GFZ

Helmholtz Centre
POTSDAM

HELMHOLTZ CENTRE POTSDAM
GFZ GERMAN RESEARCH CENTRE
FOR GEOSCIENCES

Boris Radosavljevic, Kirsten Elger, Roland Bertelmann, Christian Haberland, Susanne Hemmleb, Gerard Muñoz, Javier Quinteros, Angelo Strollo

Report on the Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences

Scientific Technical Report STR19/02

www.gfz-potsdam.de

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

Please cite this report as:

Radosavljevic, Boris; Elger, Kirsten; Bertelmann, Roland; Haberland, Christian; Hemmleb, Susanne; Muñoz, Gerard; Quinteros, Javier; Strollo, Angelo (2019): Report on the Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences. (Scientific Technical Report STR; 19/02), Potsdam: GFZ German Research Centre for Geosciences.
DOI: http://doi.org/10.2312/GFZ.b103-19029

Data generated in the course of the survey and utilized herein are available at GFZ Dataservices:

Radosavljevic, Boris; Quinteros, Javier; Bertelmann, Roland; Hemmleb, Susanne; Elger, Kirsten; Haberland, Christian; Muñoz, Gerard; Strollo, Angelo (2019): Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences. GFZ Data Services.
DOI: http://doi.org/10.5880/GFZ.LIS.2019.001

Imprint

HELMHOLTZ CENTRE POTSDAM
GFZ GERMAN RESEARCH CENTRE
FOR GEOSCIENCES
Telegrafenberg
D-14473 Potsdam

Published in Potsdam, Germany
February 2019

ISSN 2190-7110

DOI: http://doi.org/10.2312/GFZ.b103-19029
URN: urn:nbn:de:kobv:b103-19029

This work is published in the GFZ series
Scientific Technical Report (STR)
and electronically available at GFZ website
www.gfz-potsdam.de

Report on the Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences

Boris Radosavljevic, https://orcid.org/0000-0001-6095-9078
Kirsten Elger, https://orcid.org/0000-0001-5140-8602
Roland Bertelmann, https://orcid.org/0000-0002-5588-0290
Christian Haberland, https://orcid.org/0000-0002-2981-7087
Susanne Hemmleb
Gerard Muñoz, https://orcid.org/0000-0003-0111-626X
Javier Quinteros, https://orcid.org/0000-0001-9993-4003
Angelo Strollo https://orcid.org/0000-0001-9602-6077

SPONSORED BY THE

Federal Ministry of Education and Research

The GeoDataNode Project is financed by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) under the grant 16FDM026: Ausbau eines fachspezifischen Knotenpunkts in einer Nationalen Dateninfrastruktur - geo_data_node_gfz (Build-up of a Disciplinary Node in a National Data Infrastructure).

Table of Contents

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

Executive Summary

The GeoDataNode project, funded by the Federal Ministry for Research and Education (BMBF), conducted a survey of data management practices at the Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences (GFZ). The aim was to assess the state of current practices and needs, and their alignment to institutional and national guidelines for data management. The target audience included scientific and technical employees at all levels. A response rate of 24% of the demographic target was achieved.

The survey revealed a general need for improvement and structuring of research data handling. This includes provision of adequate storage space, back-up schedules, and the familiarization with good scientific practice. The most important results are summarized below.

Data safety: 55% of the respondents regard back-up within their personal responsibility, while already 41% rely on dedicated staff. Although most back-ups are carried out daily or weekly, a fifth of respondents do not have a back-up schedule. Back-ups are made on external hard drives, the GFZ back-up service, or group and section servers. During active project phases, most respondents (76%) rely on local data storage on GFZ computers, central group server (60%) or external hard drives (42%).

Data acquisition: The broad spectrum of data acquisition (from sensor to modelling data) is reflected in the large variety of data types and formats that often require proprietary or own software.

Data documentation: Almost one quarter of respondents do not record any metadata. Respondents indicate that for improving the individual data documentation they would benefit from data documentation tools, clearly defined workflows, community agreed standards, and project related data management resources. The respondents also indicate a variety of needs in their data management process (e.g. archiving, metadata standards, general research data management, back-up, and licensing, among others).

Data preservation and storage: The majority of respondents generated more than 100 GB of data in the past five years. Only 40% are aware that research data underlying a scientific publication must be archived at the institution of origin for at least ten years. Over 50% of PhD students are not aware of this requirement. 32% of respondents indicated data from completed projects are archived in a research data repository.

Data sharing: During the active project phase, data are often shared within project and research groups only. Data sharing requirements by journals are largely unknown (48%). Even if most respondents prefer to publish data as an article supplement, already 42% follow the best practice of data sharing via a domain specific research data repository. The majority of respondents (67%) rely on recommendations from colleagues or use the institute's infrastructure, i.e. GFZ Data Services (58%). 65% of the respondents state that they would share their data publicly if they were given credit (e.g. via citation). 62% would require dedicated embargo periods or personal permission as a condition for data sharing. Interestingly, 20% of respondents are not aware of the possibility to obtain data via research data repositories.

Data management: workflows, practices and policies: A large percentage of GFZ researchers are not aware that existing data policies or guidelines apply to them. About 67% of all respondents) are not familiar with the *DFG Guidelines for Safeguarding Good Scientific Practice* (DFG, 2006), while 56% are not familiar with the *Guidelines on Research Data at the GFZ German Research Centre for Geosciences* (GFZ, 2016).

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

1. Introduction

In March 2016, as first Helmholtz Centre, the GFZ adopted the *Guidelines on Research Data at the GFZ German Research Centre for Geosciences* (GFZ 2016). This document recognizes quality-assured research data as a basic pillar of scientific knowledge. It further states that provision of research data is a service that not only benefits science but also society as a whole, and that GFZ accepts this additional effort as part of the scientific endeavor.

Open Data, i.e. the availability of data and metadata underlying scientific outcome, is an international request that was already addressed in the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (Berlin Declaration, 2003). Internationally, the largest impact and a testimony for open science was the *G8 Science Ministers Statement* from June 2013 (G8 Science Ministers, 2013). This statement was followed by several national and international initiatives like the *EU Implementation of the Open Data Charter* (European Commission, 2013), which requires the use of open formats, semantic interoperability, to ensure data quality and documentation, and a clear definition of intellectual property rights among others.

In 2016, the *FAIR Guiding Principles for scientific data management and stewardship* were published (Wilkinson et al., 2016). "*The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data*" (GoFAIR, Website). The FAIR Principles are now regarded as overarching guidelines for general research data management: data should be findable, accessible, interoperable and reusable.

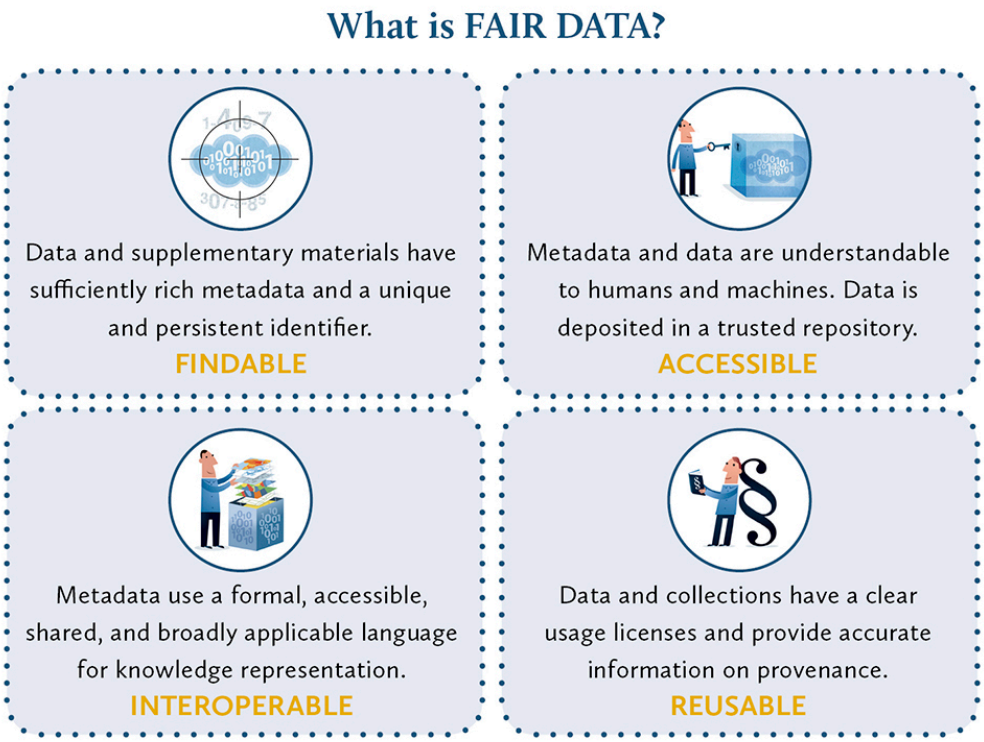


Figure 1 FAIR Principles at a glance (Association of European Research Libraries, Website)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

The call for FAIR data is supported by many funding agencies in Germany and abroad which recognize the importance of sustainable research data legacy through data curation (e.g. EU, DFG, BMBF, NSF, etc.) and require data management plans to be included in project proposals.

In recent years, several initiatives developed strategies, guidelines and practical tools for enabling FAIR data, e.g. the report *Turning FAIR Into Reality* by the European Commission Expert Group on FAIR Data (Collins et al., 2018) Especially in the Earth and Space Sciences, the push is supported by a broad coalition of research data repositories, publishers, funders, societies, scientific communities, institutions, research data infrastructure, and researchers publishers (COPDESS 2015; Enabling FAIR Data Community, 2018) through a formal commitment statement, to which GFZ is signatory. The statement calls for data underlying scientific articles are ideally deposited in domain repositories and cited within the article.

In 2017, the Helmholtz Association and GFZ adopted the *Rules for Safeguarding Good Scientific Practice* (Helmholtz Association 2017; GFZ, 2017), which lean on the *Proposals for Safeguarding Good Scientific practice* put forth by the German Research Foundation (DFG, 2006). One of the rules calls for archiving of primary data for least ten years at the institution of origin. "*Primary data include measurement results, collections, surveys, cell cultures, specimens of material, archaeological finds and questionnaires*" (DFG, 2006).

Although some communities at GFZ are already handling their data management according to the FAIR principles and the more general *Guidelines on Research Data at the GFZ German Research Centre for Geosciences* (GFZ, 2016), the full implementation of these guidelines remains a challenge in many research groups. The project *Build-up of a Disciplinary Node in a National Data Infrastructure* GeoDataNode GFZ (GDN), funded by the German Ministry of Education and Research (BMBF), aims at assessing and improving the current state of data management practices at GFZ. As a first step, GDN hosted the GFZ Data Forum for section and research group leaders to discuss the current practices on June 15, 2018. The next step represented the general online survey described here, targeting all personnel producing or working with research data. The rationale behind this survey was to identify common gaps and needs in general. The results of the survey serve as basis for the next steps of this project which envision the development of guidelines and tools with the different communities.

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

1.1 The data life cycle

The data life cycle (Figure 2) provides a general and overview of the stages involved in successful data management and preservation for use and reuse. The multiple versions of data life cycles reflect the variation in practices across domains or scientific communities. It describes the different stages research data undergoes from their generation to processing and analyses steps, to the preservation and, following the requests of open sciences described above, availability of the data for reuse. These publicly available data may be the origin of new investigations, e.g. as input data for models or during data mining, etc.

It is important to keep in mind that the responsibility for the data may change during the data life cycle. While data creation, processing and analyzing is performed by the scientist or research group, for example, data preservation and publication is often done by staff of research data repositories or libraries. Furthermore, the graphic may suggest that the stages in the cycle are clearly delimited. This is not always the case. For instance, giving access to data may refer to data access as part of a publication, or it can occur in the processing and analyzing stages as different partners in a project may work together. Storage considerations should be part of the planning effort to ensure capacities, including backups, during the processing and analysis stages, but also when the data is passed on to archives. Documenting data is penultimate for data to be reproducible or reusable and should occur at all stages of the data life cycle.

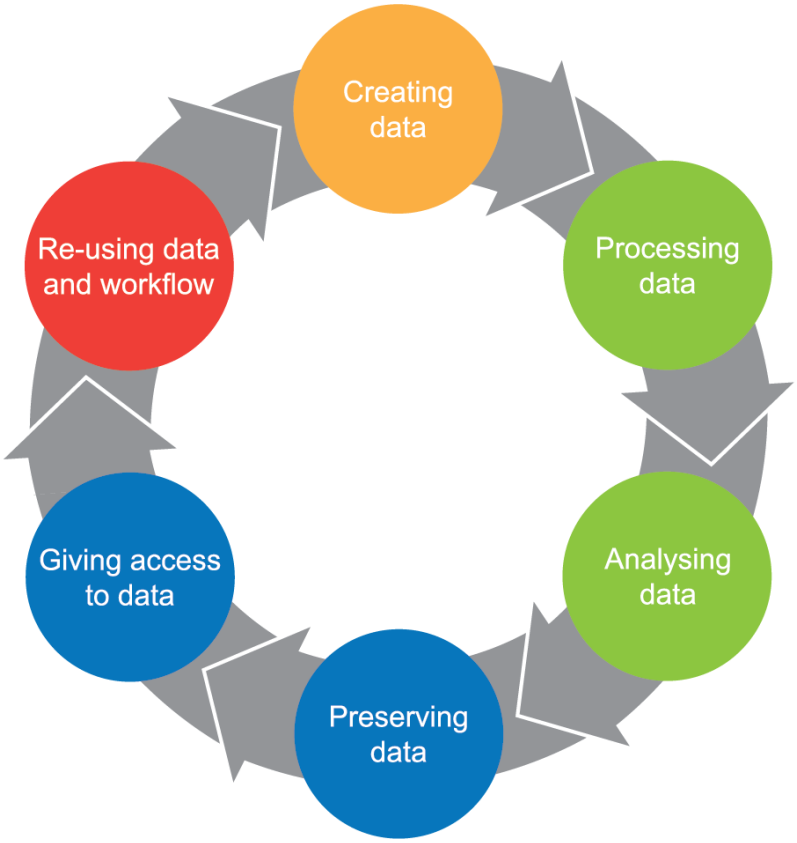


Figure 2 Data Management and the Data Life Cycle

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

2. Methodology

The survey design leaned on similar surveys carried out at German universities and research institutions (e.g. Paul-Stüve et al., 2015; Simukovic et al., 2013). It was addressed to all levels of scientific personnel. It queried aspects of the full data life cycle - from the planning stage to data reuse. Questions about the familiarity with the aforementioned guidelines, as well as usage and needs of data infrastructures were also included. An experimental run with representatives of the demographic target was carried out beforehand to further improve the survey.

The survey consisted of 37 questions. In particular, 16 single responses (SR), 20 multiple responses (MR), and one free text question about general comments. However, not all questions were necessarily presented to the interviewees. For instance, if respondents did not perform backups (Question 8) the following three questions were skipped. If no metadata was collected (Question 14), the following two questions were skipped.

