. The boxes indicate the question group (blue solid boxes) and the questions in each question group (white boxes below). Sequence of question was from top to bottom, following along gray solid arrows. Blue split boxes at the bottom of some questions indicate conditional logical flow following this question depending on the entered response. This ensured that the difficulty of the survey was adaptive to the level of the respondents. Abbreviations: DTPUB - Data Publishing; RDMPR - Research Data Management Practices; SERVC - Services

Services

In the last section of the survey, question group “Services” (SERVC) addresses the respondents’ perceived need for support in various topics of research data management, as well as related service formats they are potentially interested in.

Information from this question group can be correlated with answers from previous question groups to develop and address community-specific service formats.

In this question group, question SERVC1 was adopted from references [Fane et al. [2019], Paul-Stüve et al. [2015], Weng et al. [2018], Simukovic et al. [2013], Krähwinkel [2015]], and SERVC2 from references [Bauer et al. [2015]].

Implementation and dissemination

For each of the six Helmholtz research fields, community-specific subsets of questions were derived from the complete catalogue of questions and answers. Hence, community-specific questions and answers were included or excluded for each subset.

A landing page served as an entry point to the survey containing general information on the survey and data privacy[3]. To enter the community-specific surveys, participants were asked to associate themselves with one of the Helmholtz research fields.

The survey was implemented in LimeSurvey[4], a browser-based open source survey software for online surveys. Survey data was collected in a fully anonymized way. Upon submission of the survey data, respondents were offered to optionally enter their contact information. This information was stored separately and cannot be connected or traced back to the respondents’ individual survey answers.

An invitation for participating in the survey was distributed to researchers across all Helmholtz centres. Furthermore, it was distributed by the HMC hubs using various dissemination strategies and no incentive was provided for participating in the survey. Survey data was collected from September to November 2021.

Data analysis

Data analysis was performed in Python, using the python framework “hifis-surveyval”[5] developed by the Helmholtz Federated IT-Services (HIFIS). The responses from hub-specific surveys were merged and used for analysis.

[3] https://helmholtz-metadaten.de/en/hmc-community-survey-2021

[4] https://github.com/LimeSurvey/LimeSurvey

[5] https://pypi.org/project/hifis-surveyval/

The raw data was processed in the following ways before analysis:

- Free text answers which contained sensitive information about the participant were removed manually. Long free-text answers were rephrased and answers in German were translated to English.
- Survey replies that were not completed until submission on the last page of the main survey were removed. 647 out of 1211 (53.4%) responses were found to be complete. 346 responses (28.6%) were incomplete and 218 responses (18%) were empty.
- Sixteen answers from respondents, who could not be associated with the target group of this survey, were removed, such as respondents working primarily in management and coordination.
- Free-text answers were mapped to existing answer categories, wherever applicable. During this process, new categories were identified, which could be used in future surveys.

The final data set contains 631 responses and is publicly available (Gerlich et al. [2022]). Before the data publication the following information was removed or anonymized from the survey data in order to prevent the identification of individuals:

1. Any information – including that might reveal a respondent's institutional affiliation
2. Names of software that is used by less than 4 respondents
3. Any information about institutional repositories

Custom python scripts[6] were added to the HIFIS framework to analyse the data. These scripts generated normalized count plots for all variables to understand the overall trend. These plots were also generated individually for each research field. For multiple choice questions, UpSet[7] plots were generated to identify which combinations were popular among the respondents.

[6] https://codebase.helmholtz.cloud/hmc/hmc-public/surveys/hmc-community-survey-2021/analysis

[7] https://upset.app/

3 Survey Results

Background of survey respondents

The HMC Community Survey 2021 was directed at understanding the research data management practices of researchers within the Helmholtz Association. Research programs in the Helmholtz association are distributed over six research fields, covering a wide variety of scientific expertise. It is therefore crucial to understand the (scientific) background of the survey respondents in order to develop community- and career-specific support for the scientific staff of the Helmholtz Association in handling their meta- and research data.

Responses were obtained from all Helmholtz research centres and the Helmholtz Institute in Mainz:

- Alfred-Wegener-Institute (AWI)
- Helmholtz Center for Information Security (CISPA)
- Deutsches Elektronen-Synchrotron (DESY)
- German Cancer Research Center (DKFZ)
- German Aerospace Center (DLR)
- German Center for Neurodegenerative Diseases (DZNE)
- Forschungszentrum Jülich (FZJ)
- Helmholtz Centre for Ocean Research Kiel (GEOMAR)
- German Research Centre for Geosciences (GFZ)
- Helmholtz Centre for Heavy Ion Research (GSI)
- Helmholtz-Zentrum Hereon
- Helmholtz Institut Mainz (HIM)
- Helmholtz Zentrum München - German Research Center for Environmental Health (HMGU)
- Helmholtz-Zentrum Berlin für Materialien und Energie (HZB)
- Helmholtz-Zentrum Dresden-Rossendorf (HZDR)
- Helmholtz Centre for Infection Research (HZI)
- Karlsruhe Institute of Technology (KIT)
- Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)
- Helmholtz Centre for Environmental Research (UFZ)

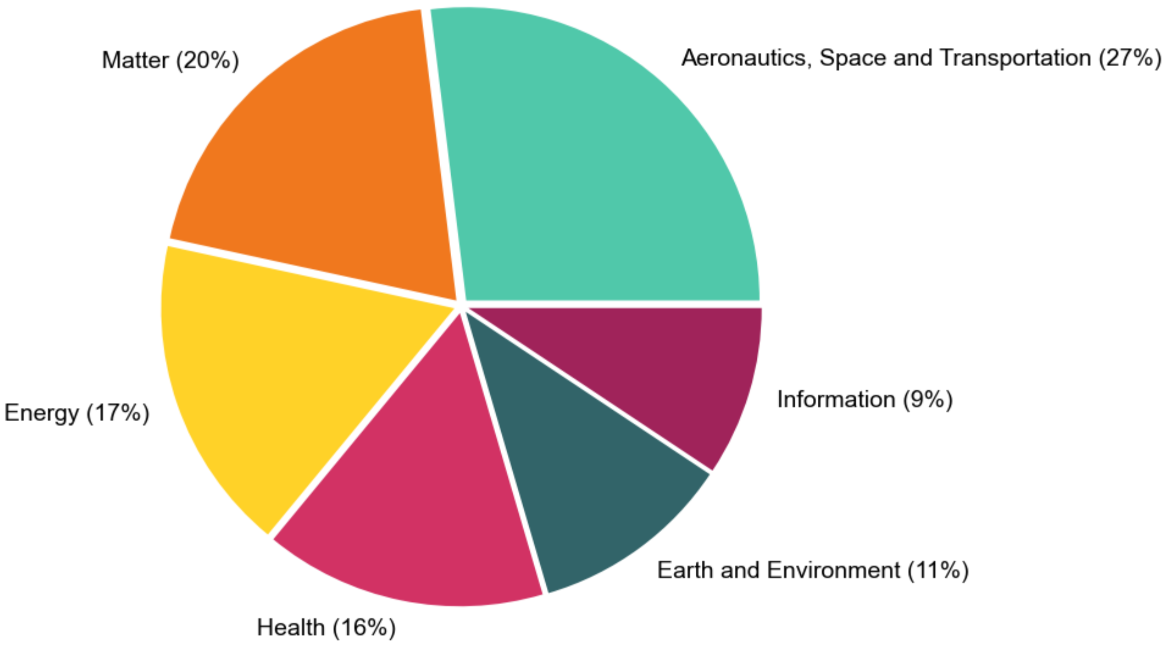


Figure 4: "Please select the Helmholtz research field you associate yourself with." (Single-choice question, available to all respondents, number of respondents who answered this question: n = 631, relative amounts refer to n)

Figure 4 shows the relative amount of survey replies for each of the six Helmholtz research fields. Most respondents (27%) assigned themselves to the research field "Aeronautics, Space and Transportation" (AST). 20% of all respondents assigned themselves to the research field "Matter", followed by the research field "Energy" (17%), research field "Health" (16%) and research field "Earth and Environment" (E&E) (11%). 9% of all respondents assign themselves to research field "Information".

Figure 5 shows the respondents' primary research discipline for each of the six Helmholtz research fields. In the research field E&E, most respondents assigned themselves to the discipline Earth Science (58.4%), followed by Life Science (20%). In the research field Health, most respondents came from Life Sciences (77.7%), followed by Engineering Science (8.5%). Most respondents from research field Matter assigned themselves to Physics (78.2%), followed by Engineering Science (10.5%). Research fields Energy, Information and AST resemble one another closely. In all three fields, most respondents assigned themselves to Engineering Science, followed by Physics.

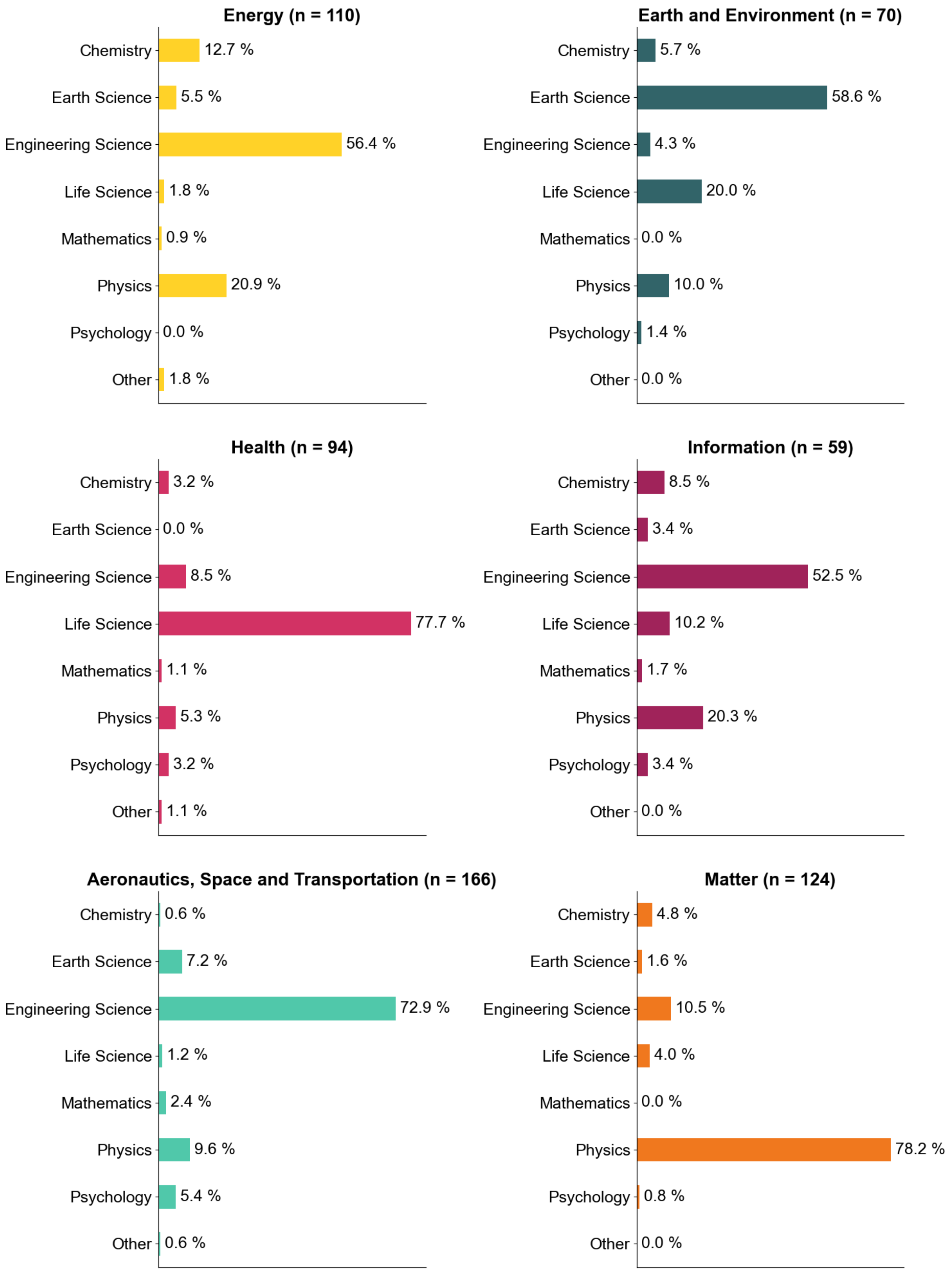


Figure 5: “Please select your principle research area.”. In each panel, relative amounts refer to the total number of respondents who previously assigned themselves to that Helmholtz research field and then answered this question. (Single-choice question, available to all respondents, number of respondents who answered this question: n = 623)

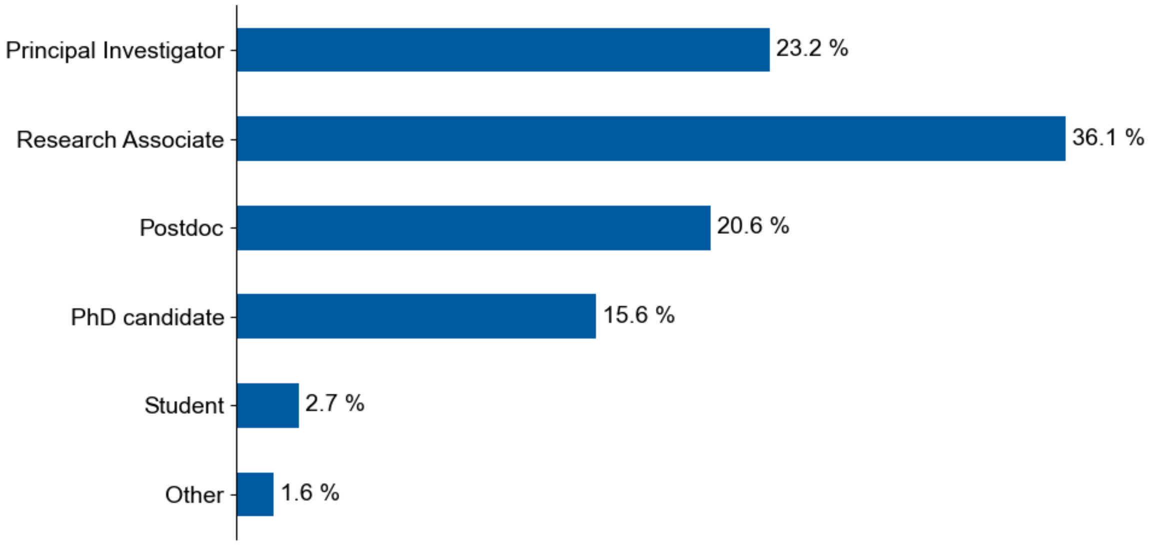


Figure 6:"Which is your current career level?". (Single-choice question, available to all respondents, number of respondents who answered this question: n = 620, relative amounts refer to n)

Figure 6 depicts the career level of the respondents. Most responses were obtained from Research Associates (36.1%), Principal Investigators (23.2%) and Postdocs (20.6%). 15.6% of all respondents are currently working on their PhD thesis.

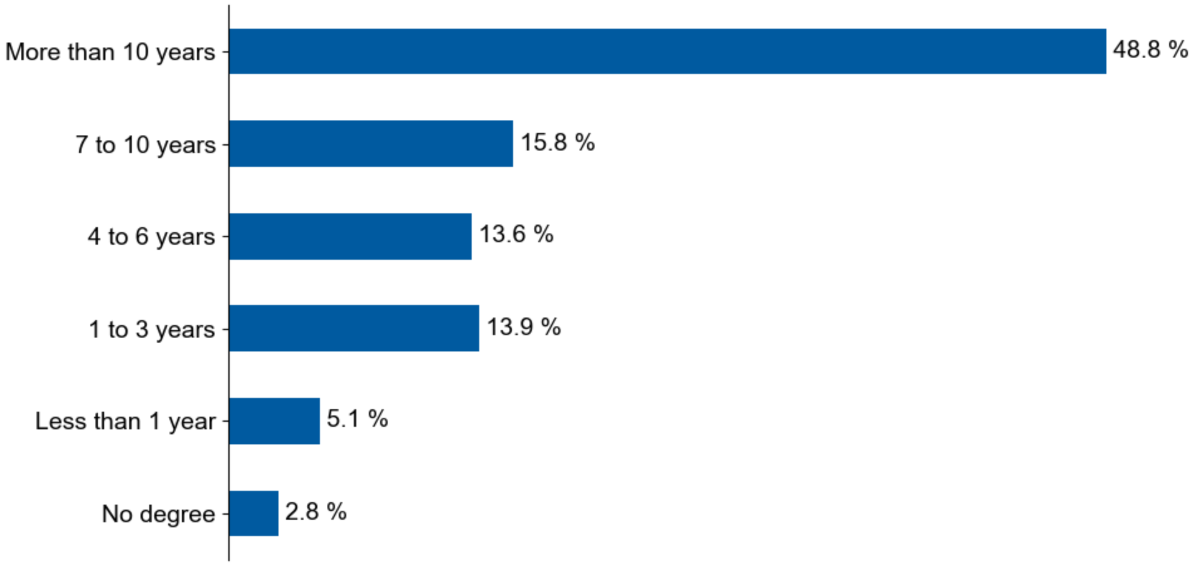


Figure 7: "How many years have you been working in research?". Question not available to respondents who previously assigned themselves to the Helmholtz research field Health. (Single-choice question, number of respondents who answered this question: n = 531, relative amounts refer to n)

The large number of survey respondents from later career stages observed in Figure 6 is also reflected in the distribution of the respondents' research experience (Fig. 7). About two-thirds (64.6%) of the participants have been working in research for 7 years and longer.

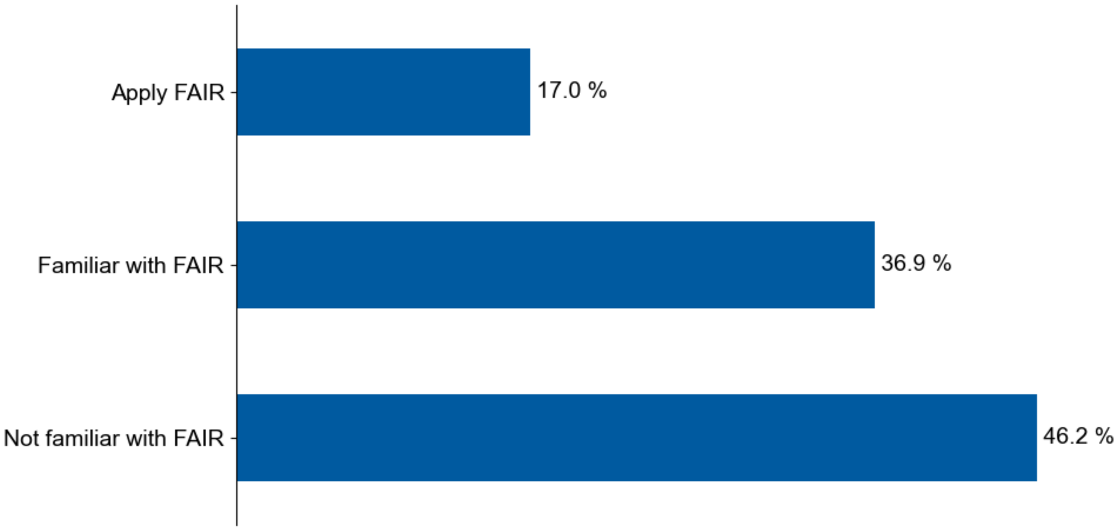


Figure 8: “How familiar are you with the FAIR data guidelines?”. Question not available to respondents who previously assigned themselves to the Helmholtz research field Information. (Single-choice question, number of respondents who answered this question: n = 559, relative amounts refer to n)

By answering the question “How familiar are you with the FAIR data guidelines?” (Fig. 8), the participants were challenged to reflect on their knowledge about data annotation following the FAIR guidelines. Intriguingly, almost half of the respondents concluded, that they are not familiar with the FAIR guidelines. On the other hand, it is noticeable that 36.9% of the respondents considered themselves familiar with the FAIR guidelines and 17% of the respondents already apply the FAIR guidelines. The large number of participants expressing unfamiliarity with the FAIR guidelines highlights the need for training on this topic.

Research data properties

Data origin and formats play a crucial role for annotating research data with expressive (structured) metadata. Hence, accessing the properties of research data generated in the Helmholtz research centres is of great importance, in order to understand the research communities in depth. To inspect the community-specific habits of research data acquisition and to identify service demands accordingly, participants were asked to specify the origin of their research data, the methods used for data acquisition, and their most commonly generated data formats.

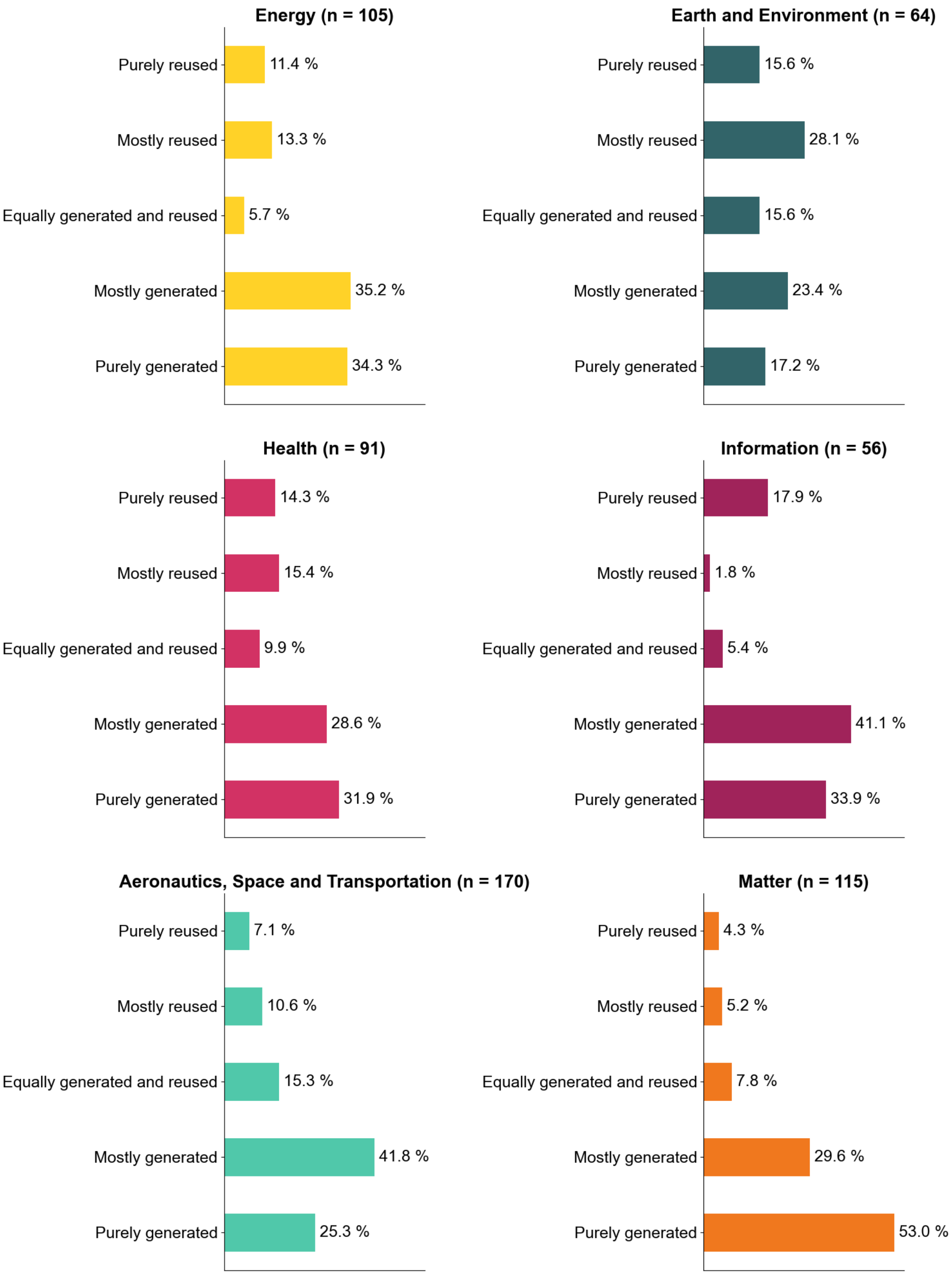


Figure 9: "Please characterize the origin of your research data.". Self-assessed origin of the respondents' research data, on a scale from "purely reused" to "purely self-generated". In each panel, relative amounts refer to the total number of respondents who previously assigned themselves to that Helmholtz research field and then answered this question. (Single-choice "slider" question, available to all respondents, number of respondents who answered this question: n = 601)