The completely anonymous online survey was carried out with the Questback EFS Survey platform (Questback EFS Survey, Website). A short invitation to participate was emailed to the general GFZ mailing list with the following text:

> **Subject: 10 Minute Survey of Data Management Practices at GFZ - Please Respond**
>
> Dear colleague,
>
> We invite you to complete a short survey about research data management at GFZ. The GeoDataNode project, a project funded by the German Ministry of Education and Research (BMBF), seeks to help researchers with their research data management tasks with custom solutions and information material. This is only possible if we have sufficient information about the baseline. We would appreciate your quick response as the survey will be active only from August 27 to September 27, 2018.
>
> The anonymous survey is directed at all levels of the scientific personnel, defined as anyone working with or producing data.
>
> The survey should take about 10-20 minutes and you may access it here: [link]

Three reminders were sent in the course of the survey. The assessment of the participation, i.e. the representativeness of the survey sample was compared to actual roles and departments provided by GFZ Human Resources. However, in the months following the survey, GFZ's organizational structure was changed. In January 2019, i.e. after the survey was completed and before the presentation of the results, the former 6 departments (i.e. Geodesy, Geophysics, Geochemistry, Geomaterials, Geoarchives and Geotechnologies) were re-structured into four new departments: Geodesy, Geophysics, Geochemistry and Geosystems. Sections of the former departments Geoarchives and Geotechnologies were mostly associated to the new departments of Geochemistry and Geosystems. Consequently, the results reflect the old GFZ structure.

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

3. Results

The survey was carried out from August 27 to September 27. The completion rate was 55%, i.e. 226 participants out of 411 completed the survey (Figure 3), corresponding to a participation rate of 24% in the target demographic at GFZ (Figure 4). The participation rate is comparable or even higher than similar surveys (e.g. Paul-Stüve et al., 2015; Simukovic et al., 2013). Participation among section heads or group leaders, postdocs or senior scientists and PhD students was representative, i.e. >20%, with 36%, 29% and 25% participating, respectively. However, no representative samples were obtained among infrastructure support employees (9%), bachelor's and master's students, and student assistants (5%). Consequently, replies falling into these categories were grouped as "others" herein.

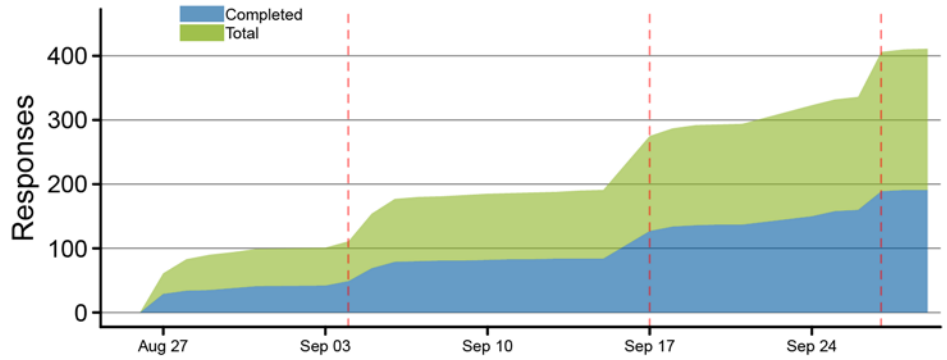


Figure 3 Survey participation showing the total number of attempted and completed interviews. Dashed lines indicate the dates of email reminders.

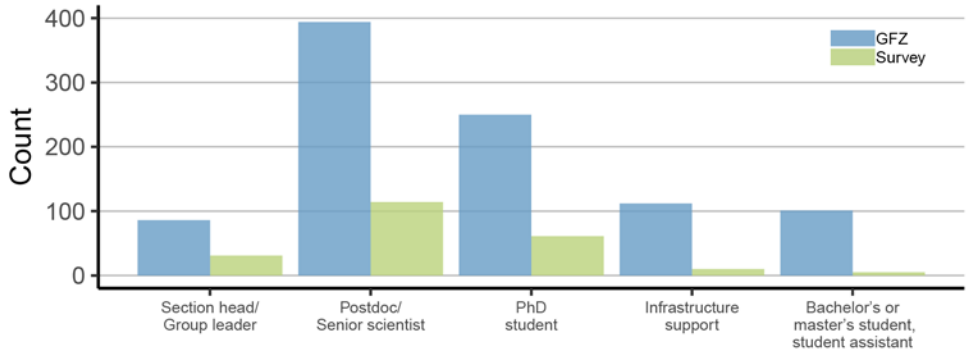


Figure 4 Comparison of survey participation with employees in the survey period.

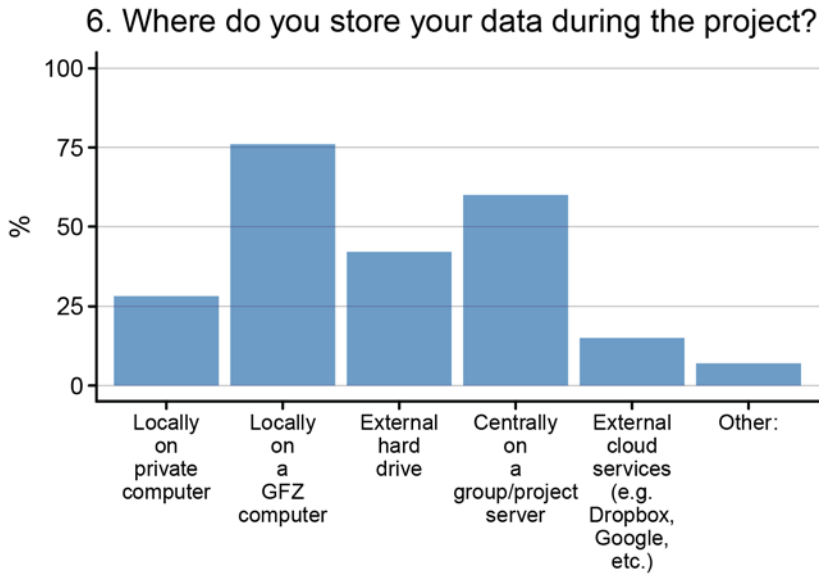
Motivated by the Data Life Cycle, the results are presented with a focus on storage, safety, sharing, documentation and publication of data. In addition, the familiarity of respondents with guideline documents was surveyed. Results are presented matching the survey outline: the subtitles represent sections of the survey, and the graphics showing the question number as it appeared in the survey.

Selected results are presented in this section for the GFZ as whole (not associated to the departments). In addition, the extended results section with all answers grouped by department, role, and employment length are provided in the Appendix. The data are also made available by the authors (Radosavljevic et al, 2019). Detailed results are only mentioned here when they strongly deviate from the overall results. Due to data safety issues, free text answers were summarized and paraphrased to ensure the anonymity of the survey.

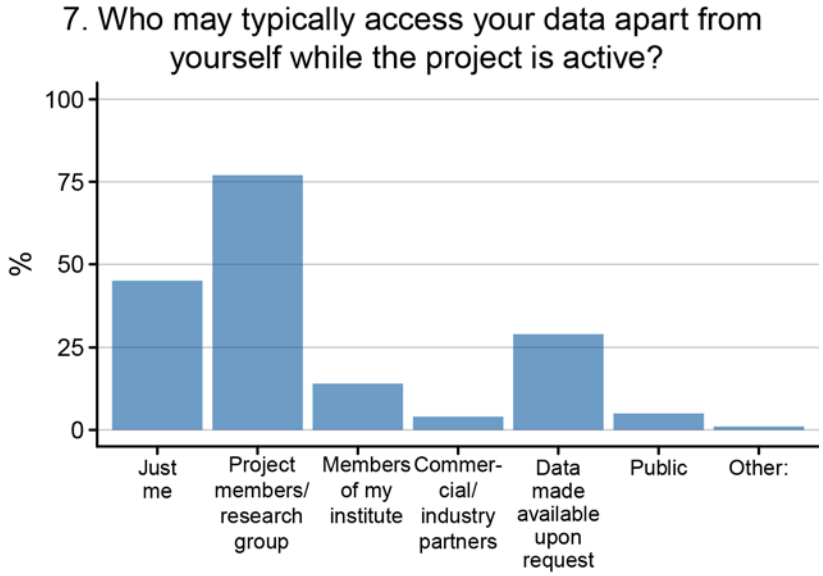
STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

3.1 Data safety

Securing data during a project is an important concern of data management. We inquired about storage use and backup practices. Most respondents rely on their GFZ computers and central group or project servers to store their data. In addition, multiple users may need to access data during the processing and analysis stage. The needs of the community to share data before publication are also revealed in the usage of different sharing approaches.

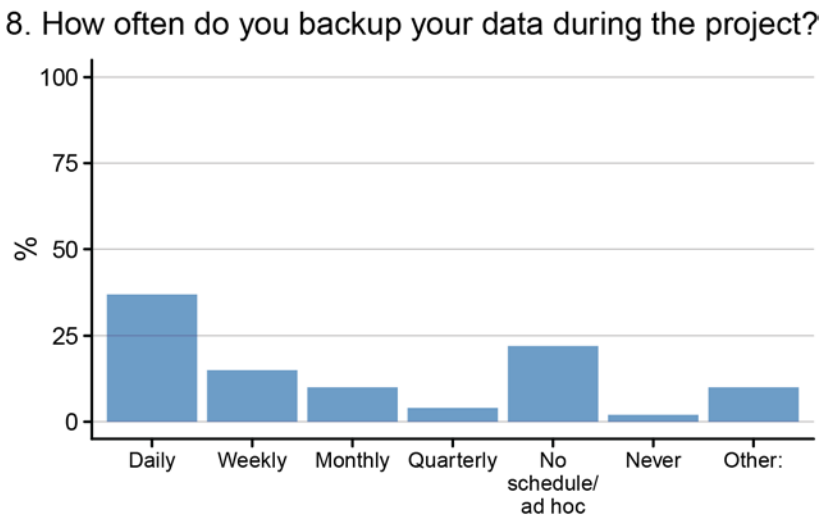


Question 6, multiple response: For active storage, the majority of interviewees rely on local data storage on GFZ computers (76%), on group or project servers (60%), and on external hard drives (42%). Although the usage of different storage is fairly uniform across departments, there are differences. In addition, interviewees specified their usage of Git and the GFZ PowerFolder, among others. (n=226)

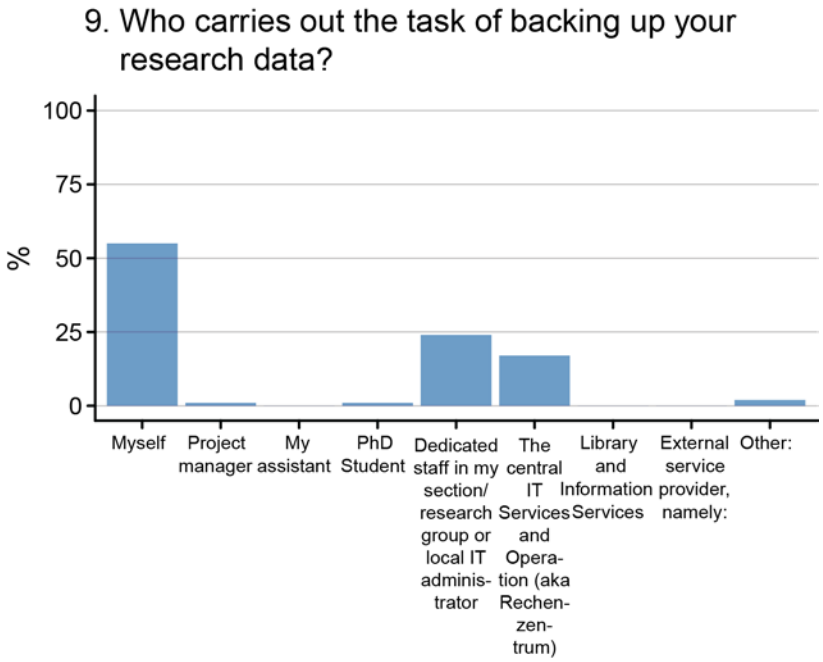


Question 7, multiple response: During the active project, data access by most respondents includes project members, only themselves, or upon request. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

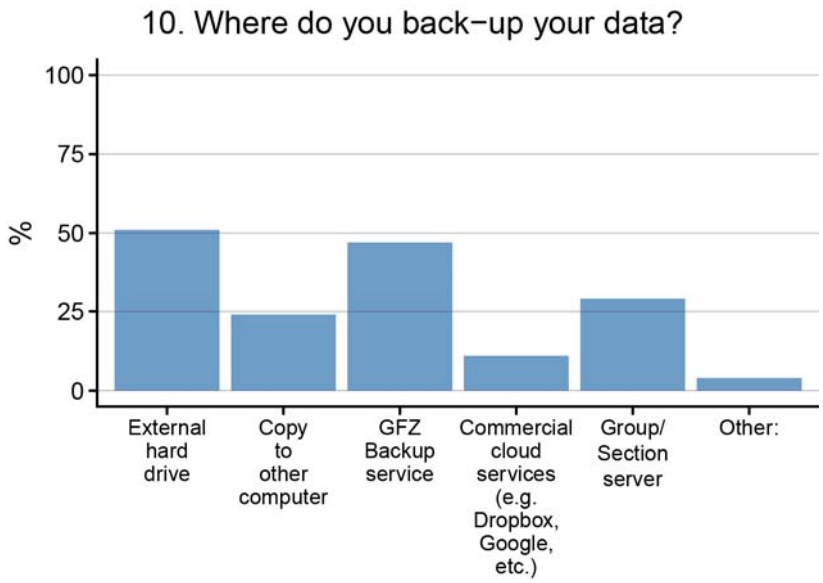


Question 8, multiple response: The largest percentage (37%) of the respondents backup data daily, followed by weekly backups (15%). 22% of respondents do not have a schedule. Free text comments reveal that many rely on services provided by the central IT Services and Operation. (n=226)



Question 9, single response: 55% of the respondents regard back-up within their personal responsibility, while already 41% rely on dedicated staff and the central IT Services and Operation (n=221)

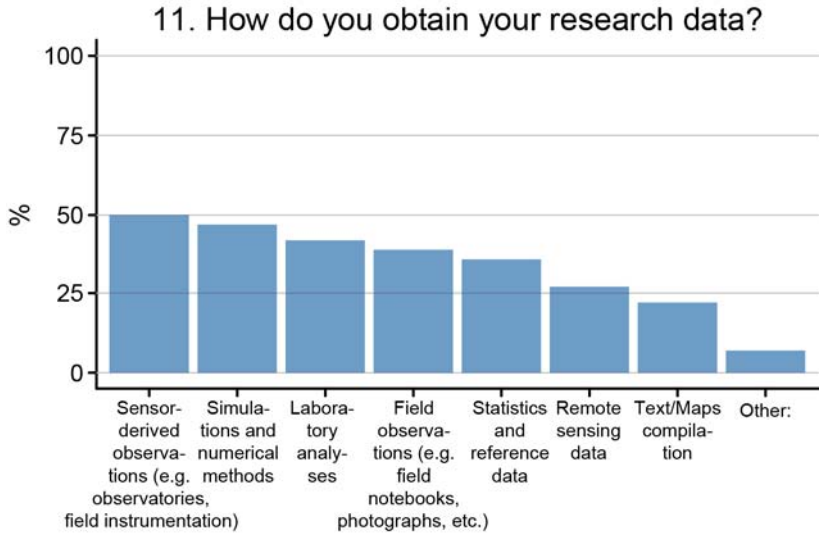
STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 10, multiple response: Most respondents use external hard drives for backup followed by the GFZ backup service, group/section server or another computer. Those employed >2 years rely on GFZ backup service, while those employed <2 years rely on commercial cloud services to a greater extent. Among others, GitLab is used for backup (). (n=221)