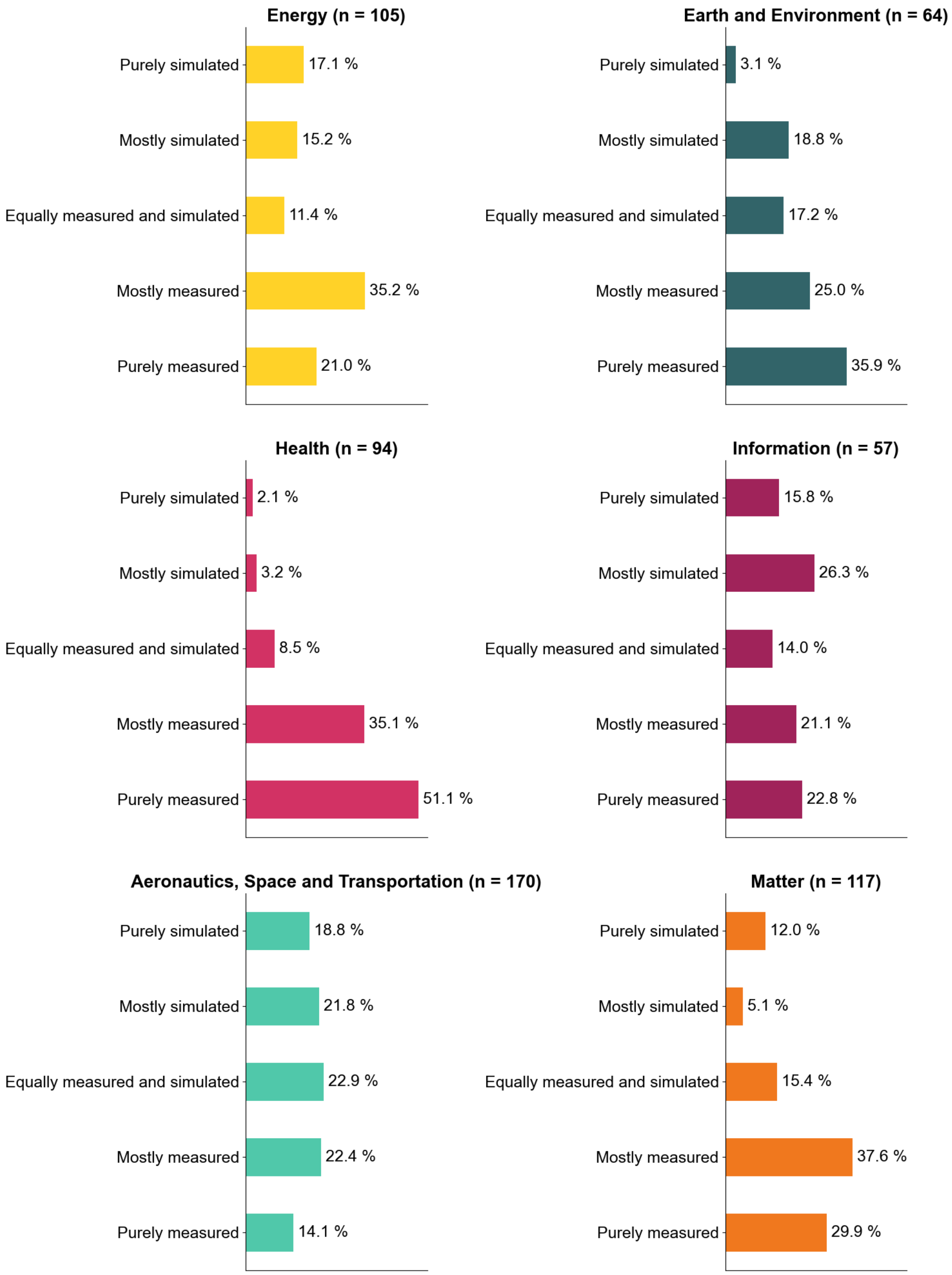


Figure 10: "Please characterize the origin of your research data.". Self-assessed origin of the respondents' research data on a scale from "purely simulated" to "purely experimental". In each panel, relative amounts refer to the total number of respondents who previously assigned themselves to that Helmholtz research field and then answered this question. (Single-choice "slider" question, available to all respondents, number of respondents who answered this question: n = 607)

In most research fields, responding researchers indicated to predominantly work with self-generated data (Fig. 9). The research field Earth & Environment constitutes an exception in this regard, as responses are roughly evenly distributed, spanning the whole spectrum from purely reused to purely self-generated data. Considering the research data origin in terms of measurements vs. simulations, two distinct patterns can be identified across the six Helmholtz research fields (Fig. 10). Data in research fields Energy, Earth & Environment, Health, and Matter originates predominantly from measurements. This pattern is most distinctive in the research fields Health and Matter, and moderately marked in the fields Energy and Earth & Environment, where a notable number of scientists derive their research data mostly or purely from simulations. In contrast, no notable pattern in the data origin can be detected in the research fields AST and Information. Researchers in these fields derive their data from measurements, simulations and any combination of both methods.

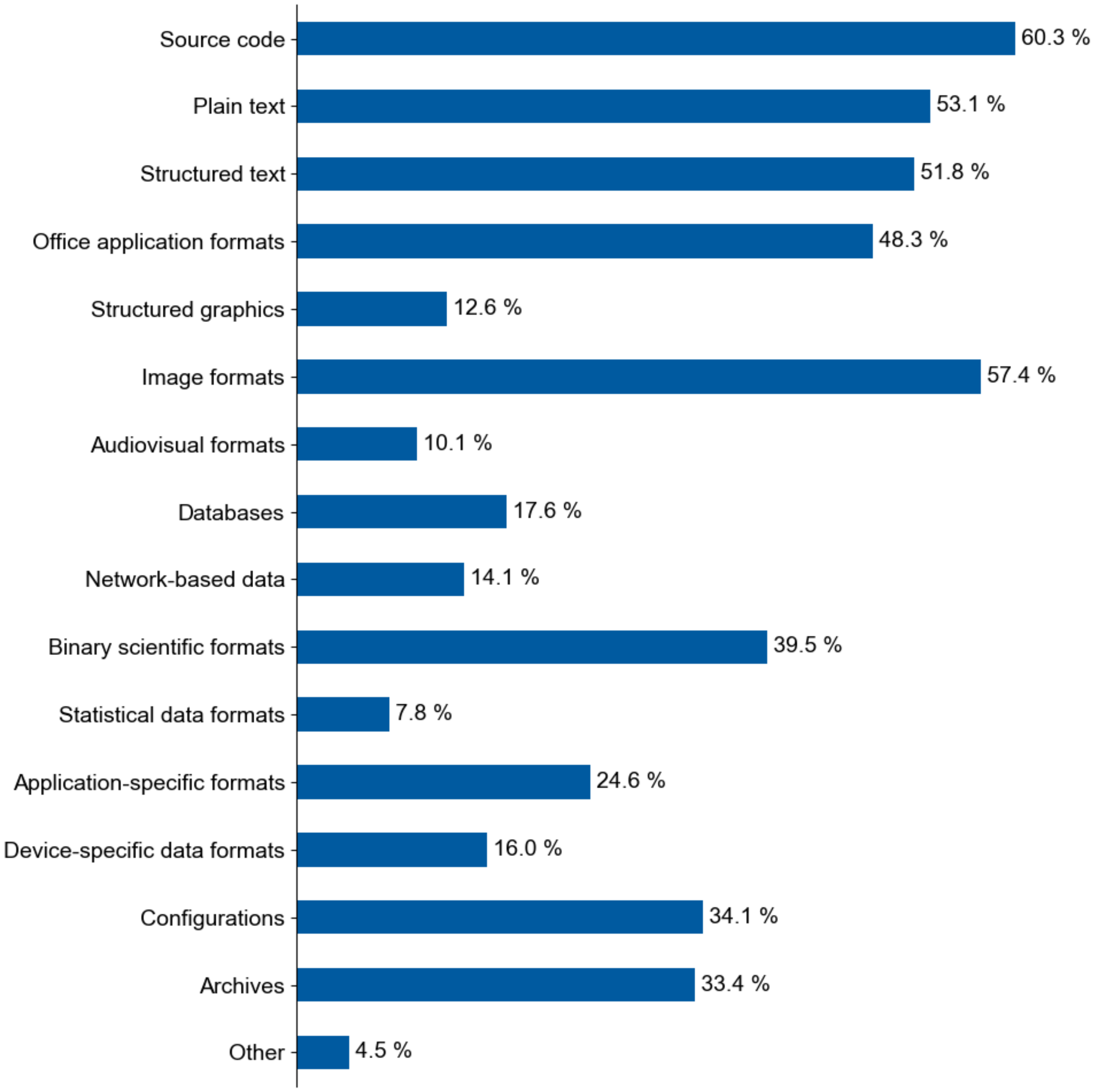


Figure 11: "Please select the data formats that you generate or use in your current research project.". (Multiple-choice question, available to all respondents, number of respondents who answered this question: n = 625, relative amounts refer to n)

Figure 11 provides information on the data formats used by the respondents. Over 50% of the respondents generate image formats, source code, plain text, and structured text files. Notably, more than one-third of the participants generate archival, configurational or binary data files during their research activities. Statistical data formats are the least generated, standing at 7.8% averaged over the six research fields.

Research data management practices

With HMC's mission to enable researchers to make their research data FAIR across the Helmholtz Association, it is crucial to understand and review the current research data management practices of Helmholtz' researchers. The question group "Research Data Management Practices", hence, focuses on research data storage routines, as well as data annotation and documentation.

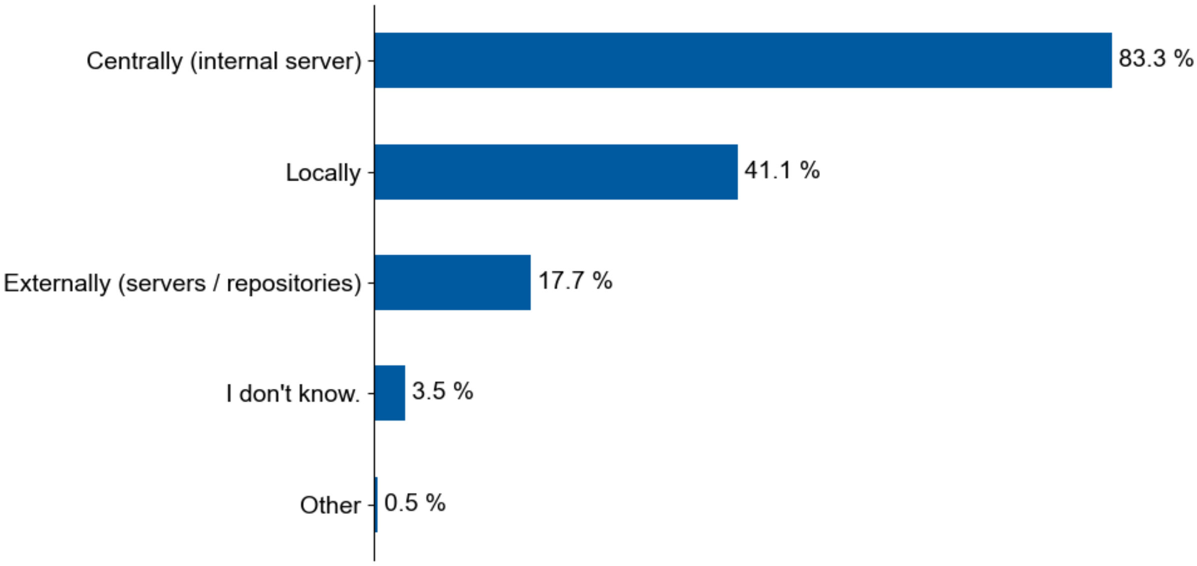


Figure 12: “Where is most of your research data stored after a project is finished?”. (Multiple respond question, available to all respondents, number of respondents who answered this question: n = 623, relative amounts refer to n)

Figure 12 illustrates the research data storage locations after a project is finished. Predominantly, respondents are using a central (internal) server (83.3%) or store their data locally (41.1%). Only 17.7% of respondents make their data publicly available by storing them on external servers or external repositories.

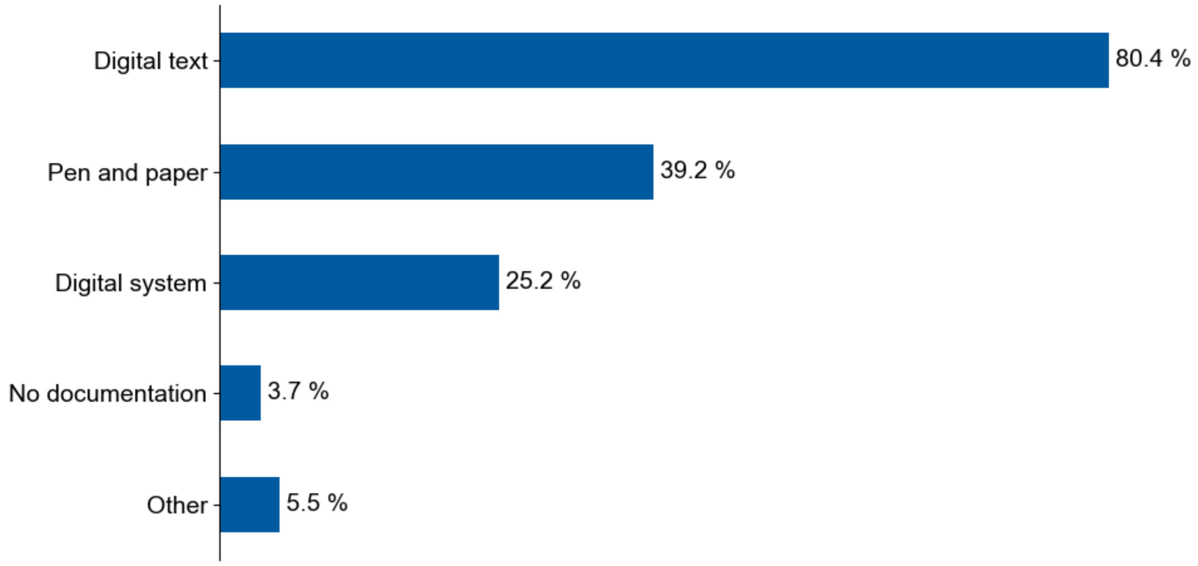


Figure 13: "In your current project, where do you document the steps used to generate and process your data?". (Multiple choice question, available to all respondents, number of respondents who answered this question: n = 622, relative amounts refer to n)

Figure 13 shows where respondents document their data generation and processing. Most respondents selected digital text (80.4%). Fewer selected pen and paper (39.2%), followed by digital system (25.2%). Some respondents reported doing no documentation (3.7%).