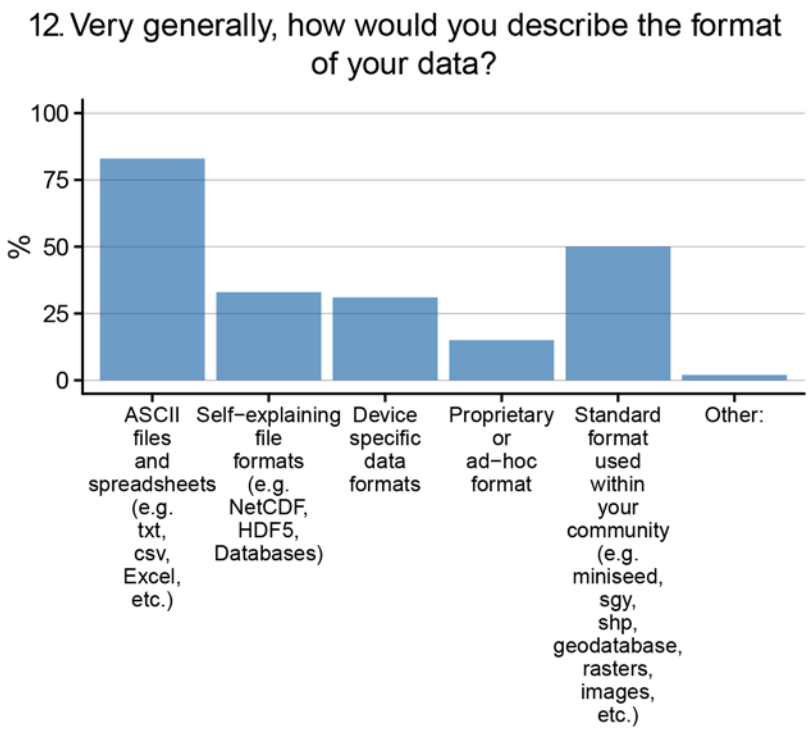
3.2 Data acquisition

During acquisition, processing and analysis stages, data is often kept in native instrument formats read by proprietary software. In the latter stages of the data life cycle increasing the interoperability of data is important and approached by using platform-independent data formats and rich metadata, even though it may degrade precision and accuracy (e.g. Library of Congress, Website). In some cases, the usage of a file format depends on community conventions and chances are that these are standardized and well documented. In other cases own formats and codes are necessary to work with the data. In these cases, documentation is particularly important to increase the interoperability and re-use of data.

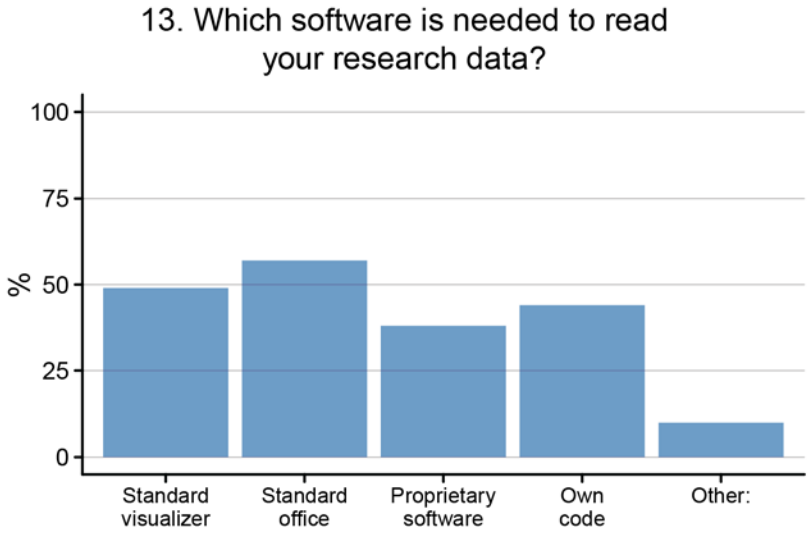


Question 11, multiple response: GFZ researchers using various methods. Consequently, there are considerable differences reflecting the department affiliation (see extended results). National and international data centers to source research data were noted among others. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 12, multiple response: majority of data are in standard formats, while device specific data formats and proprietary formats (31% and 15%) are also used. 15% use proprietary or ad-hoc formats. (n=226)

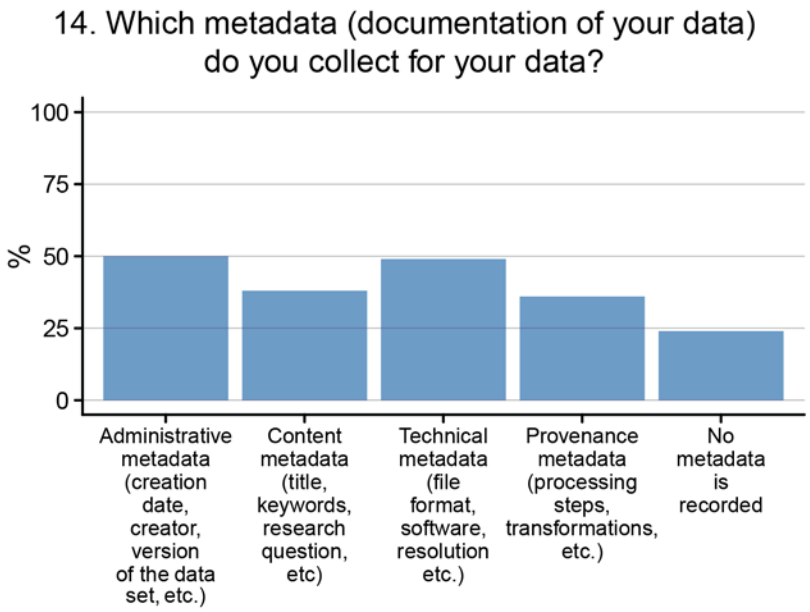


Question 13, multiple response: Most respondents indicate their research data is readable with standard software, but a significant percentage rely on proprietary software (38%) and own code (44%). (n=226)

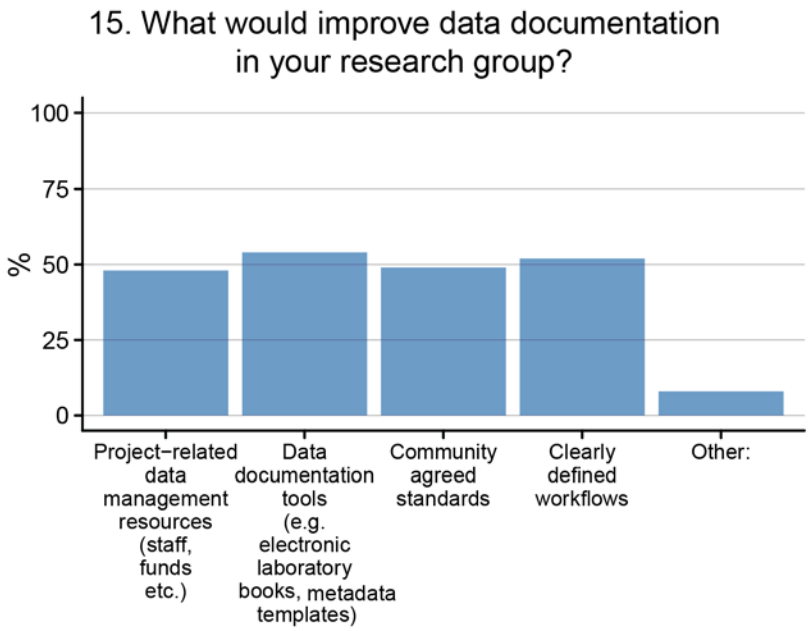
3.3 Data documentation

Documenting data is essential to enable re-tracing of processing steps and enabling the reuse of data for other purposes. Therefore, utilizing metadata standards is highly beneficial as it enables consistent documentation. The results below illustrate that there is a need to improve data documentation practices and provide support for these activities.

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

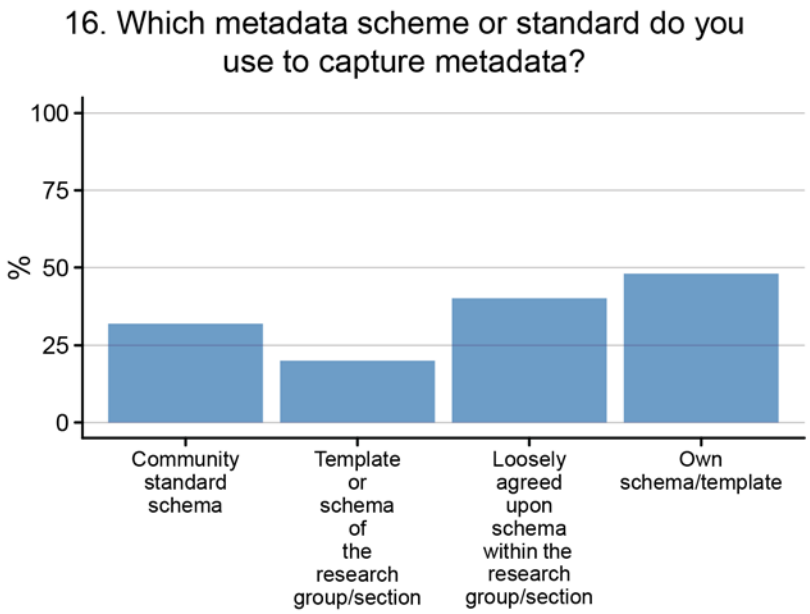


Question 14, multiple response: The figure illustrates that there is room for improvement in all metadata categories. 24% report no metadata is recorded. (n=226)

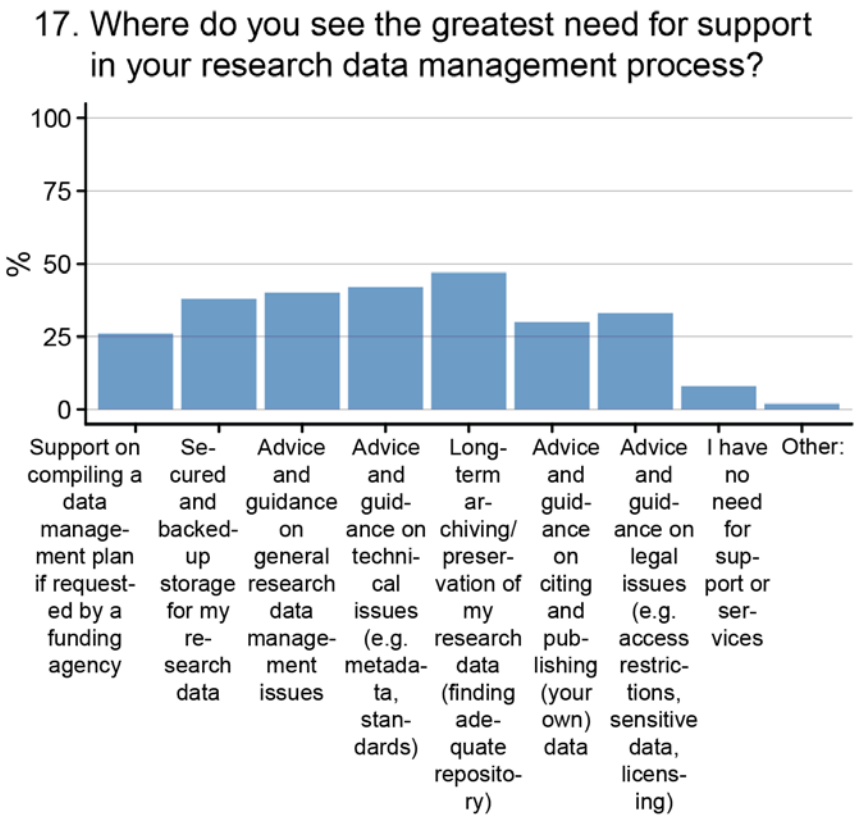


Question 15, multiple response: 54% of respondents view data documentation tools (e.g. electronic laboratory books, metadata templates) as most desirable, however clarity in workflows and responsibilities, along with community agreed standards are also seen as important. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 16, multiple response: Even though there are some standard metadata templates in use, many more respondents rely on their own schemas. (n=171)

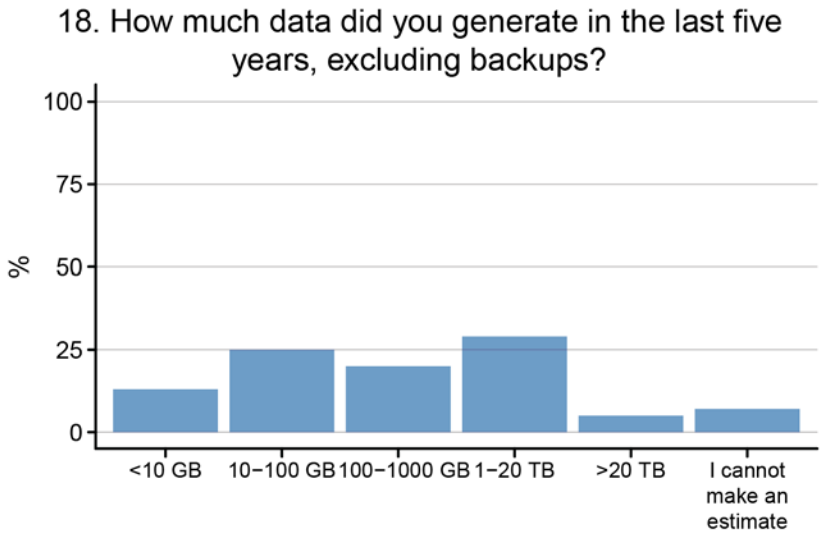


Question 17, multiple response: The greatest self-identified support needs are with long term-archiving, documentation, backup, and advice on licensing and developing a data management plan. 8% of respondents don't need support in their data management. (n=171)

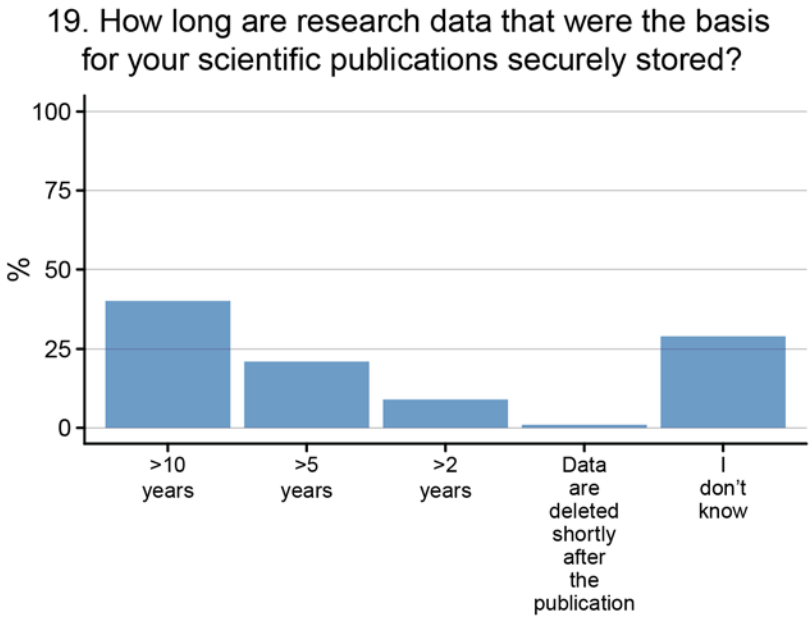
STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

3.4 Data preservation and storage

The archiving of research data is an important aspect of data management. Data that led to a publication must be archived for at least ten years at the institution of origin (DFG, 2006). Archiving encompasses depositing data and sufficient documentation (metadata) for longer periods of time and to allow the particular measurement/experiment to be reproduced or retraced. Even though individual scientists perform this task, the section heads bear the responsibility of upholding the DFG recommendation as defined in the GFZ data policy. Interestingly, 47% would like assistance with data archiving (Question 17).