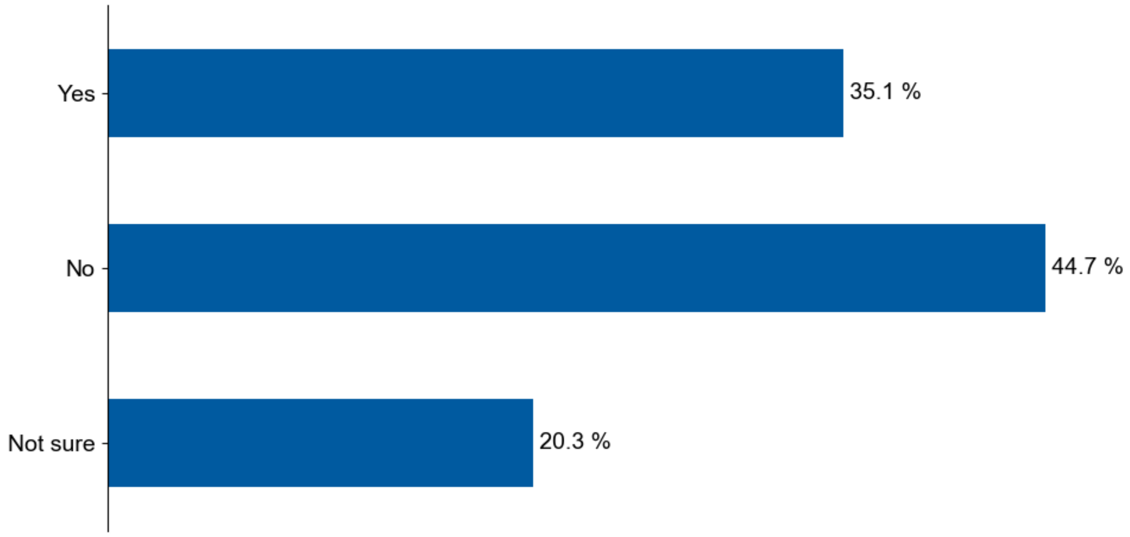


Figure 14: "Do you document your research data in a structured way? (e.g., using forms, templates or schemas)". (Single choice question, available to all respondents, number of respondents who answered this question: n = 582, relative amounts refer to n)

Figure 14 is about the structured documentation of research data. Most respondents are of the opinion to not structure their documentation (44.7%), followed by respondents who structure their documentation (35.1%). Interestingly, one-fifth of the respondents were not sure if their documentation is structured. This is likely due to an unfamiliarity with standardised metadata schemas which is also reflected in the high number of responses that were unfamiliar with the FAIR principles in general (see Figure 8).

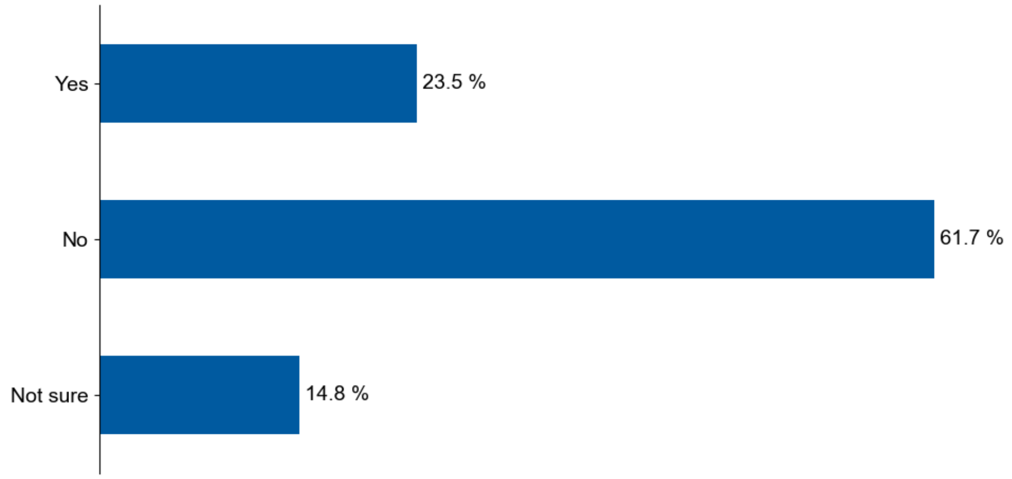


Figure 16: “Do you use internationally used templates, schemas or standards for [the structured documentation of your research data]?”. Question was only shown to those respondents, who had previously indicated that they document their research data in a structured way. (Single choice question, number of respondents who answered this question: n = 196, relative amounts refer to n)

A quarter of the respondents who document their research data in a structured way, adopt international standards, while 61.7% do not. 14.8% of the respondents are not sure if they use international standards (Fig. 15).

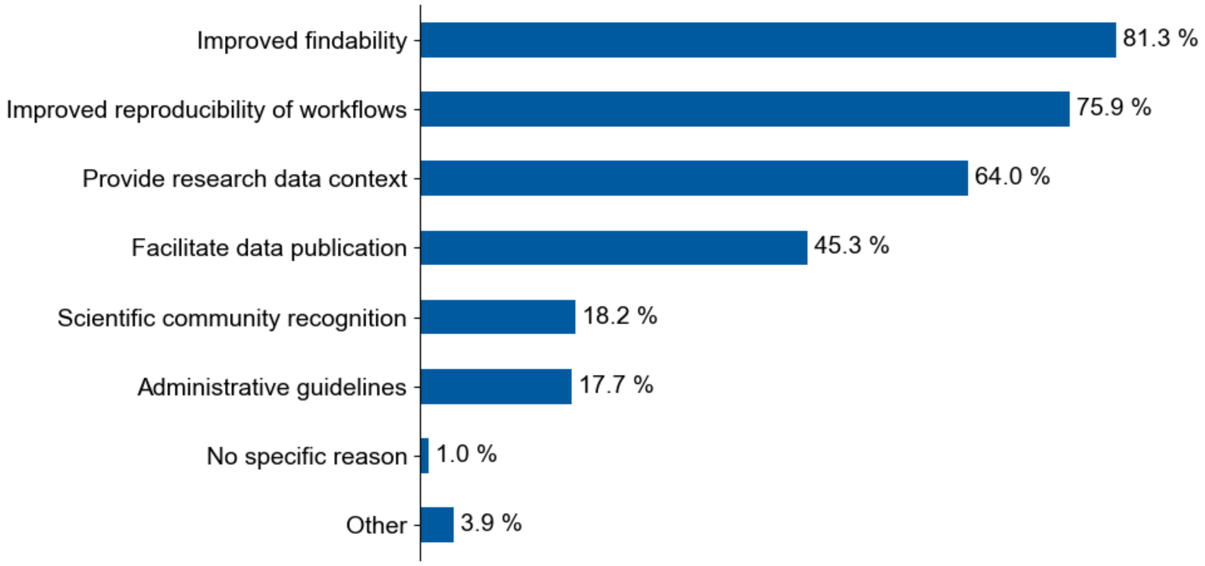


Figure 15: “Which of these reasons motivate you to document your work in a structured way?”. Question was shown only to those respondents, who had previously indicated that they document their research data in a structured way. (Multiple choice question, number of respondents who answered this question: n = 203, relative amounts refer to n)

Almost all respondents who document their research data in a structured way, have specific reasons for doing so since only 1% indicated the contrary (Fig. 16). The most frequently selected reasons were intrinsic motivations. Extrinsic motivations were less frequently selected. The reason with the highest frequency was "improved findability" (81.3%), followed by "improved reproducibility" (75.9%) and "providing context" (64%). Close to half of the respondents also felt that structured documentation of research data facilitates data publication (45.3%), while recognition by other scientists and administrative guidelines were each only motivating for around 18% of the respondents.

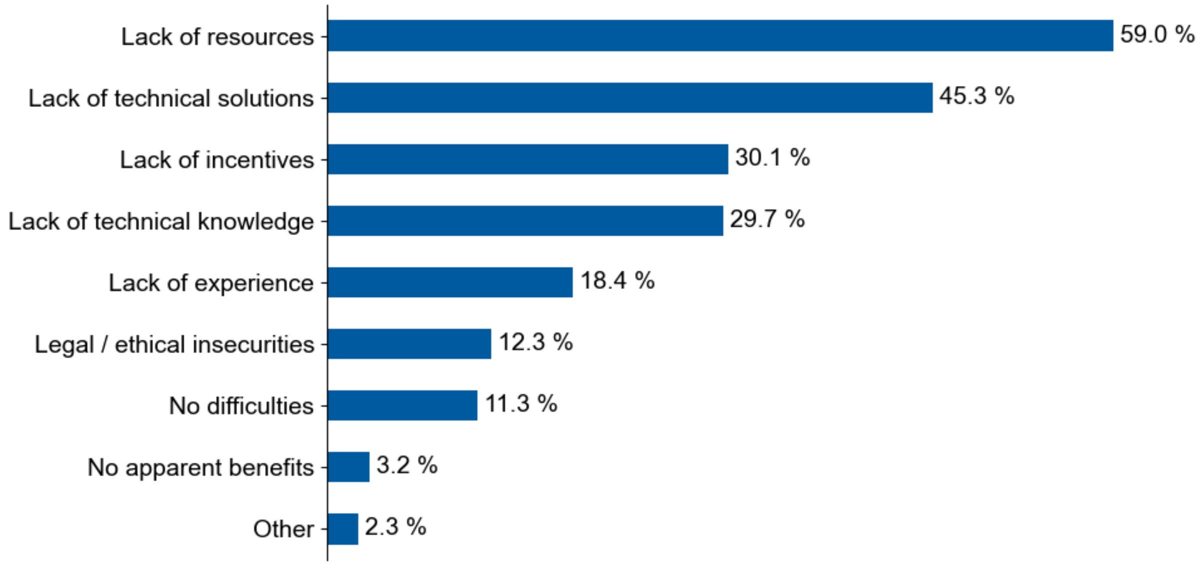


Figure 17: "What obstacles or difficulties have you encountered in collecting metadata as part of your work?". (Multiple choice question, available to all respondents, number of respondents who answered this question: n = 602, relative amounts refer to n)

While 11.3% of the respondents had no difficulties in collecting metadata, most of them were facing a lack of resources when doing so (i.e. a "lack of resources" (59%), "lack of technical solutions" (45.3%), and a "lack of technical knowledge" (29.7%)) (Fig. 17). 18.4% of the respondents also felt they were lacking experience and 12.3% indicated legal and/or ethical insecurities. Furthermore, 30.1% of the respondents perceived a lack of incentives, while 3.2% did not see any benefits in collecting metadata.

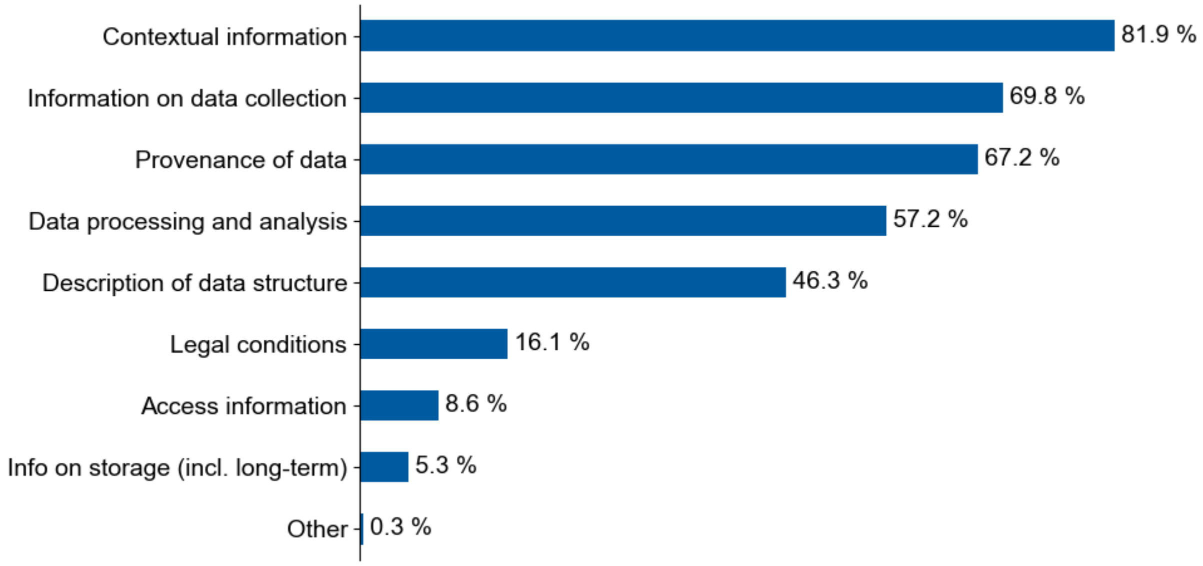


Figure 18: "Please select which information (metadata) you typically use to describe your research data.". (Multiple choice question, available to all respondents, number of respondents who answered this question: n = 603, relative amounts refer to n)

When asked for the recorded metadata types, a majority of 81.9% of the respondents had chosen the broader category of contextual information (e.g., topic, object of investigation), while close to 70% specifically document metadata on data collection (69.8%) and/or data provenance (67.2%) (Fig. 18). About half of the respondents record metadata on data processing (57.2%) and/or their data structures (46.3%). Legal conditions (16.1%), access information (8.6%) and storage information (5.3%) are documented less often.