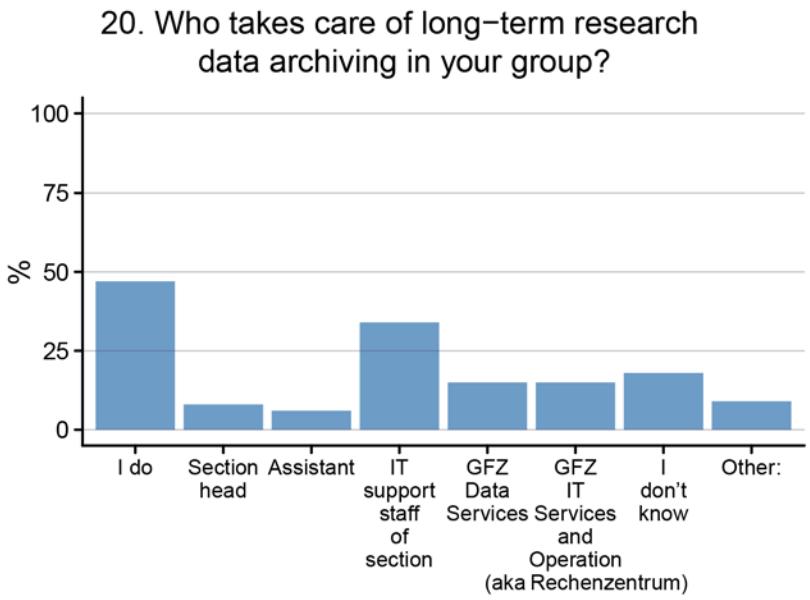


Question 18, single response: The majority of respondents generated more than 100 GB of data in the last five years. Answers among departments vary, with Geodesy, Geophysics, Geomaterials, and Geotechnologies generating more data than other departments. (n=226)

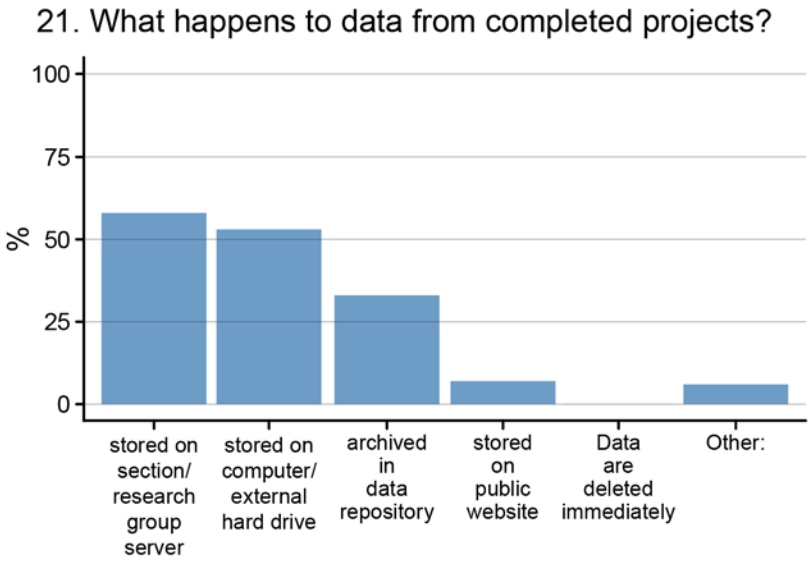


Question 19, single response: Only 40% are aware that primary data underlying publications must be stored for at least 10 years at the institution of origin, which correlates with length of employment. Over 50% of PhD students are not aware of this obligation. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 20, multiple response: 47% personally responsible for data archiving, 34% RZ, more PhD students cite the central IT Services and Operation as responsible. (n=226)

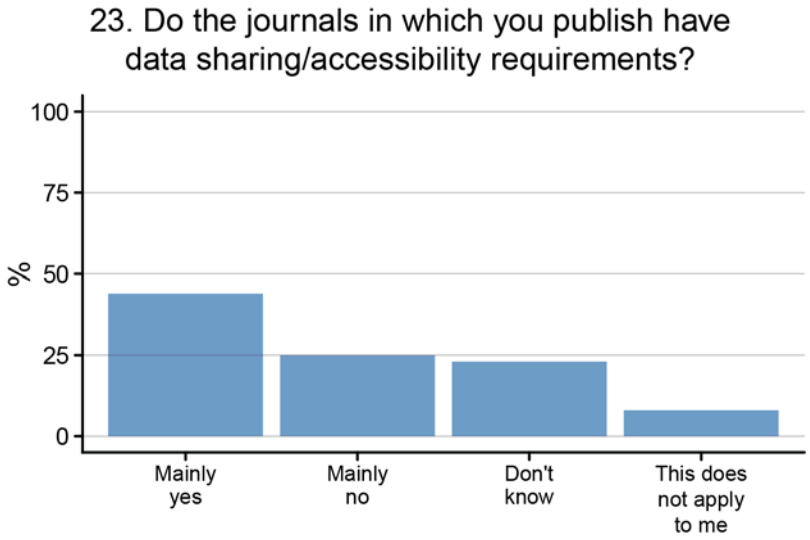


Question 21, multiple response: Just 32% of data is archived in a repository. Many respondents simply do not know what happens to data from completed projects, as expressed in the comments. (n=226)

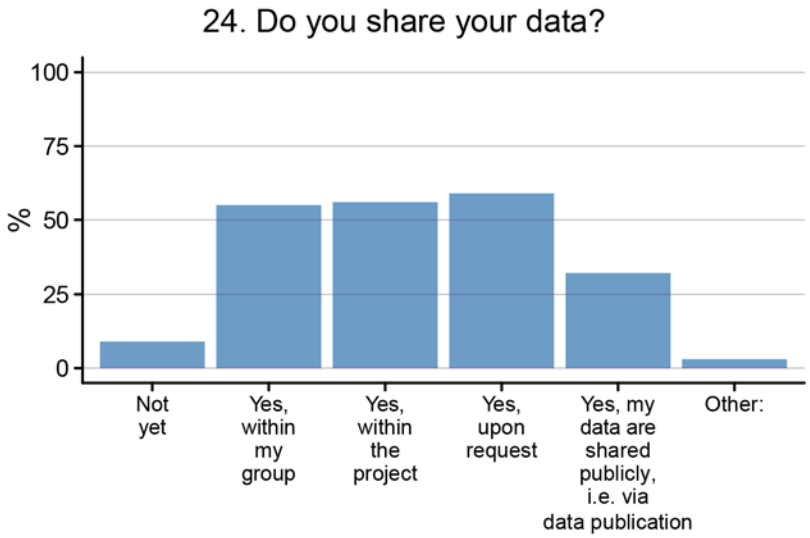
STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

3.5 Data sharing

Major publishers, funding agencies, data repositories, research institutions and many more, are signatories to the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) statement of commitment which strives toward FAIR data (Enabling FAIR Data Community, 2018). In particular, the commitment statement aims to increase data sharing by mandating publishing of data underlying a scholarly publication using the best practice of a data publication in a discipline specific domain repository which issues a digital object identifier (DOI) for the object whenever possible. Guides for selecting a suitable repository are available (Enabling FAIR Data Community, 2018), which can be used in combination with the Registry of Research Repositories (https://www.re3data.org/). Data publications should be cited in the references.

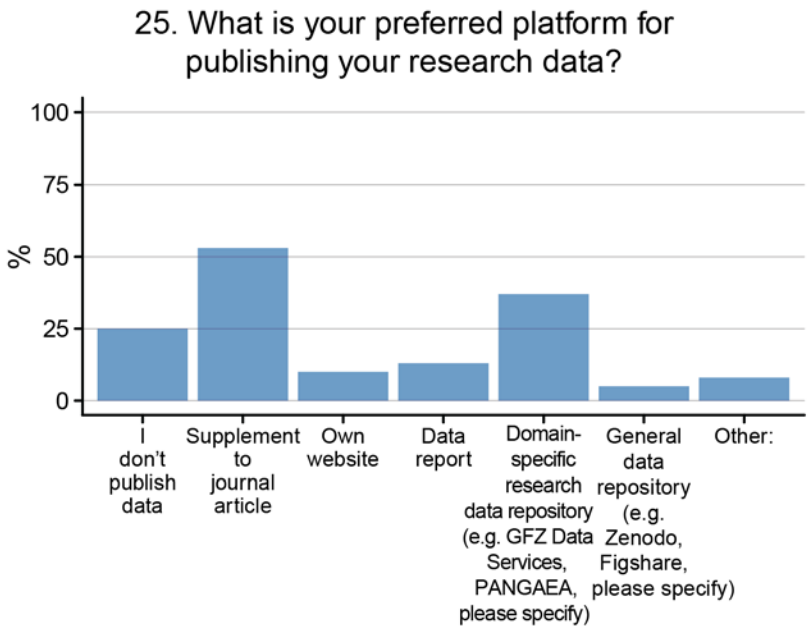


Question 23, single response: Data sharing requirements by journals are largely unknown (48%), but number of "don't know" responses decreases with employment length. (n=226)

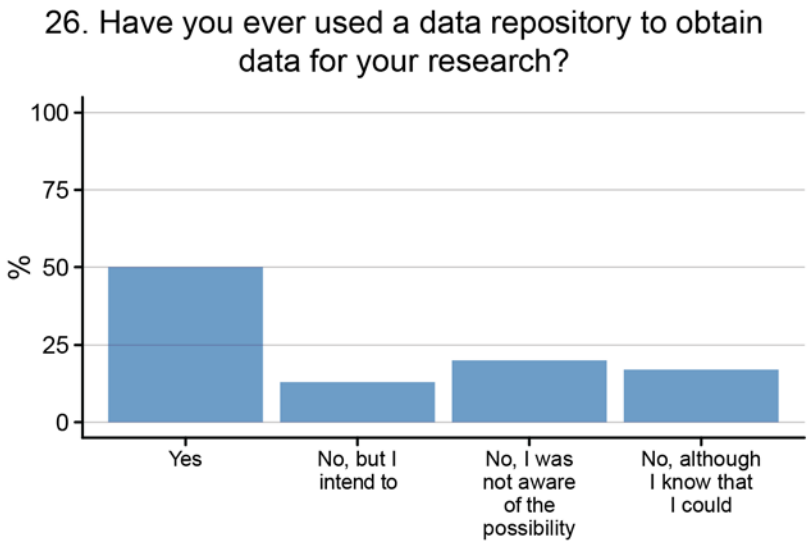


Question 24, multiple response: Only 32% of data is shared using data publications, i.e. deposited in repositories. The question does not specify what data are meant, and it is not clear what happens to data of completed projects. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

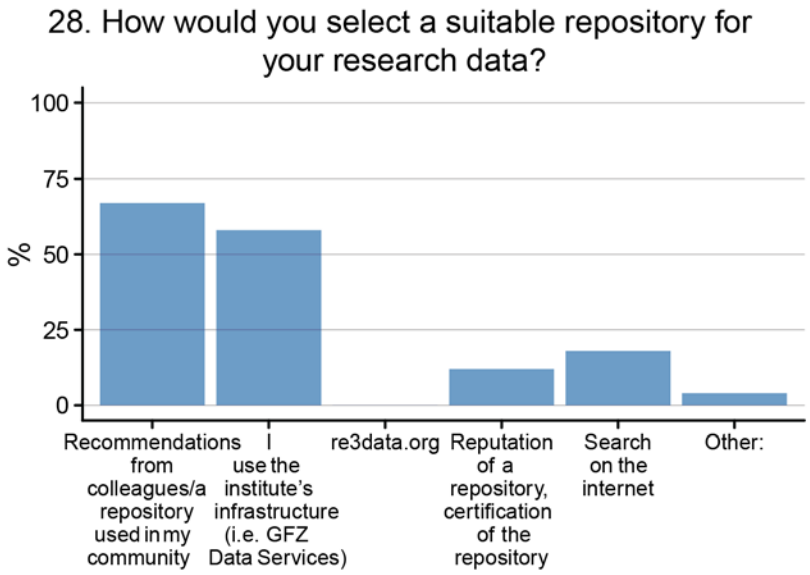


Question 25, multiple response: At the time of the survey, over half of the respondents indicated that they prefer to publish data as journal article supplement. Already 42% following the best practice of sharing via domain specific research data repositories. (n=226)

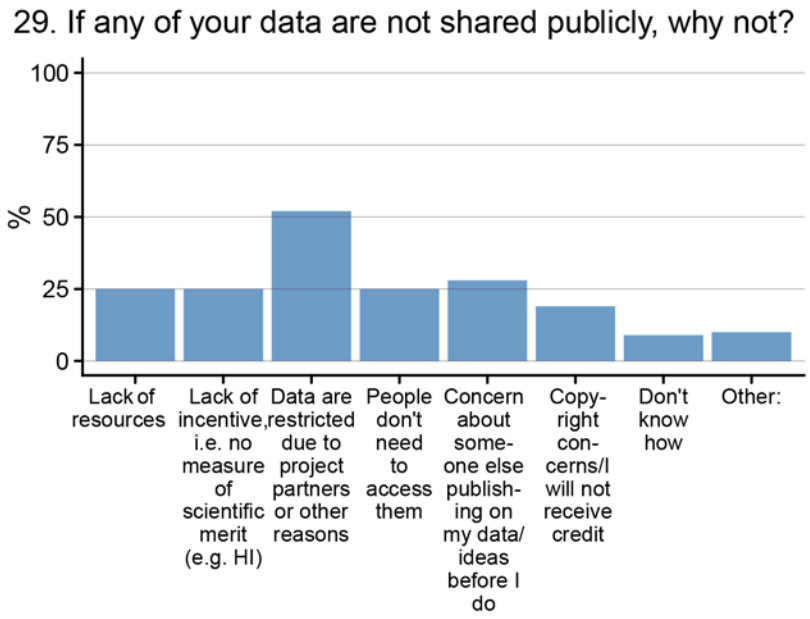


Question 26, single response: 63% have used or plan on using repositories to obtain data (section heads/group leaders and postdoc/senior scientists above average). 20 % do not know this is possible, a percentage even higher among PhD students. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 28, multiple response: 67% of respondents rely on recommendations from colleagues and the institute's infrastructure, i.e. GFZ Data Services (58%), only one respondent would use re3data.org – a database of research data repositories. (n=226)