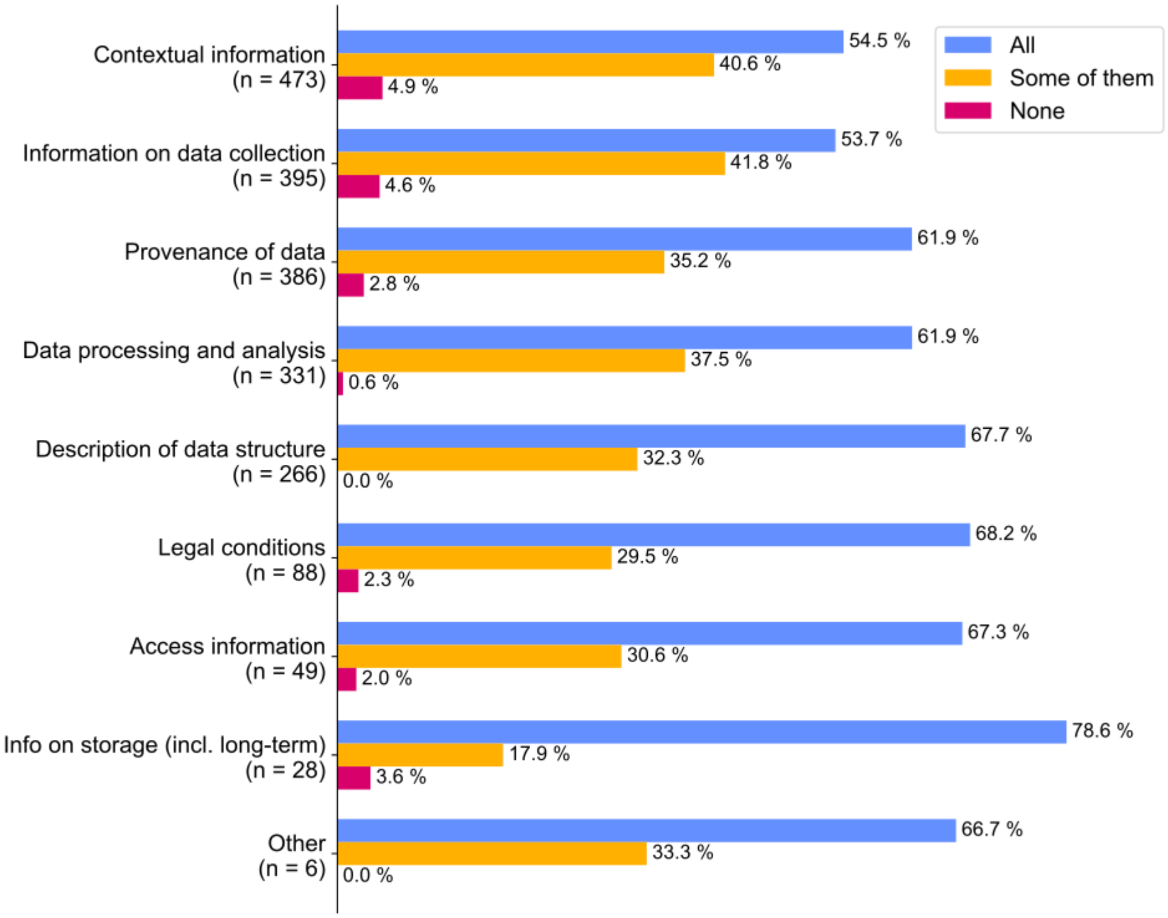


Figure 19: “Which information (“metadata”) do you typically document in a digital way?”. Respondents could rate only those metadata categories for which they had previously indicated that they typically use them to describe their research data. Relative amounts refer to the total number of answers collected for the respective metadata category. (Multiple choice question, number of respondents who answered this question: n = 573)

Scientific metadata is rarely recorded exclusively with analogue media (Fig. 19). However, for every category, the number of respondents who record some of their metadata in an analogue way is between a half and three quarters of the number of those respondents who do so fully digitally.

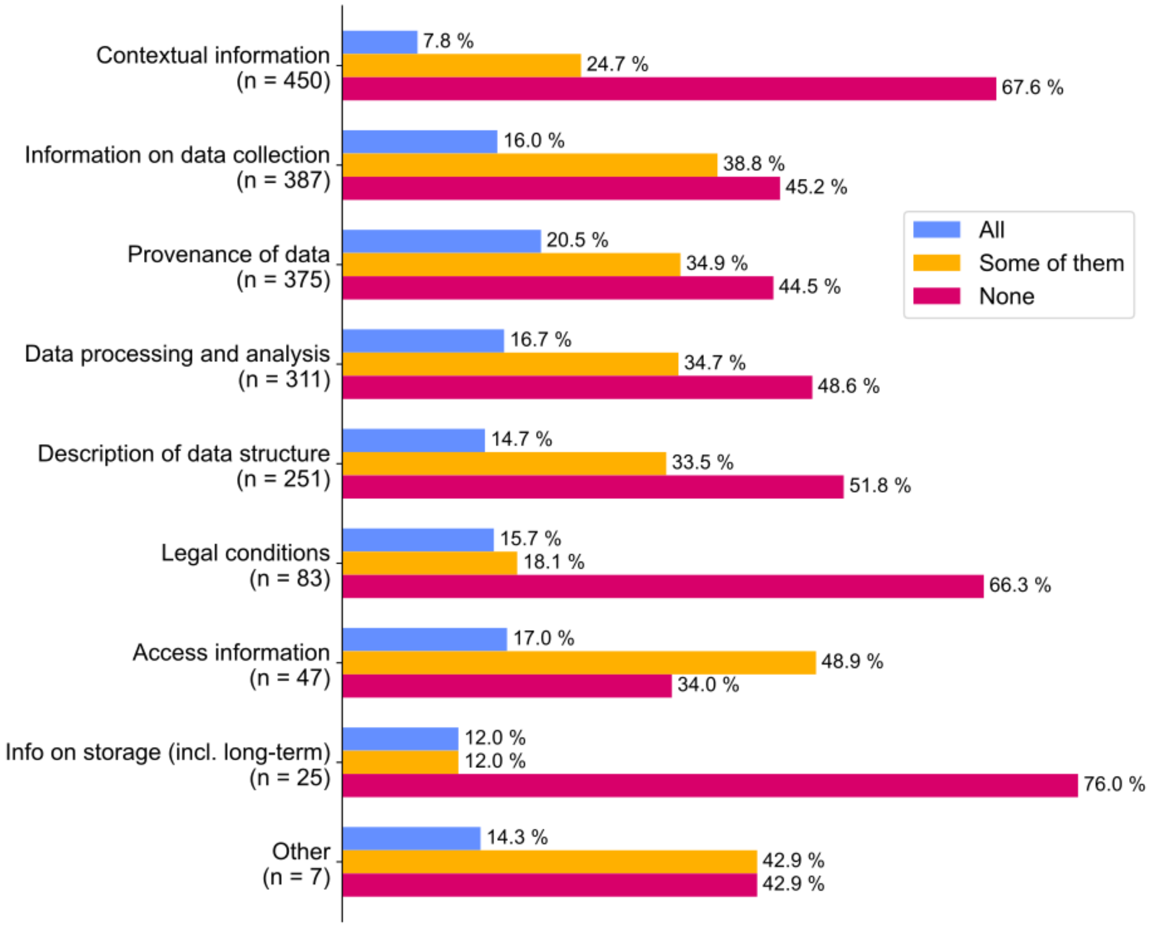


Figure 20: "Which of those information ("metadata") do you typically gather in an automated way?". Respondents could rate only those metadata categories for which they had previously indicated that they typically use them to describe their research data. Relative amounts refer to the total number of answers collected for the respective metadata category. (Multiple choice question, number of respondents who answered this question: n = 349)

In nearly every category, at least half of the respondents gather their metadata either partially or completely manually (Fig. 20). The main exception is "access information" (option text shown to respondents: "Access information about data or objects (e.g., API descriptions, authentication, log files, statistics)") which is collected by about two thirds using automated processes. Other categories with comparatively low shares in fully manual approaches are "description of data structure" (51.8%), "data processing and analysis" (48.6%), "information on data collection" (45.2%), and "provenance of data" (44.5%). Legal, contextual, and storage policy metadata on the other hand are each collected fully manually by more than two-thirds of the respondents.

Data publishing practices

Research data annotation with metadata is crucial for data publications. Many data repositories require certain metadata records for the data sets they accept. The data publishing processes within the Helmholtz Association were investigated with the following set of questions in figures 21 - 25.

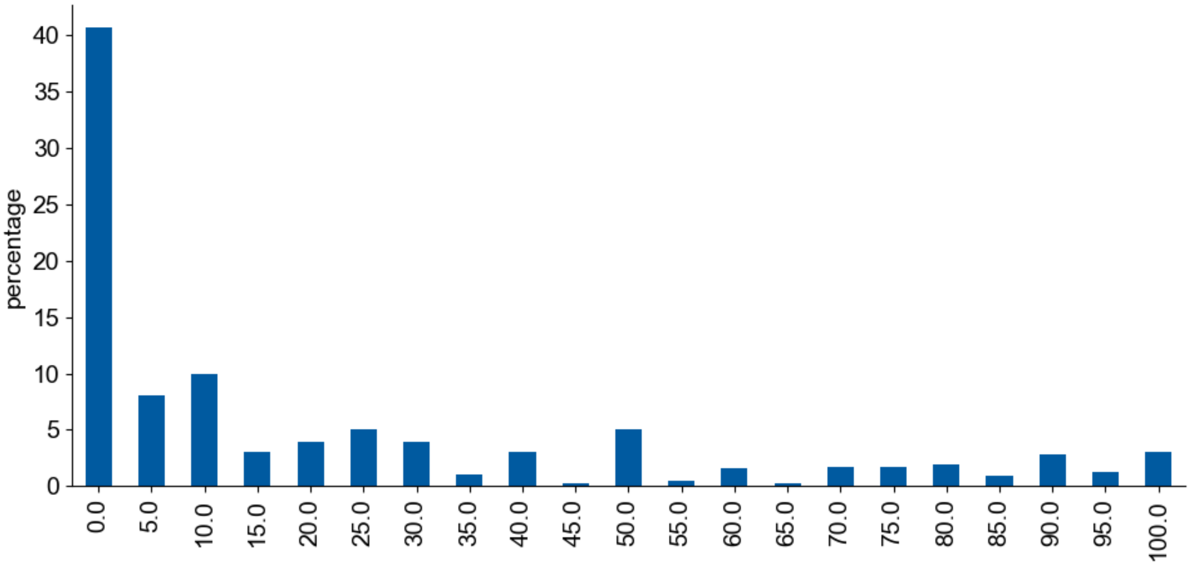


Figure 21: "Please estimate the relative amount of your data sets that you make publicly available." (Single choice question with options being available in discrete steps of 5%, available to all respondents, number of respondents who answered this question: n = 631, relative amounts refer to n)

40.6% of all respondents don't make any of their data sets publicly available and 18.1% only publish 5 to 10% (Fig. 21). The amounts of published data indicated by the remaining 41.3% of respondents are distributed quite uniformly between 15% and 100%. Out of all respondents, 21% publish at least 50% of their data.