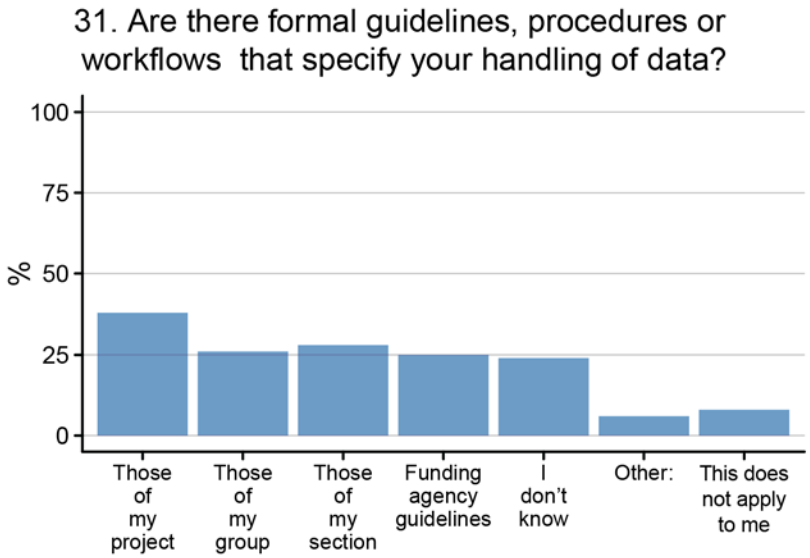


Question 29, multiple response: The main reasons data are not shared publicly are because data are restricted (52%), there is concern about others publishing on data (28%), 25% say data do not need to be accessed, followed or there is a lack of resources or incentive , copyright concerns (19%), 9% don't know how. Others have data too large for sharing, and also state a lack of time/motivation. In contrast, question 30 (extended results) asked about the ***willingness to share data*** *69% only with credit given, 33 % after embargo, 29% with permission, 19% without restrictions. (n=226)*

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

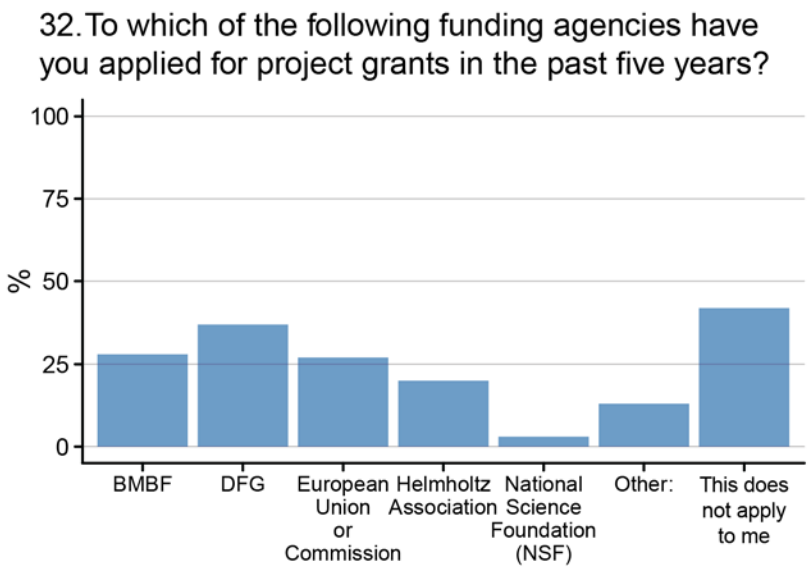
3.6 Data management: workflows, practices and policies

Organization-level data policies provide guidelines for handling data in the organization (e.g. GFZ, 2016). Formal guidelines, workflows and procedures at project, section or laboratory level clarify workflows, duties and responsibilities. As a result, research can be carried out more efficiently: work started by others can be continued easier, and the impact of data or code is increased once published, as well documented and structured workflows in data management increase the likelihood of re-use. Such documents are seen as highly desirable by the respondents of the survey (Question 34). In addition, many funding agencies are requesting a structured document specifying all aspects of data handling in the project at the proposal stage in the form of a data management plan (e.g. Alliance of German Science Organisations, 2010; DFG, 2015; GFZ, 2016; NSF). However, our survey revealed that there is room for improvement in familiarity with established guiding documents (department, section, or lab level), the requirements of funding agencies, the GFZ data guidelines (GFZ, 2016), as well as the DFG Recommendations for Safeguarding Good Scientific Practice (DFG, 2006).

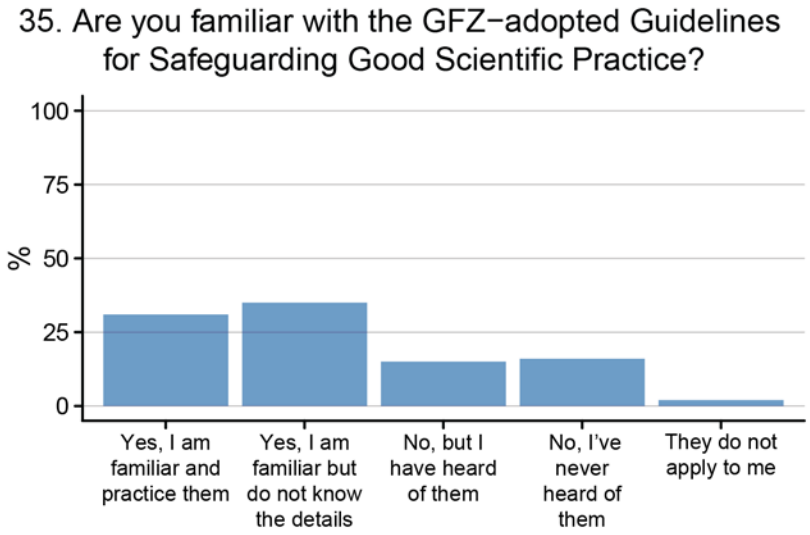


Question 31, multiple response: At least a quarter of respondents indicated that some formal guidelines specify handling of data. On average, 24% of respondents are not aware that any guidelines or workflows apply to their handling of data; this percentage is even higher among PhD students, postdocs/senior scientists, and those employed for less than two years. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

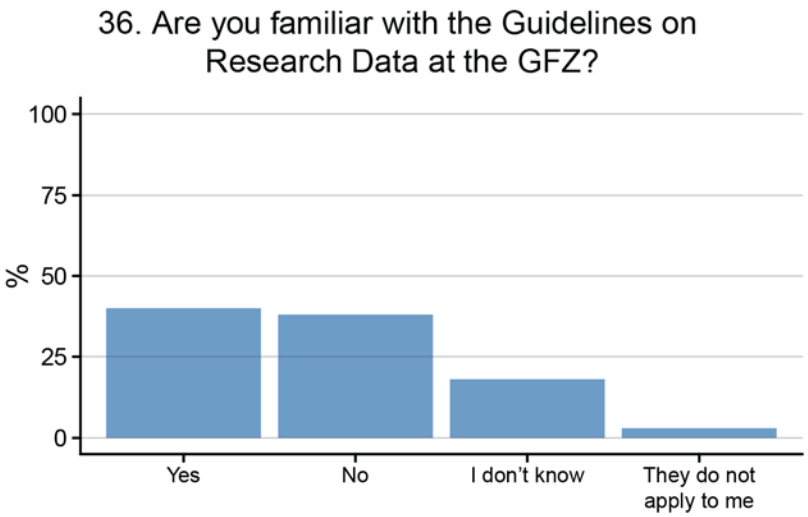


Question 32, multiple response: Respondents indicated they applied for grants at the DFG, BMBF and EU (37%, 28%, and 27%, respectively) in the past five years, all of which require data management plans to be included in grant applications. (n=226)



Question 35, single response: Only 33% of respondents are familiar with and practice the Guidelines for Safeguarding Good Scientific Practice (DFG, 2006). The percentage among new employees (<1 year) who have never heard of them is particularly high (39%) have never heard of them. (n=226)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029



Question 36, single response: The familiarity with the guidelines is relatively low. Only 40% are familiar with the guidelines while 56% of respondents stated that they either are not familiar or do not know if they are. (n=226)

3.7 General comments

The last question of the survey provided a space for respondents to voice general comments or requests regarding the survey, data management, or data publishing in general. Most comments express a desire for defined workflows and guiding documents, at least at the section level. Storage space was also one of the concerns. The importance of data management was recognized by most comments, with some stating that easy tools for documenting data are needed. A small number of respondents expressed concern that excessive requirements or a change of the status quo in terms of data management would diminish scientific output.

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

3.8 References

Alliance of German Science Organisations (2010). Principles for the Handling of Research Data. Retrieved January 29, 2019, from https://www.wissenschaftsrat.de/download/archiv/Allianz-Principles_Research_Data_2010.pdf

Association of European Research Libraries [Website] Open Consultation on FAIR Data Action Plan. Retrieved February 11, 2019 from: https://libereurope.eu/blog/2018/07/13/fairdataconsultation/

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). Retrieved February 11, 2019 from: https://openaccess.mpg.de/Berlin-Declaration.

Collins, S.; Genova, F.; Harrower, N.; Hodson, S.; Jones, S., Laaksonen, L.; Mietchen, D. Petrauskaité, R.; Magnus, V. Wittenburg, P. (2018) Turning FAIR into reality (Final Report and Action Plan from the European Commission Expert Group on FAIR Data). European Commission, Directorate-General for Research and Innovation. Retrieved February 20, 2019 from: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

COPDESS. Coalition on Publishing Data in the Earth and Space Sciences: Statement of Commitment from Earth and Space Science Publishers and Data Facilities 2015. Retrieved February 19, 2019 from: http://www.copdess.org/statement-of-commitment/

Deutsche Forschungsgemeinschaft (DFG) (2006). Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission Selbstkontrolle in der Wissenschaft. Retrieved February 11, 2019 from: https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

DFG, Deutsche Forschungsgemeinschaft. (2015). Leitlinien zum Umgang mit Forschungsdaten. Deutsche Forschungsgemeinschaft. Retrieved from http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

Enabling FAIR Data Community (2018). Commitment Statement to Enabling FAIR Data in the Earth, Space, and Environmental Sciences. https://doi.org/10.5281/zenodo.1451971

Enabling FAIR Data Community, Duerr, R., Kinkade, D., Witt, M., Yarmey, L. (2018). Data Repository Selection Decision Tree for Researchers in the Earth, Space, and Environmental Sciences. https://doi.org/10.5281/zenodo.1475430

European Commission (2013). EU Implementation of the G8 Open Data Charter. Retrieved February 11, 2019 from: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=3489

G8 Science Ministers (2013). G8 Science Ministers Statement; June 12, 2013. Retrieved - February 11, 2019 from: https://www.gov.uk/government/news/g8-science-ministers-statement

GFZ (2016). Guidelines on Research Data at the GFZ German Research Centre for Geosciences. Retrieved January 29, 2019, from http://media.gfz-potsdam.de/gfz/wv/doc/16/GFZ_Daten_Grundsaetze+Erg_en.pdf

GFZ (2017). GFZ Rules for Safeguarding Good Scientific Practice. Retrieved February 11, 2019 from: https://www.gfz-potsdam.de/en/about-us/organisation/board-bodies-administration/bodies/ombudsperson-for-safeguarding-good-scientific-practice/

GoFAIR (Website). FAIR Principles. Retrieved on February 20, 2019 from: https://www.go-fair.org/fair-principles/

Helmholtz Gemeinschaft (2017). Sicherung guter wissenschaftlicher Praxis und Verfahren bei wissenschaftlichem Fehlverhalten. Retrieved February 11, 2019 from: https://www.helmholtz.de/fileadmin/user_upload/01_forschung/wiss_Praxis/HGF_Verfahren_bei_wiss_Fehlverhalten.pdf

Library of Congress (Website). Recommended Formats Statement. Retrieved January 28, 2019, from https://www.loc.gov/preservation/resources/rfs/data.html

National Science Foundation (NSF). Dissemination and Sharing of Research Results. Retrieved January 11, 2019, from https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Paul-Stüve, T., Rasch, G., Lorenz, S. (2014). Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Christian-Albrechts-Universität zu Kiel, 2015. Zenodo, https://doi.org/10.5281/zenodo.32582

Questback EFS Survey (Website) Online Befragungstool. Retrieved August 20. 2018 from: https://www.questback.com/de/online-befragungstool

Radosavljevic, B.; Elger, K.; Bertelmann, R.; Haberland, C.; Hemmleb, S.; Munoz, G.; Quinteros, J.; Strollo, A. (2019): Report on the Survey of Digital Data Management Practices at the GFZ German Research Centre for Geosciences. Scientific Technical Report STR; 19/02. Deutsches GeoForschungsZentrum GFZ, Potsdam, http://doi.org/10.2312/GFZ.b103-19029

Simukovic, E., Kindling, M., Schirmbacher, P. (2013). Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. Humboldt-Universität zu Berlin, Zentraleinrichtung Computer- und Medienservice (Rechenzentrum) https://doi.org/10.18452/13568

Appendix

a. Survey questions

1. What describes your role at GFZ? (Single response question)
    a. Section head/Group leader
    b. Postdoc/Senior scientist
    c. PhD student
    d. Infrastructure support
    e. Bachelor's or master's student, student assistant
    f. Other:
2. Which department do you belong to? (Single response question)
    a. 1. Geodesy
    b. 2. Geophysics
    c. 3. Geochemistry
    d. 4. Geomaterials
    e. 5. Geoarchives
    f. 6. Geotechnologies
    g. 7. Data, Information, and IT Services
    h. Other:
3. How long have you been working at GFZ? (Single response question)
    a. 1-2 years
    b. 2-5 years
    c. > 5 years
4. If you need to re-use a dataset you created some time ago, how much time do you need to find and understand it? (Single response question)
    a. 1-2 hours
    b. A day
    c. 2-4 hours
    d. More than one day
5. Please use the star ranking to assess how easy it would be for a colleague to use, understand and continue working on your research data based on the categories below. (Ranking question)
    a. Folder Structure
    b. File versioning continuity
    c. File naming
    d. Documentation/Metadata
6. Where do you store your data during the project? (Multiple response question)
    a. Locally on private computer
    b. Locally on a GFZ computer
    c. External hard drive
    d. Centrally on a group/project server
    e. External cloud services (e.g. Dropbox, Google, etc.)
    f. Other:
7. Who may typically access your data apart from yourself while the project is active? (Multiple response question)
    a. Just me
    b. Project members/research group
    c. Members of my institute
    d. Commercial/industry partners
    e. Data made available upon request
    f. Public
    g. Other:
8. How often do you backup your data during the project? (Single response question)
    a. Daily

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

b. Weekly
c. Monthly
d. Quarterly
e. No schedule/ad hoc
f. Never
g. Other:

9. Who carries out the task of backing up your research data? (Single response question)
    a. Myself
    b. Project manager
    c. My assistant
    d. PhD Student
    e. Dedicated staff in my section/research group or local IT administrator
    f. The central IT Services and Operation (aka Rechenzentrum)
    g. Library and Information Services staff
    h. External service provider, namely:
    i. Other:
10. Where do you back-up your data? (Multiple response question)
    a. External hard drive
    b. Copy to other computer
    c. GFZ Backup service
    d. Commercial cloud services (e.g. Dropbox, Google, etc.)
    e. Group/Section server
    f. Other:
11. How do you obtain your research data? (Multiple response question)
    a. Field observations (e.g. field notebooks, photographs, etc.)
    b. Sensor-derived observations (e.g. observatories, field instrumentation)
    c. Remote sensing data
    d. Laboratory analyses
    e. Simulations and numerical methods
    f. Statistics and reference data
    g. Text/Maps compilation
    h. Other:
12. Very generally, how would you describe the format of your data? (Multiple response question)
    a. ASCII files and spreadsheets (e.g. txt, csv, Excel, etc.)
    b. Self-explaining file formats (e.g. NetCDF, HDF5, Databases)
    c. Device specific data formats
    d. Proprietary or ad-hoc format
    e. Standard format used within your community (e.g. miniseed, sgy, shp, geodatabase, rasters, images, etc.)
    f. Other:
13. Which software is needed to read your research data? (Multiple response question)
    a. Standard visualizer
    b. Standard office software
    c. Proprietary software
    d. Own code
    e. Other:
14. Which metadata (documentation of your data) do you collect for your data? (Multiple response question)
    a. Administrative metadata (creation date, creator, version of the data set, etc.)
    b. Content metadata (title, keywords, research question, etc)
    c. Technical metadata (file format, software, resolution etc.)
    d. Provenance metadata (processing steps, transformations, etc.)
    e. No metadata is recorded
15. What would improve data documentation in your research group? (Multiple response question)
    a. Project-related data management resources (staff, funds etc.)
    b. Data documentation tools (e.g. electronic laboratory books, metadata templates)
    c. Community agreed standards
    d. Clearly defined workflows

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

e. Other:

16. Which metadata scheme or standard do you use to capture metadata? (Multiple response question)
    a. Community standard schema
    b. Template or schema of the research group/section
    c. Loosely agreed upon schema within the research group/section
    d. General or community specific metadata standards (e.g. Datacite, ISO19115, DublinCore, FGDC, etc.), namely:
    e. Own schema/template
17. Where do you see the greatest need for support in your research data management process? (Multiple response question)
    a. Support on compiling a data management plan if requested by a funding agency
    b. Secured and backed-up storage for my research data
    c. Advice and guidance on general research data management issues
    d. Advice and guidance on technical issues (e.g. metadata, standards)
    e. Long-term archiving/preservation of my research data (finding adequate repository)
    f. Advice and guidance on citing and publishing (your own) data
    g. Advice and guidance on legal issues (e.g. access restrictions, sensitive data, licensing)
    h. I have no need for support or services
    i. Other:
18. How much data did you generate in the last five years, excluding backups? (Single response question)
    a. 10-100 GB
    b. 100-1000 GB
    c. 1-20 TB
    d. > 20 TB
    e. I cannot make an estimate
19. How long are research data that were the basis for your scientific publications securely stored? (Single response question)
    a. > 10 years
    b. > 5 years
    c. > 2 years
    d. Data are deleted shortly after the publication
    e. I don’t know
20. Who takes care of long-term research data archiving of your data in your group? (Multiple response question)
    a. I do
    b. Section head
    c. Assistant
    d. IT support staff of section
    e. GFZ Data Services
    f. GFZ IT Services and Operation (aka Rechenzentrum)
    g. I don’t know
    h. Other:
21. What happens to data from completed projects? (Multiple response question)
    a. Data are deleted immediately
    b. stored on computer/external hard drive
    c. stored on section/research group server
    d. stored on public website
    e. archived in data repository
    f. Other:
22. Do your research data contain sensitive personal information? (Single response question)
    a. Yes
    b. No
    c. I don’t know
23. **Do the jo**urnals in which you publish have data sharing/accessibility requirements? (Single response **question)**
    a. Mainly yes
    b. Mainly no

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

c. Don't know
d. This does not apply to me

24. Do you share your data? (Multiple response question)
    a. Not yet
    b. Yes, within my group
    c. Yes, within the project
    d. Yes, upon request
    e. Yes, my data are shared publicly, i.e. via data publication
    f. Other:
25. What is your preferred platform for publishing your research data? (Multiple response question)
    a. I don’t publish data
    b. Supplement to journal article
    c. Own website
    d. Data report
    e. Domain-specific research data repository (e.g. GFZ Data Services, PANGAEA, please specify)
    f. General data repository (e.g. Zenodo, Figshare, please specify)
    g. Other:
26. Have you ever used a data repository to obtain data for your research? (Single response question)
    a. Yes
    b. No, but I intend to
    c. No, I was not aware of the possibility
    d. No, although I know that I could
27. How would you cite data from other researchers if you are using them in a paper? (Multiple response question)
    a. I include the citation of the data publication including the DOI in the reference list of the paper
    b. I would cite the journal article where the data was used
    c. I am not citing my data sources
    d. Other:
28. How would you select a suitable repository for your research data? (Multiple response question)
    a. Recommendations from colleagues/a repository used in my community
    b. I use the institute’s infrastructure (i.e. GFZ Data Services)
    c. re3data.org
    d. Reputation of a repository, certification of the repository
    e. Search on the internet
    f. Other:
29. If any of your data are not shared publicly, why not? (Multiple response question)
    a. Lack of resources
    b. Lack of incentive, i.e. no measure of scientific merit (e.g. HI)
    c. Data are restricted due to project partners or other reasons
    d. People don't need to access them
    e. Concern about someone else publishing on my data/ideas before I do
    f. Copyright concerns/I will not receive credit
    g. Don't know how
    h. Other:
30. I am willing to share my data publicly... (Multiple response question)
    a. without restrictions
    b. only if there is credit given, i.e. via citation
    c. only with my permission
    d. I would not share my data
    e. Only after a period of my exclusive usage (i.e. embargo)
    f. Other:
31. Are there formal guidelines, procedures or workflows that specify your handling of data? (Multiple response question)
    a. Those of my project
    b. Those of my group
    c. Those of my section
    d. Funding agency guidelines

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

e. I don’t know
f. Other:
g. This does not apply to me

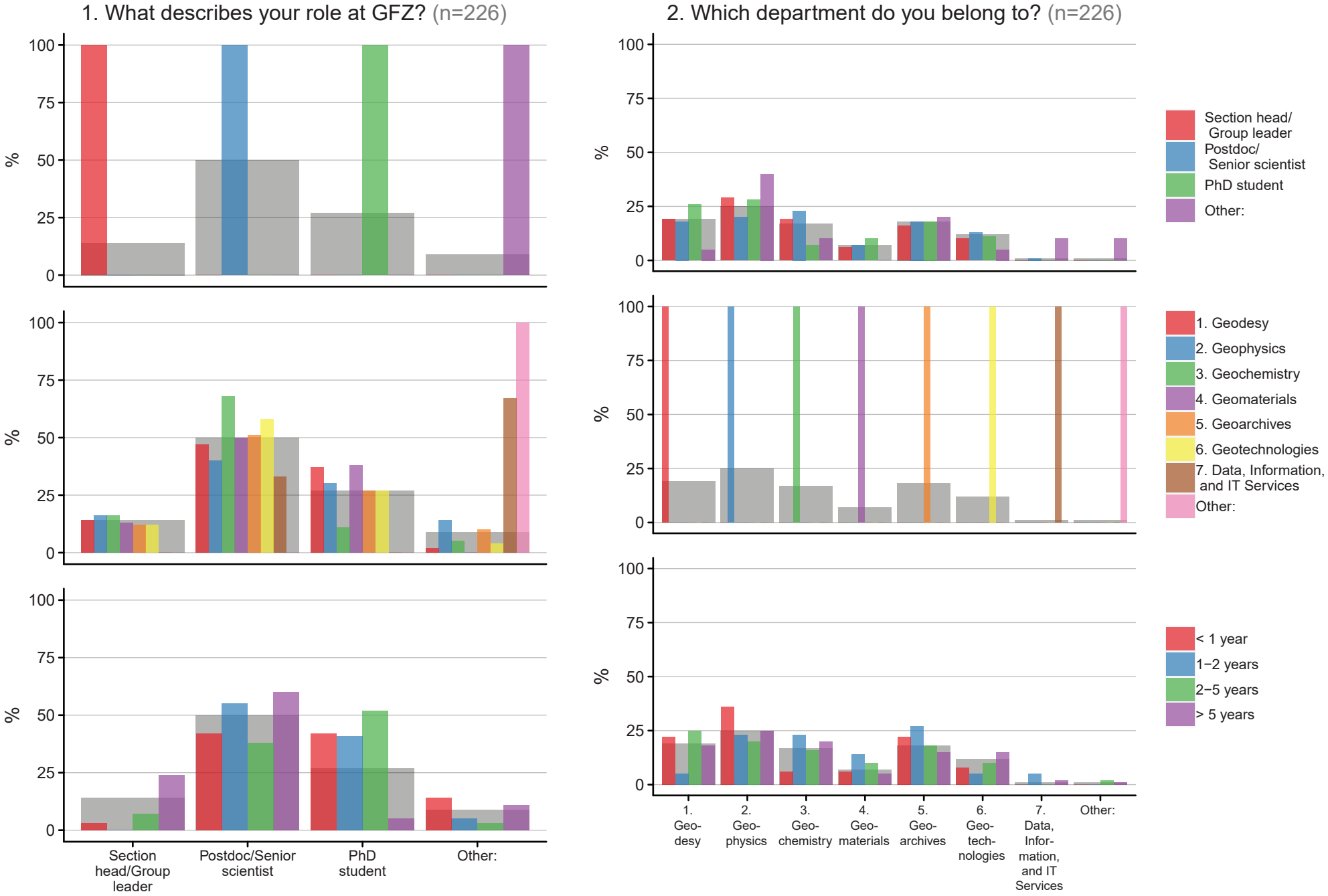
32. To which of the following funding agencies have you applied for project grants in the past five years? (Multiple response question)
    a. BMBF
    b. DFG
    c. European Union or Commission
    d. Helmholtz Association
    e. National Science Foundation (NSF)
    f. Other:
    g. Does not apply to me
33. Did the funding agency/agencies require a data management plan for your project? (Single response question)
    a. Yes
    b. No
    c. This does not apply to me
34. Creating and implementing data management plans for all research projects (e.g. PhD project) or laboratory is beneficial... (Rating question)
35. Are you familiar with the GFZ-adopted Guidelines for Safeguarding Good Scientific Practice? (Single response question)
    a. Yes, I am familiar and practice them
    b. Yes, I am familiar but do not know the details
    c. No, but I have heard of them
    d. No, I’ve never heard of them
    e. They do not apply to me
36. Are you familiar with the “Guidelines on Research Data at the GFZ German Research Centre for Geosciences”? (Single response question)
    a. Yes
    b. No
    c. I don't know
    d. They do not apply to me
37. Do you have any comments or requests regarding this survey, data management, or data publishing in general? These comments are for internal use only and will not be made public. (Free text)

STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029

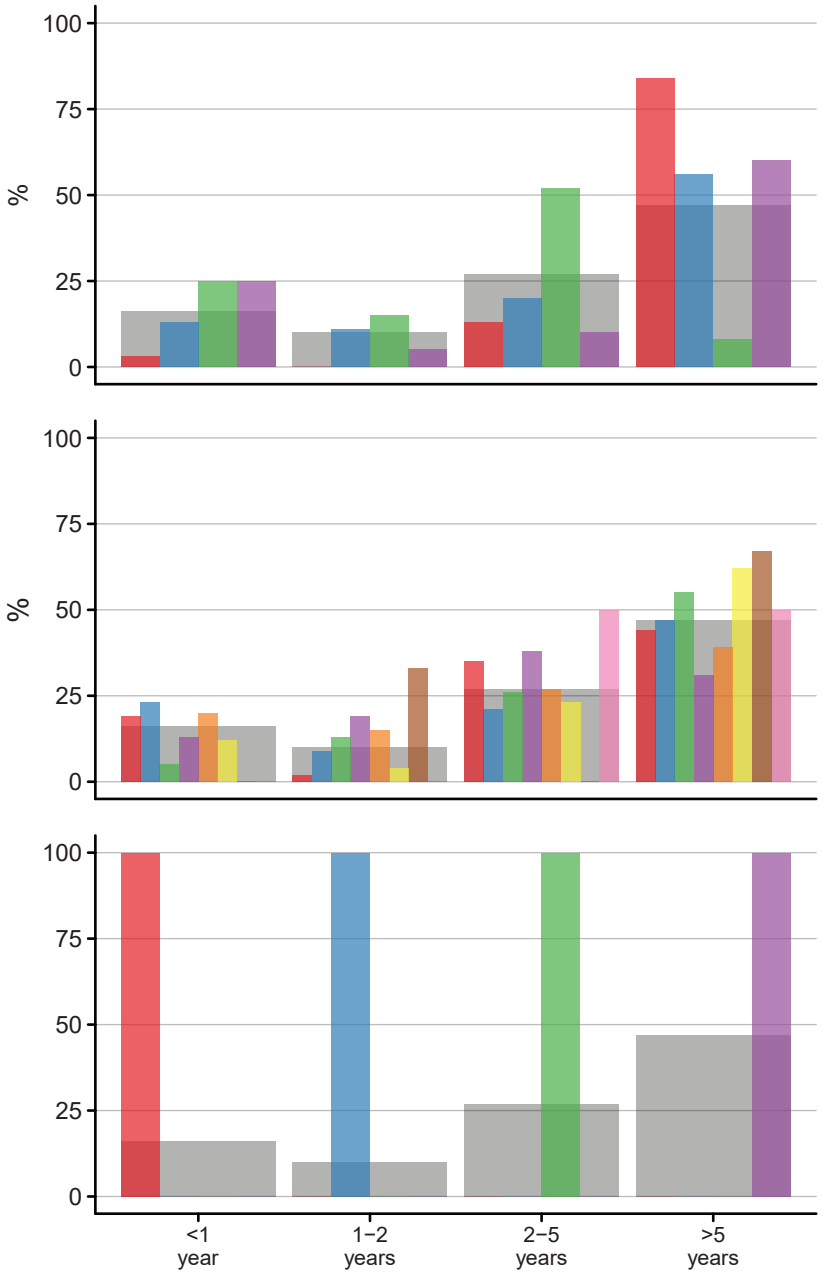
b. Extended Results

The following section provides graphs of each question split by role, department and employment length of each interviewee. The grey shaded area behind the graphs shows the mean without the split. Questions asking respondents for a ranking are shown as arithmetic means of the responses (Questions 5 and 34).