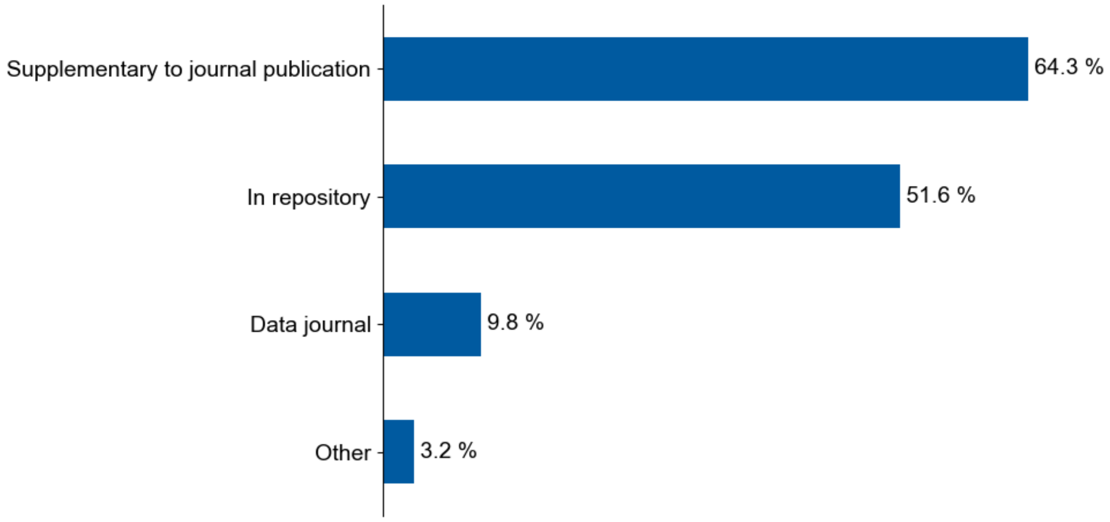


Figure 22: "How did you publish your data?". This question was shown only to those respondents who had previously indicated that they make a non-zero amount of their data publicly available. (Multiple choice question, number of respondents who answered this question: n = 386, relative amounts refer to n)

More than half of the respondents who have experience in data publication, tend to make their data available as a data supplement to their journal publication (64.3%) and/or in a repository (51.6%) (Fig. 22). Data publication in specialized data journals is less common (9.8%).

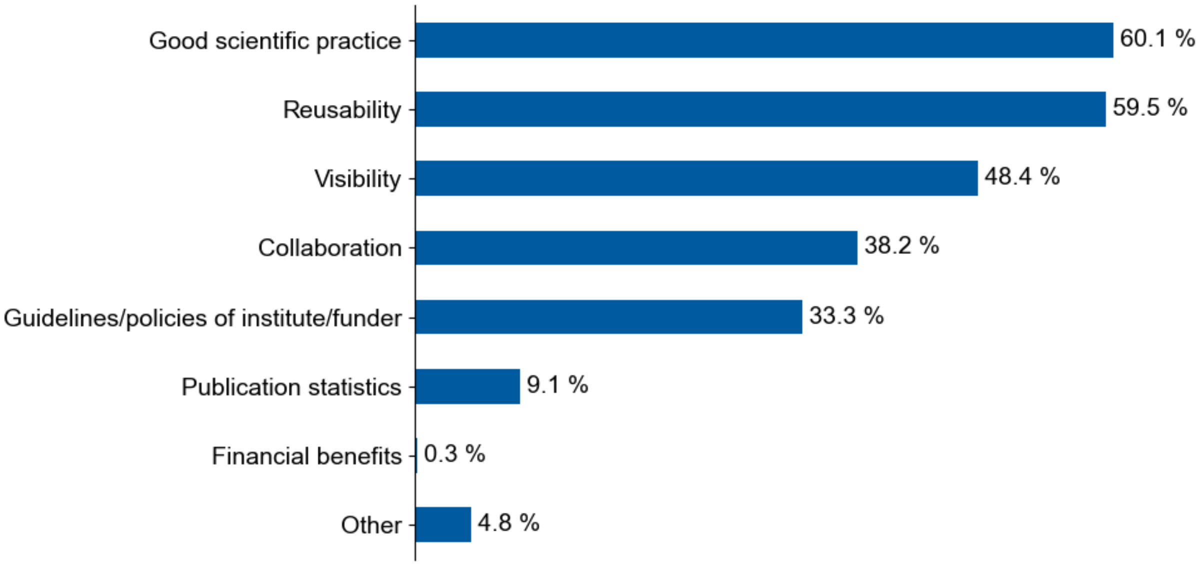


Figure 23: "Which of the following motivated you to publish your data? (Please choose up to 3 options)". This question was shown only to those respondents who had previously indicated that they make a non-zero amount of their data publicly available. (Multiple choice question, number of respondents who answered this question: n = 386, relative amounts refer to n)

Notably, most respondents indicated that their data publication practices follow intrinsic motivation (good scientific practice (60.1%), reusability (59.5%), and visibility (48.4%)) (Fig. 23). For about one-third of the respondents, collaboration or funding/institute policies are important motivators to publish their data. Improvement of the publication statistics (9.1%) and financial reasons (0.3%) were far less frequently selected as data publication motivations.

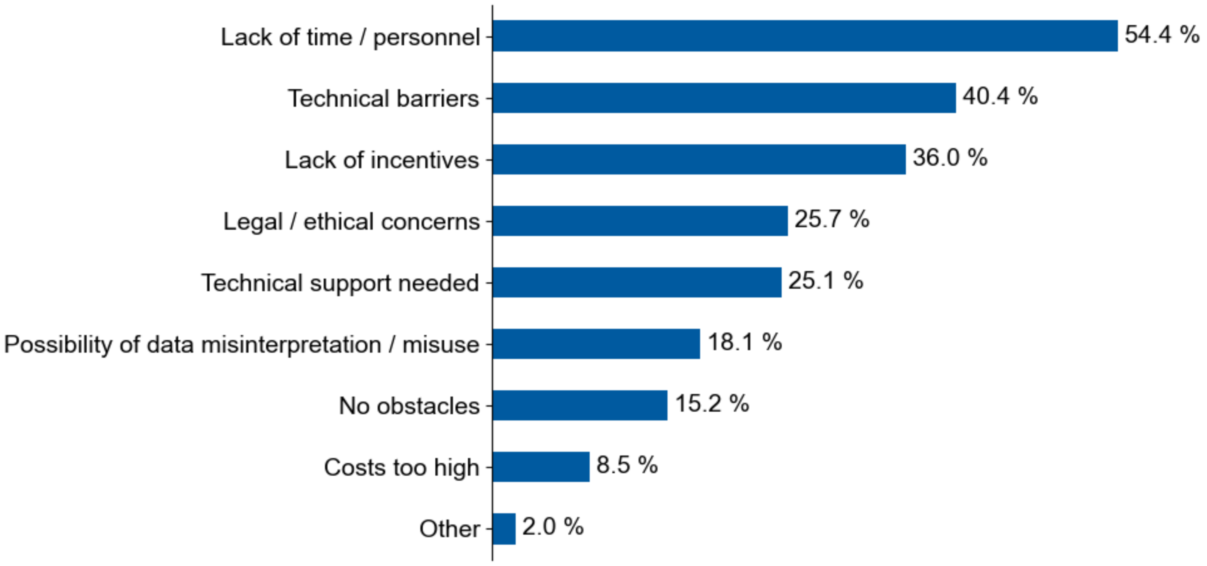


Figure 24"What obstacles have you encountered in publishing your research data?". This question was shown only to those respondents who had previously indicated that they make a non-zero amount of their data publicly available. (Multiple choice question, number of respondents who answered this question: n = 377, relative amounts refer to n)

Despite their experience in and motivation for data publication, more than half of the respondents named a lack of time and/or personnel as an obstacle followed by technical barriers (40.4%) (Fig. 24). A quarter of the respondents indicated a need for technical support and 8.5% said that high costs were also a problem for them. Furthermore, 25.7% have legal and/or ethical concerns to publish their research data. 15.2% said that they have not encountered any obstacles when publishing their data.

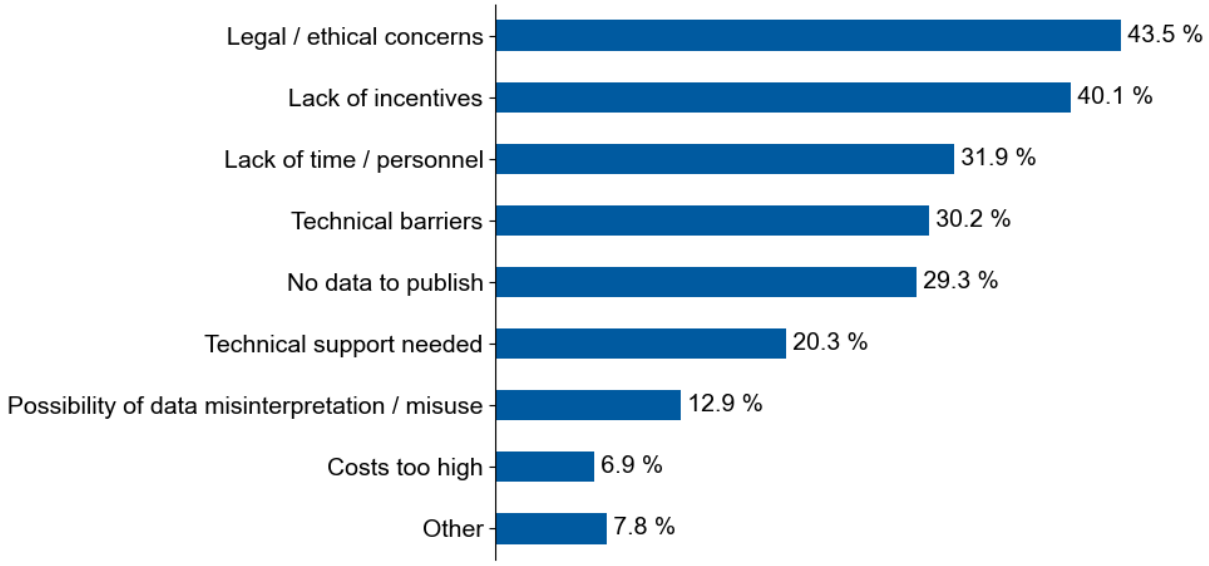


Figure 25: “What concerns or obstacles have discouraged you from publishing your research data so far?”. This question was shown only to those respondents who had previously indicated that they do not make any data publicly available. (Multiple choice question, number of respondents who answered this question: n = 208, relative amounts refer to n)

In parallel, respondents who did not publish any data yet, were asked to indicate the barriers that kept them from doing that. A third of these respondents does not have any data which they could publish and about 40% of them don't see sufficient reasons as to why they should do it in the first place (Fig. 25). More than that however, legal and ethical concerns keep 43.5% of the respondents from publishing their data. 31.9% face a lack of personnel or time and/or face technical barriers and 20.3% are missing technical support. 6.9% also lack financial resources and 12.9% do not publish any data for fear of misinterpretation or misuse.

The 12 most frequently used metadata categories (Fig. 26), between about 25% and 50% were mostly related to how data was recorded ("research method" (49.9%), "data collection software" (33.6%), "data collection workflows" (22.8%)). The exceptions from this trend were "author/producer of the data" (43.6%), "research subject" (42.2%), "analysis software/scripts" (31.3 %), and "name of data set" (30.2%). 11.1% of respondents indicated that they use all of the asked for metadata categories, while 6.8% don't use any of them. The least selected option was the data retention period (0.3%)

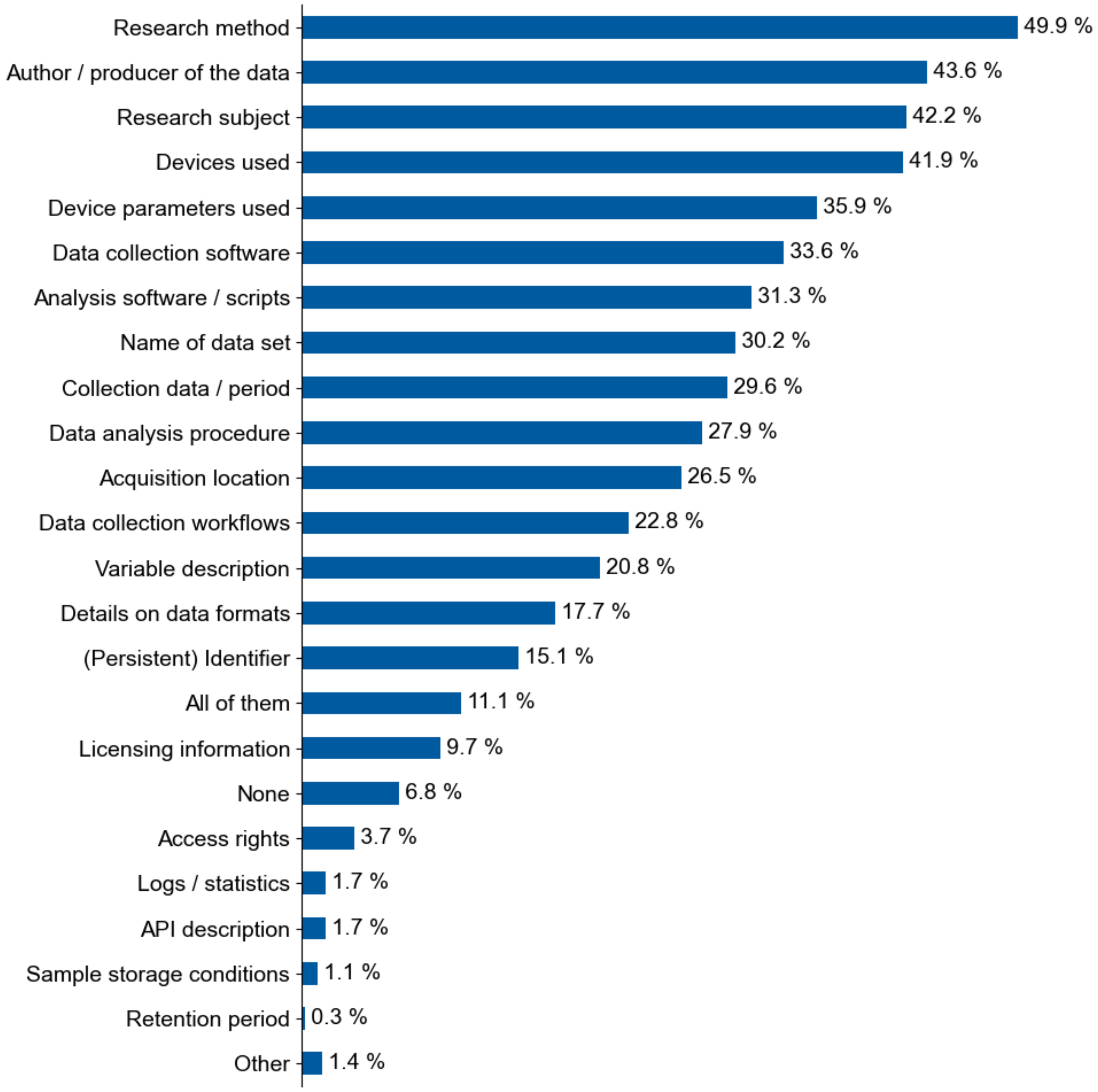


Figure 26: “Which of these metadata do you publish along with your research data?”. This question was shown only to those respondents who had previously indicated that they make a non-zero amount of their data publicly available and document at least one metadata category. (Multiple choice question, number of respondents who answered this question: n = 386, relative amounts refer to n)

Gaps and needs

To specifically understand the needs of respondents for HMC services we asked about their support requirements for metadata and research data management as well as service formats before concluding the survey.

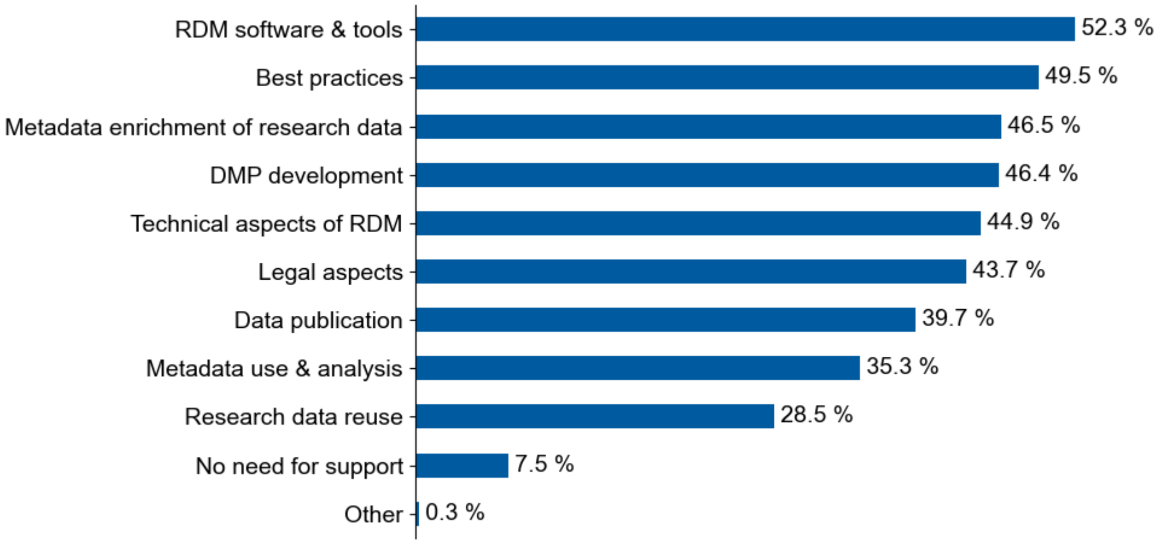


Figure 28: "In which areas of research data management do you perceive a need for supporting services?". (Multiple-choice question, available to all respondents, number of respondents who answered this question: n = 604, relative amounts refer to n)

When the respondents were asked within which areas they require support, most of the available options were selected by nearly half of respondents (Fig. 27). Options which relate to

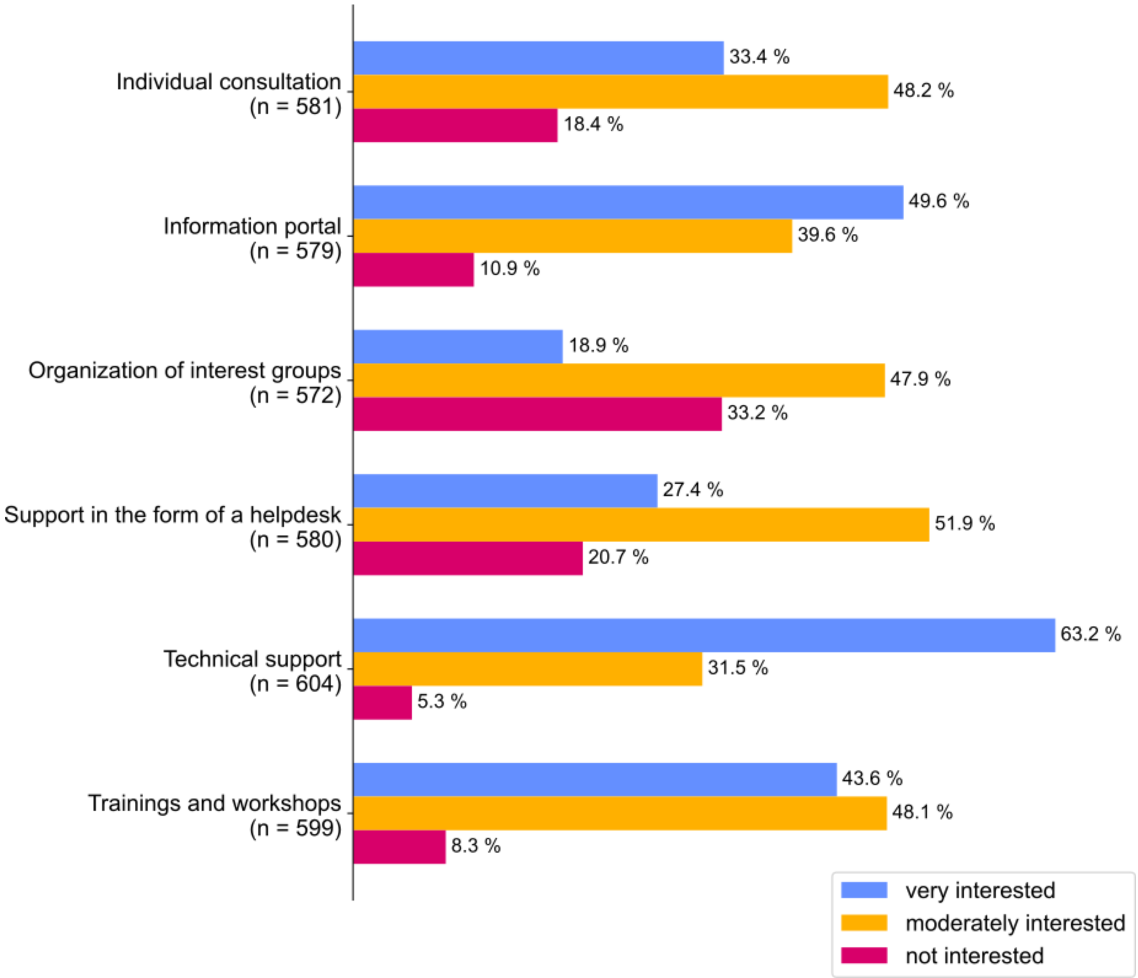


Figure 27: "Please rate your interest in the following service formats.". (One non-mandatory interest rating per service format, available to all respondents, number of respondents who answered this question: n = 612, relative amounts refer to n)

research data management before data publication were selected somewhat more often than the rest. There were also respondents who do not require any support, although comparatively few (7.5%).

Out of a selection of possible service formats, more than 65% of the respondents were at least moderately interested in every one of them (Fig. 28). Nonetheless, the format of "technical support" was ranked highest among them (94.7%). The least interesting format was the organization of interest groups (65.8%).

4 Summary

In this section, the results presented in the previous section are summarized, in line with the survey goals defined in the Survey Design chapter.

Status quo

Responses were obtained from researchers in 18 Helmholtz centres in Germany and representing the six Helmholtz research fields. It was observed that more than 65% of the participants' research data is mostly self-generated. Projects carried out by nearly 20% of the participants mostly involved the reuse of existing data. This highlights the potential of fostering data reuse practices in specific research communities. Various types of data are being used by the respondents, with the most widespread ones being source code, images, plain text, structured text (e.g., JSON, XML, CSV) and Office application data. Among all research fields, it was observed that the application of the FAIR principles is low. 54% of the participants were aware of the FAIR principles. However, those who then also apply them in their research work only amounted to 17% of all respondents.

It was observed that over 80% of the participants store some of their metadata in digital formats. About 40% of the respondents also use pen and paper to document parts of their metadata. 35% of the participants document data in a structured way, out of which 23% use an international standard. 65% of the participants do not store metadata in a structured way, and among those who do, more than three-quarters do not use international standards. The top motivations for documenting metadata were observed to be intrinsic - improved findability, improved reproducibility, and providing context to the data. Scientific community recognition and administrative guidelines were the weakest motivating factors, with less than 20% of the respondents choosing these options.

More than 60% of the participants are documenting metadata such as contextual information, information on data collection, and provenance. Legal conditions, access information, and information on long term storage are not extensively documented. However, more than 67% of those who document metadata in these categories reported storing complete information on these categories. This is higher than the completeness of information documented for the other categories of metadata, except for the description of data structure.

After their projects are concluded, more than 83% of the respondents store their research data in an internal server, thus making the data inaccessible to a large group of researchers. External servers and repositories are being used by only about 17% of the respondents. Nearly 41% of the respondents never published any data, with 21% of the respondents publishing at least 50% of their data. A quarter of the respondents publish less than one-quarter of their data. Most of these datasets were published either as a supplement to journal publications or in a repository. Data journals were found to be a less popular choice, with only about 10% of the respondents choosing this option. While publishing their data sets, more than 40% of the respondents published metadata such as research method, author of data, devices used, and research subject. Details such as access rights, retention period, API information, and sample storage conditions, that are observed to be less documented, were also the least included categories while publishing datasets.

Community demands

Only 3% of the respondents feel that there is no benefit in documenting metadata, hinting that a great proportion of researchers do find metadata to be valuable; but face other difficulties in documenting them. The main challenge faced by the respondents in collecting metadata is a lack of resources, followed by other difficulties such as a lack of technical solutions, incentives, and technical knowledge. This corresponds to the areas in which the respondents express a need for support, with topics such as research data management tools, and metadata enrichment being selected by about half of the respondents. 35% to 55% of the respondents perceived a need for support in most of the topics that were provided as options in the survey, except for research data reuse with only 28%.

In the area of data publication, among the respondents who have never published a dataset, a lack of incentives, and legal and ethical concerns were found to be major obstacles which prevented the respondents from making data publicly accessible. Those who have published data, at least once, reported a lack of time and personnel and technical barriers as their major problems. Nevertheless, neither the possibility of misinterpretation or misuse, nor high publication costs were considered as barriers, regardless of the data publication practices.

Major motivations for data publication are found to be reusability and good scientific practice, chosen by close to 60% of the participants. Other factors such as visibility, collaboration, and guidelines of the institute or funder stand between 33% and 49%. The least motivating factors are publication statistics and financial benefits. These motivations also correspond with the areas in which the researchers have expressed an interest and a need for support, which again indicates that they are indeed motivated to engage in FAIR data.

Community feedback on potential services

The community has expressed a need for support in several topics such as RDM software and tools, best practices, metadata enrichment of research data, DMP development, technical aspects of RDM, legal aspects, data publication, and metadata use and analysis. Communication can be established in one of the formats preferred by the community, such as by providing technical support, organizing training and workshops, and establishing an information portal. Addressing these topics would improve awareness and knowledge of the FAIR principles, in addition to motivating researchers to store and publish datasets enriched with metadata.

The format that most of the researchers would be interested in is technical support, with just over 63% of the respondents being very interested and about 95% being at least moderately interested. This is followed by trainings and workshops as well as an information portal. Individual consultation and support in the form of a helpdesk are less popular options, with about 20% not being interested and only about 30% being very interested in these supportive offers. The least interesting format is the organization of interest groups, with close to one-third of the respondents not being interested in it.

Credits and acknowledgements

Author contributions

In the following paragraph, author contributions are summarized following the CRediT taxonomy[8].