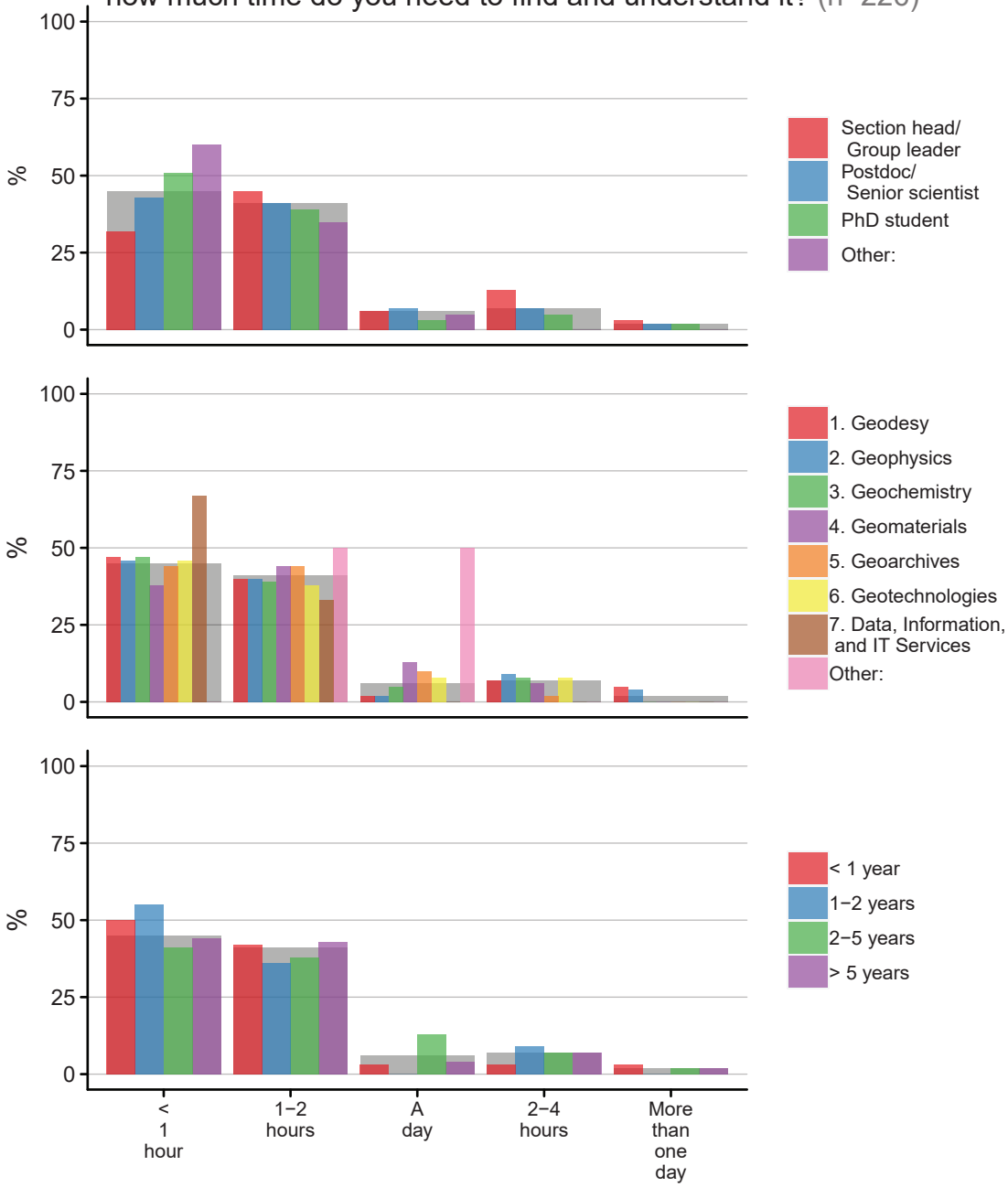
STR 19/02. GFZ German Research Centre for Geosciences.
DOI: 10.2312/GFZ.b103-19029


1. What describes your role at GFZ? (n=226)
2. Which department do you belong to? (n=226)
%
100
75
50
25
0
Section head/Group leader
Postdoc/Senior scientist
PhD student
Other:
1. Geodesy
2. Geophysics
3. Geochemistry
4. Geomaterials
5. Geoarchives
6. Geotechnologies
7. Data, Information, and IT Services
< 1 year
1−2 years
2−5 years
> 5 years
1. Geo-desy
2. Geo-physics
3. Geo-chemistry
4. Geo-materials
5. Geo-archives
6. Geo-tech-nologies
7. Data, Infor-mation, and IT Services

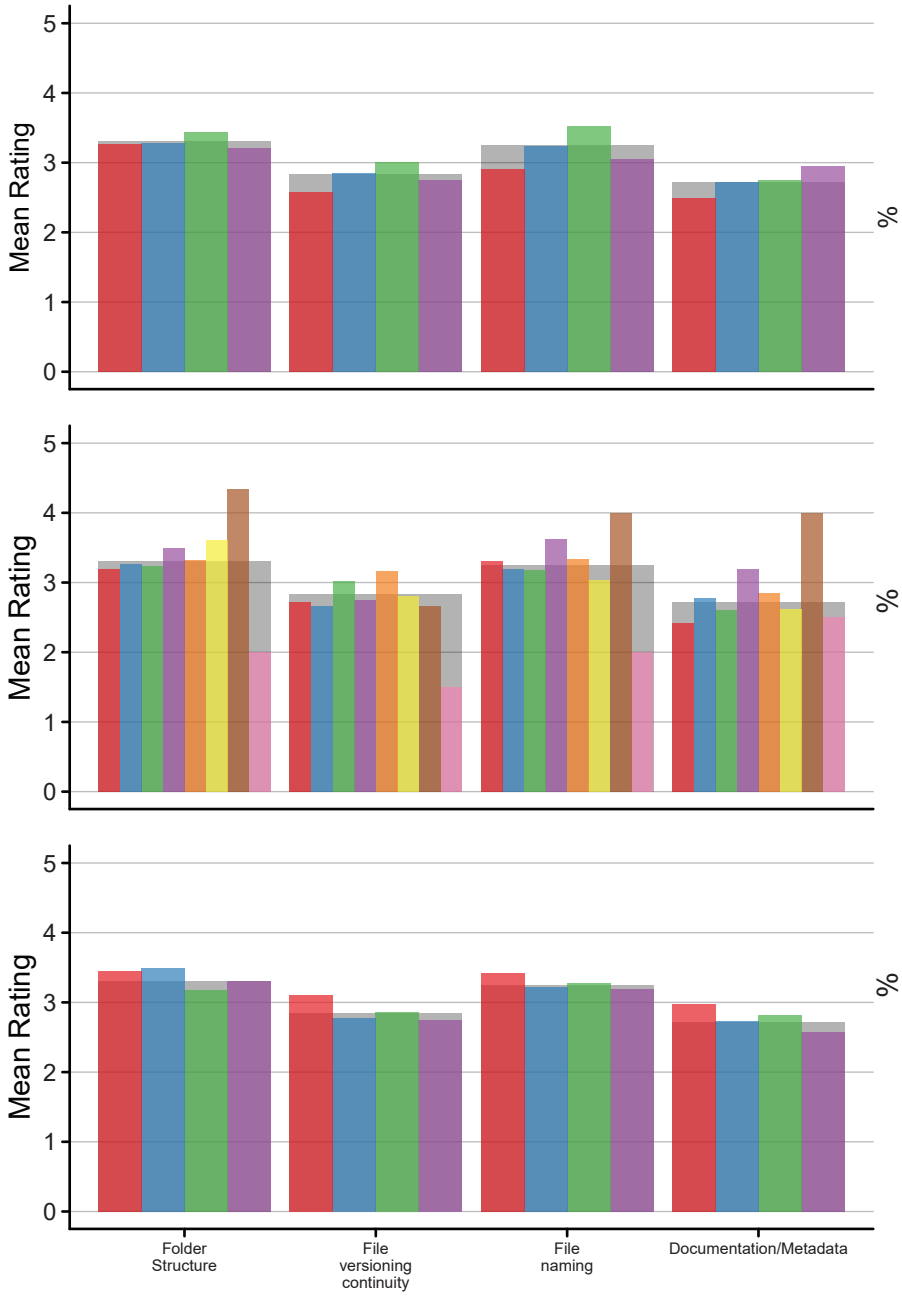
3. How long have you been working at GFZ? (n=226)



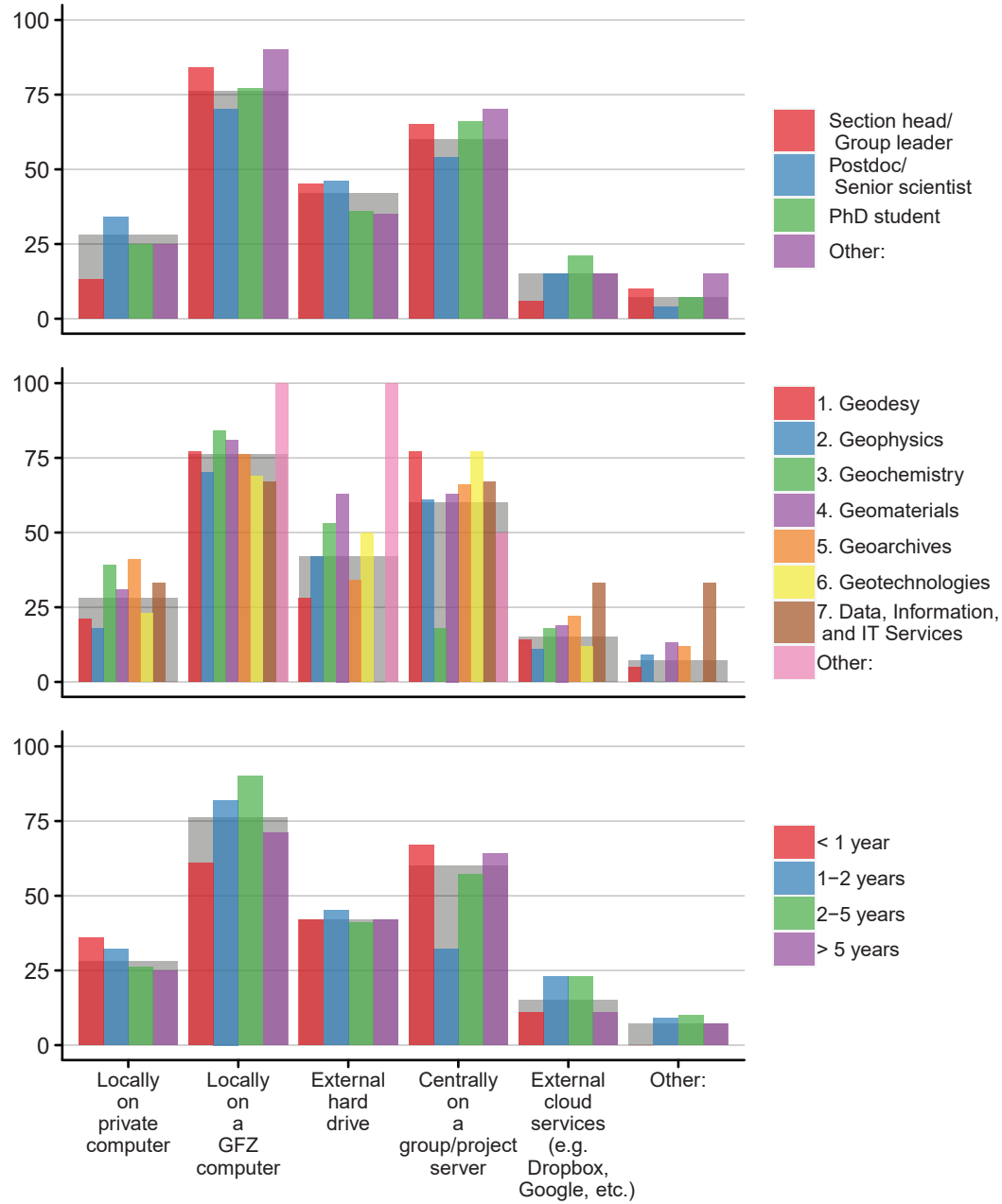
4. If you need to re−use a dataset you created some time ago, how much time do you need to find and understand it? (n=226)



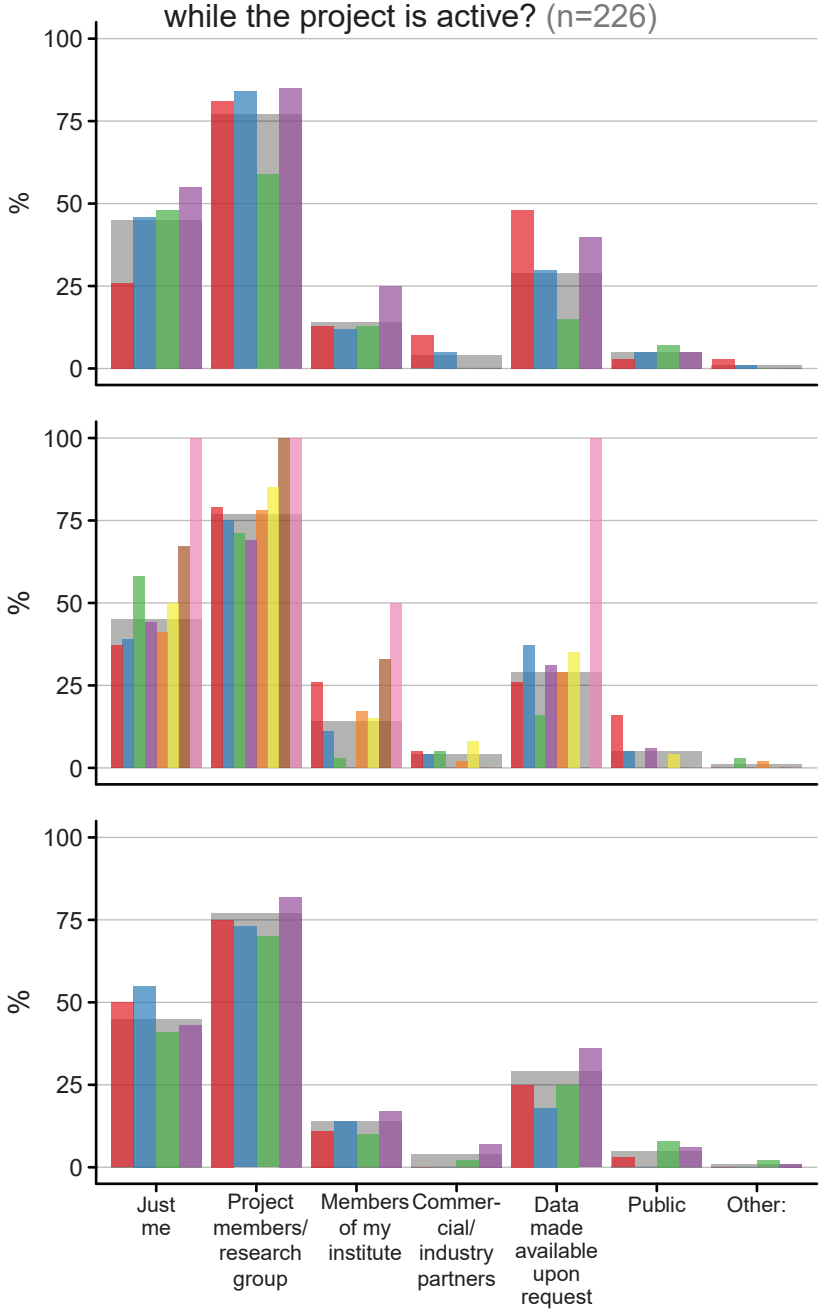
5. Please rank from low to high (1-5) how easy it would be for a colleague to use, understand and continue working on your research data based on the categories below. (n=226)