W.A., S.C.G., V.H., M.K., L.K., C.L., O.M., K.R., J.S., S.S., E.S. and W.S. conceptualized the study.

S.C.G., M.K., L.K., C.L., J.S., S.S. and L.S. curated the data.

S.C.G. and S.S. formally analysed the data.

S.C.G., M.K., L.K., C.L. J.S., and S.S. conducted the investigation.

S.C.G., V.H., M.K., L.K., C.L., K.R., J.S., S.S. and E.S. developed the methodology.

M.K. and C.L. administrated the project.

W.A., M.K., C.L., and E.S. provided study materials (resources).

S.C.G., M.K., J.S. and S.S. worked on the software.

W.A., V.H., O.M., M.N., E.S. and W.S. supervised the project.

J.S. and S.S. validated the data.

S.C.G. and S.S. visualized the data.

S.C.G., M.K., L.K., C.L., S.S. and L.S. wrote the original draft of this report.

W.A., S.C.G., V.H., M.K., L.K., C.L., O.M., M.N., S.S. and L.S. reviewed and edited the report.

Acknowledgements

We thank various members from the Helmholtz Federated IT Services (HIFIS) for consultation and technical support. In particular, we would like to thank Katharina Dworatzyk and Meike Schaller for feedback on the questionnaire and Christian Hüser and Fredo Erxleben for supporting us in adopting the HIFIS Surveyval framework.

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

References

Oleksandra Arndt, Laura Glatz, Benedikt Hummel, Magdalena Porst, Wassili Schabalowski, and Sophia Skubatz. Umfrage zum Forschungsdatenmanagement an der FH Potsdam: Projektbericht. Technical report, September 2018. DOI:10.5281/zenodo.1161792.

Bruno Bauer, Andreas Ferus, Juan Gorraiz, Veronika Gründhammer, Christian Gumpenberger, Nikolaus Maly, Johannes Michael Mühlegger, José Luis Preza, Barbara Sánchez Solís, Nora Schmidt, and Christian Steineder. Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung – Report 2015. Technical report, October 2015. DOI:10.5281/zenodo.31935.

Stephane Berghmans, Helena Cousijn, Gemma Deakin, Ingeborg Meijer, Adrian Mulligan, Andrew Plume, Sarah de Rijcke, Alex Rushforth, Clifford Tatum, Thed van Leeuwen, and Ludo Waltman. Open Data: the researcher perspective - survey and case studies. Technical report, 2017. DOI:10.17632/bwrnfb4bvh.1.

Bela Brenger, Stephanie Rehwald, Konstantin L. Wilms, Ania López, and Stefan Stieglitz. UNEKE: Forschungsdatenspeicherung - Praxis und Bedarfe: Online-Survey 2019. Technical report, Aug 2019. DOI:10.17185/duepublico/70259.

Briony Fane, Paul Ayris, Mark Hahnel, Iain Hrynaszkiewicz, Grace Baynes, and Emily Farrell. The State of Open Data Report 2019, 2019. DOI:10.6084/m9.figshare.9980783.v2.

Inken Feldsien-Sudhaus and Beate Rajski. Digitale Forschungsdaten für die Zukunft sichern: Umfrage zum Umgang mit Forschungsdaten an der TU Hamburg: Auswertung. Technical report, 2016. DOI:10.15480/882.1326.

Silke Christine Gerlich, Volker Hofmann, Markus Kubin, Lucas Kulla, Christine Lemster, Oonagh Mannix, Katharina Rink, Jan Schweikert, Sangeetha Shankar, Emanuel Söding, Leon Steinmeier, and Wolfgang Süß. HMC Community Survey 2021 – A survey on research data management practices among researchers in the Helmholtz Association (Dataset), 2022. DOI:10.7802/2433.

Reingis Hauck, Reiko Kaps, Hans Georg Krojanski, Anneke Meyer, Janna Neumann, and Volker Soßna. Der Umgang mit Forschungsdaten an der Leibniz Universität Hannover. Auswertung einer Umfrage und ergänzender Interviews 2015/16. Technical report, 2016. DOI:10.15488/265.

Maurice Heinrich, Anne Sieverling, Felix Schäfer, and Sabine Jahn. Digitale Forschungsdaten in den Altertumswissenschaften. Stakeholderanalyse 2013 zu Forschungsdaten in den Altertumswissenschaften. Teil 2: Kombinierte Auswertung & Interpretation. Technical report, 2015. DOI:10.13149/000.jah37w-q.

Sonja Herres-Pawlis, Johannes C Liermann, and Oliver Koepler. Research Data in Chemistry – Results of the first NFDI4Chem Community Survey. Zeitschrift für anorganische und allgemeine Chemie, 646(21):1748–1757, 2020. DOI:10.1002/zaac.202000339.

Bettina Hesse, Markus Baaske, Roman Gerlach, and Birgitta König-Ries. Forschungsdatenmanagement an der Universität Jena: Interviews zum Stand und Bedarf bei Verbundprojekten. Technical report, 2017. DOI:10.22032/dbt.33434

Esther Krähwinkel. Forschungsdatenmanagement an der Philipps-Universität Marburg. Die Ergebnisse der Umfrage zum Forschungsdatenmanagement im November 2014. Technical report, 2015. DOI:10.17192/es2015.0019

Marina Lemaire, Yvonne Rommelfanger, Jan Ludwig, Alexander Lürken-Uhl, Benjamin Merkler, and Peter Sturm. Umgang mit Forschungsdaten und deren Archivierung. Bericht zur Online-Bedarfserhebung an der Universität Trier. Technical report, 2016. URN:urn:nbn:de:hbz:385-10156.

Thilo Paul-Stüve, Georg Rasch, and Sören Lorenz. Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Christian-Albrechts-Universität zu Kiel (2014). Technical report, 2015. DOI:10.5281/zenodo.32582.

Boris Radosavljevic, Kirsten Elger, Roland Bertelmann, Christian Haberland, Susanne Hemmleb, Gerard Muñoz, Javier Quinteros, and Angelo Strollo. Report on the Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences. Technical report, 2019. DOI:10.5880/GFZ.LIS.2019.001.

Elena Simukovic, Maxi Kindling, and Peter Schirmbacher. Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. Technical report, 2013. DOI:10.18452/13568.

Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8):1–24, 08 2015. DOI:10.1371/journal.pone.0134826

Jon Treadway, Mark Hahnel, Sabina Leonelli, Dan Penny, David Groenewegen, Nobuko Miyairi, and et al. The State of Open Data Report. Technical report, 2016. DOI:10.6084/m9.figshare.4036398.v1.

Franziska Weng, Stella Thoben, Thilo Paul-Stüve, Stefan Farrenkopf, Holger Marten, Kerstin Helmkamp, and Katja Barth. Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten in Schleswig-Holstein (2018). Technical report, 2018. DOI:10.5281/zenodo.1216810

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1):1–9, 2016. doi:10.1038/sdata.2016.18.

Appendix

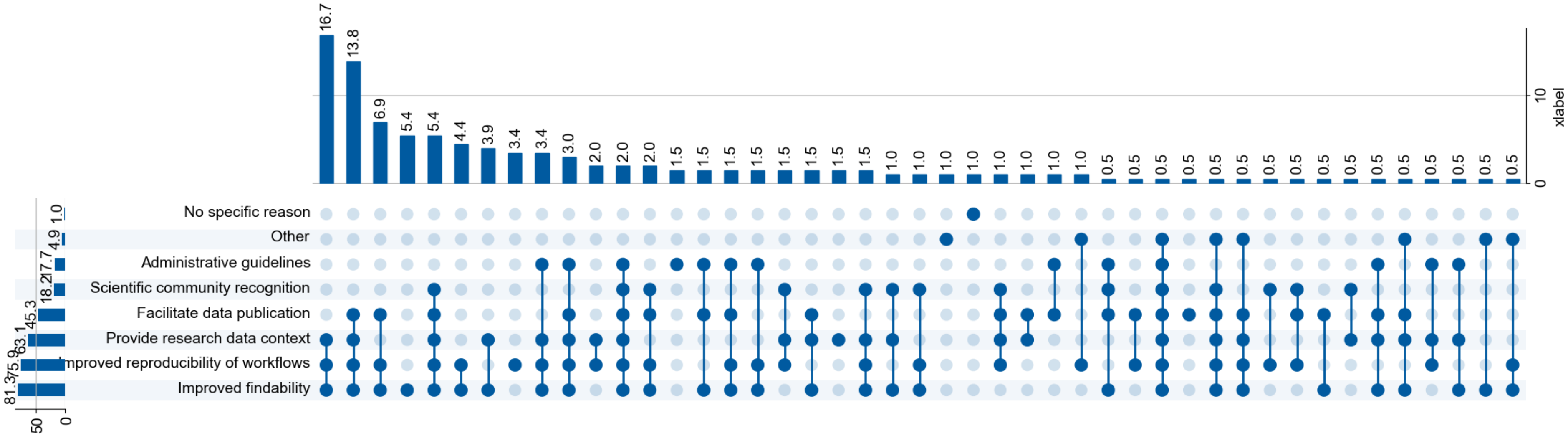


Figure 29: Answers to the question **“Which of these reasons motivate you to document your work in a structured way?”.** This question was shown only to those respondents, who had previously indicated that they document their research data in a structured way. Relative numbers refer to the total number of respondents who chose at least one answer option for this question. (Multiple choice question, n = 203)

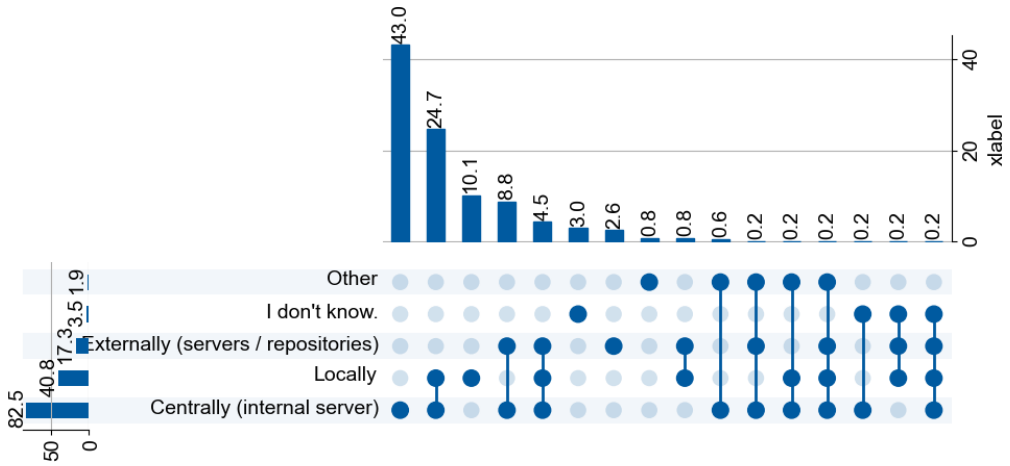


Figure 30: "Where is most of your research data stored after a project is finished?". (Multiple choice question, available to all respondents, number of respondents who answered this question: n = 623, relative amounts refer to n)

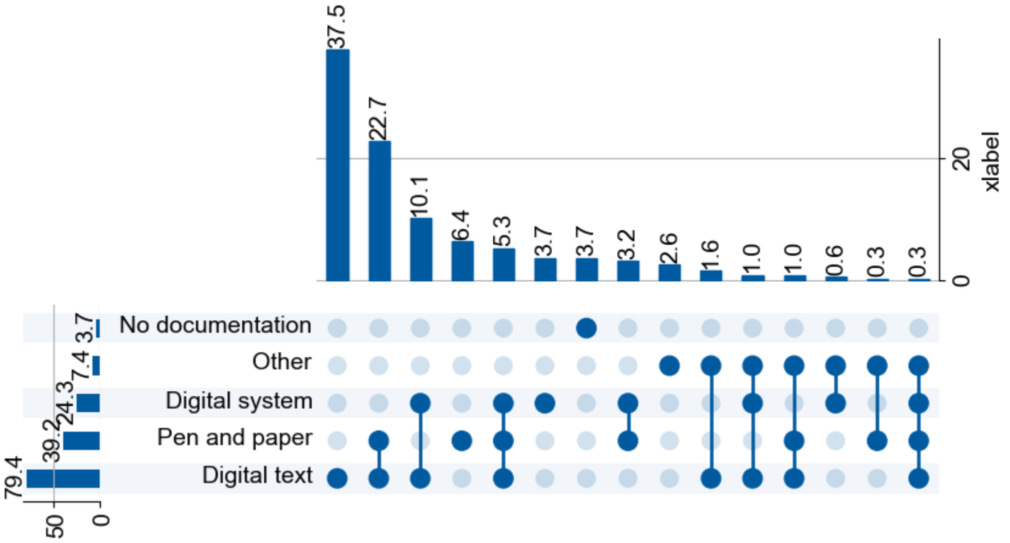


Figure 31: "In your current project, where do you document the steps used to generate and process your data?". (Multiple choice question, available to all respondents, number of respondents who answered this question: n = 622, relative amounts refer to n)