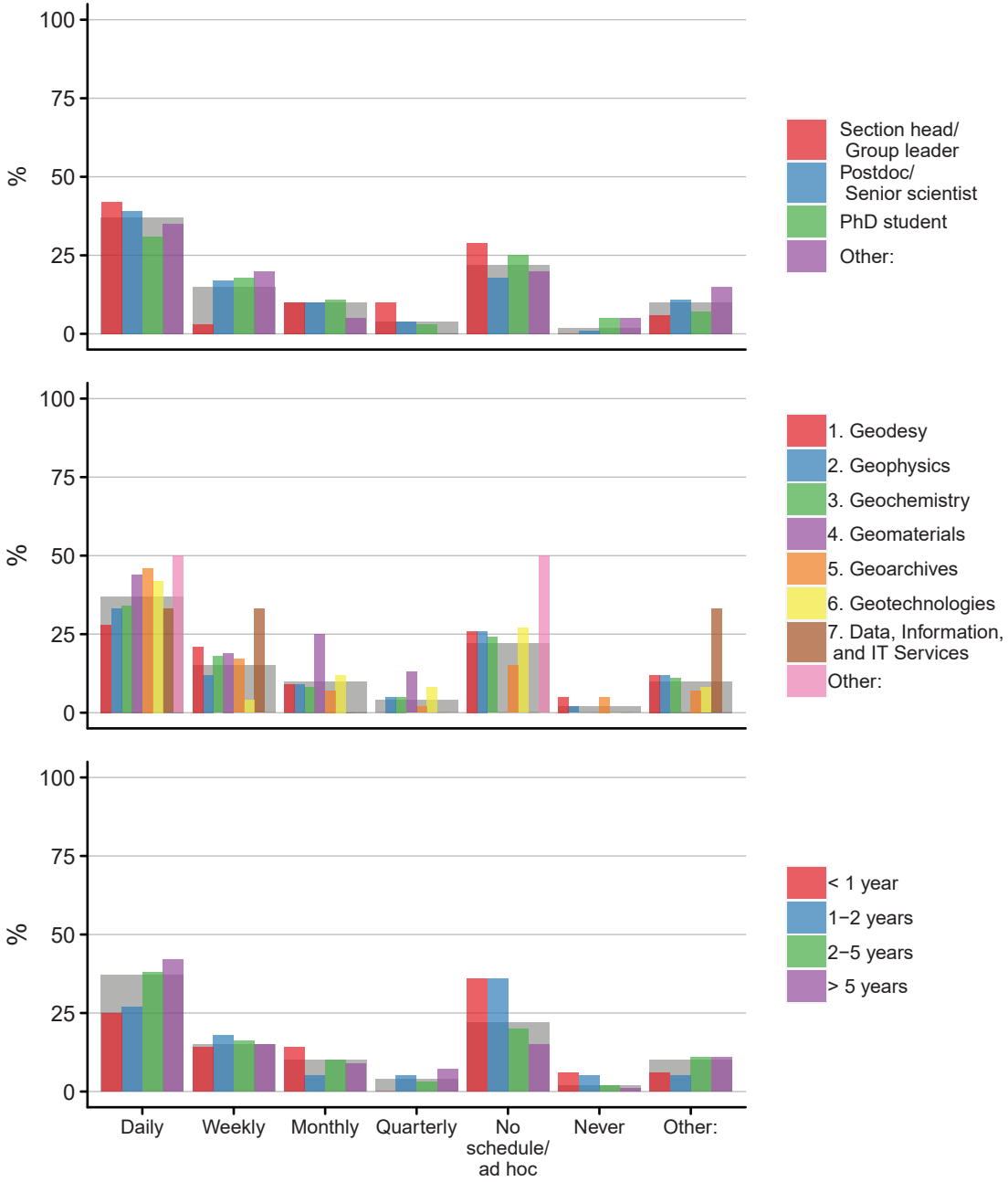
6. Where do you store your data during the project? (n=226)



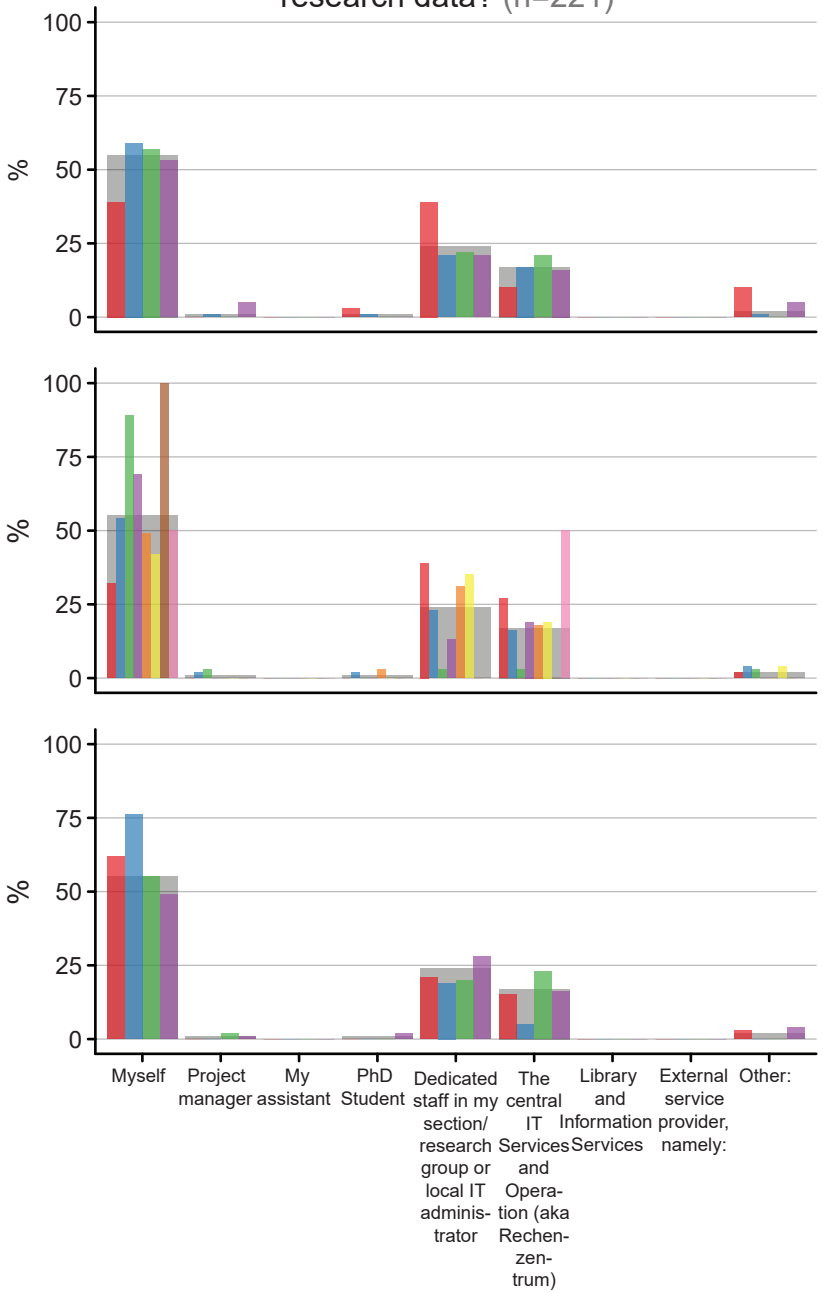
7. Who may typically access your data apart from yourself while the project is active? (n=226)



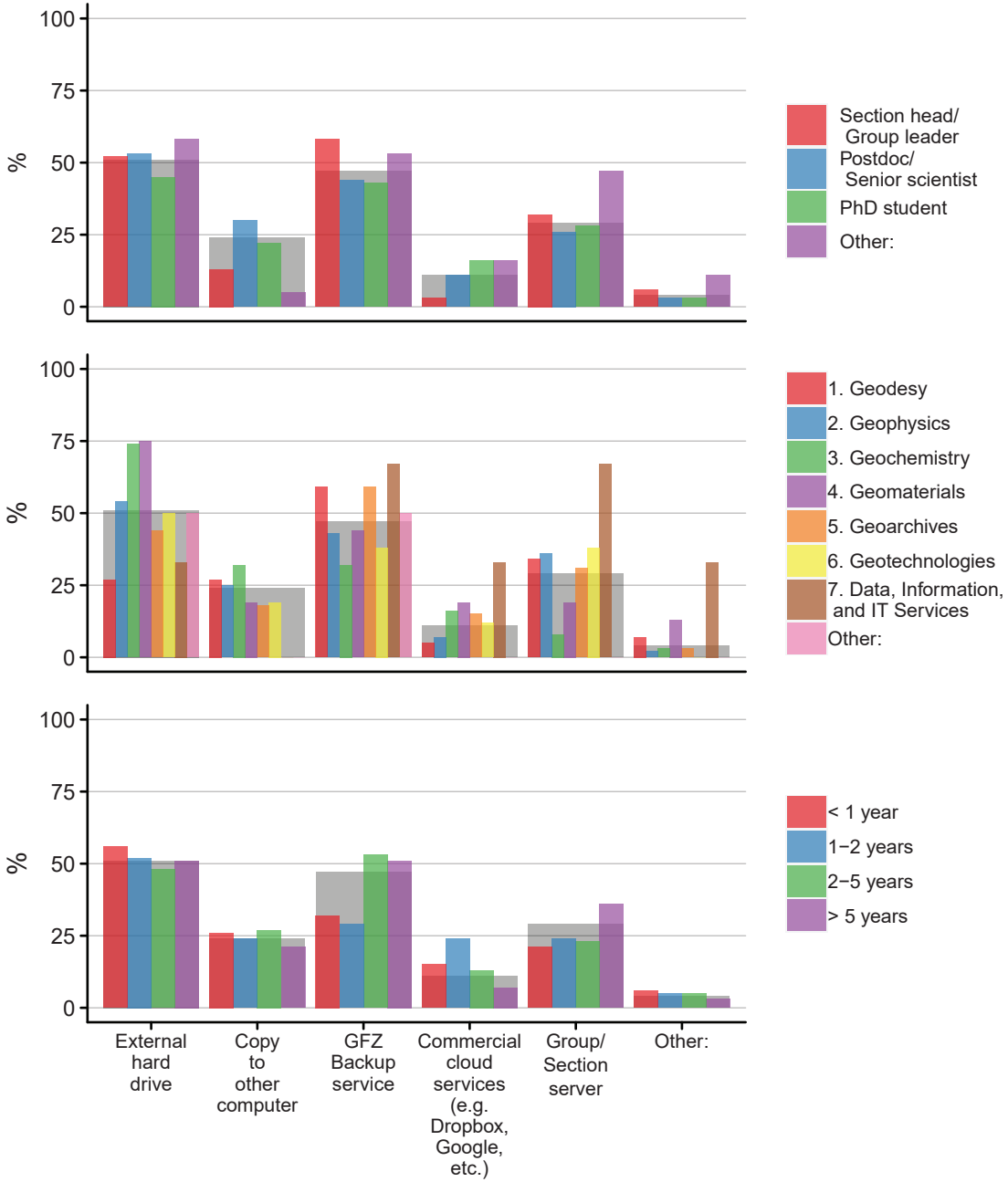
8. How often do you backup your data during the project? (n=226)



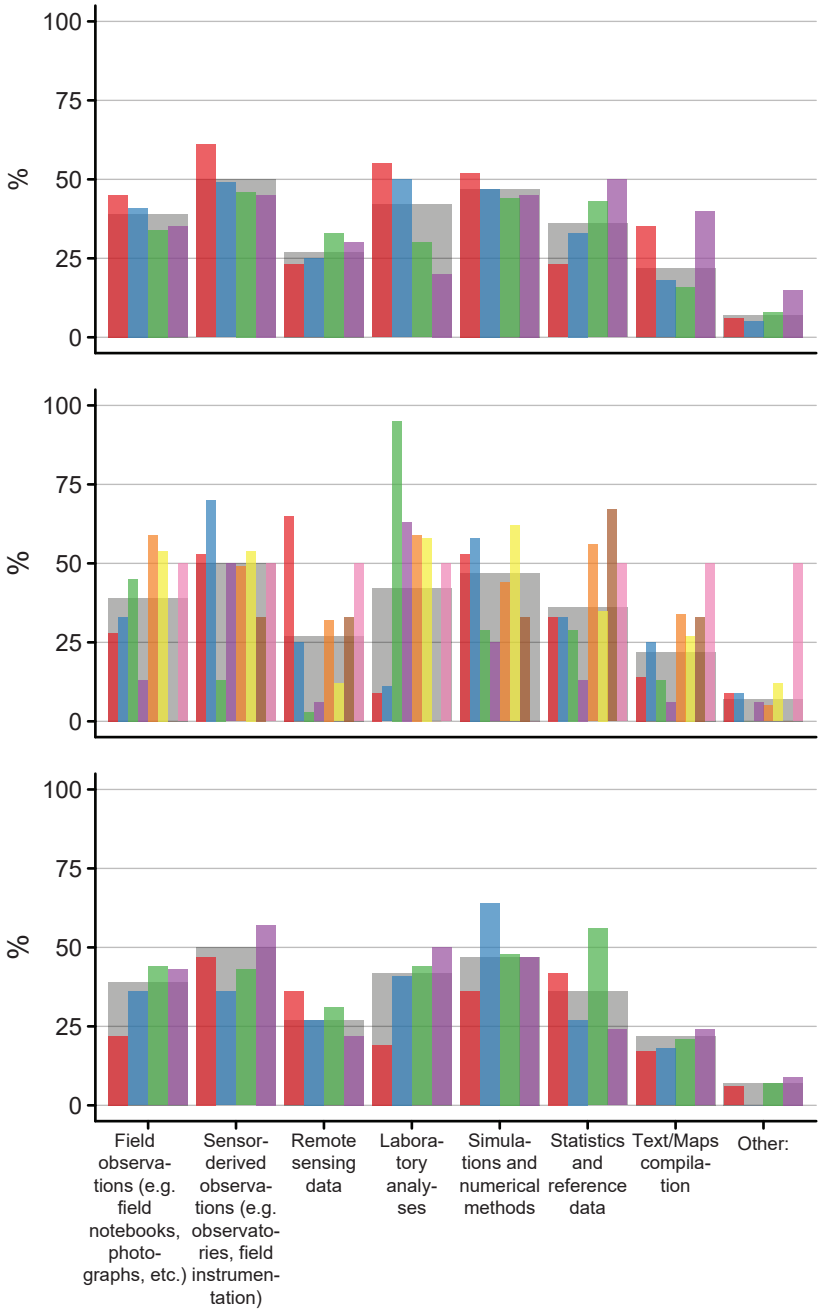
9. Who carries out the task of backing up your research data? (n=221)



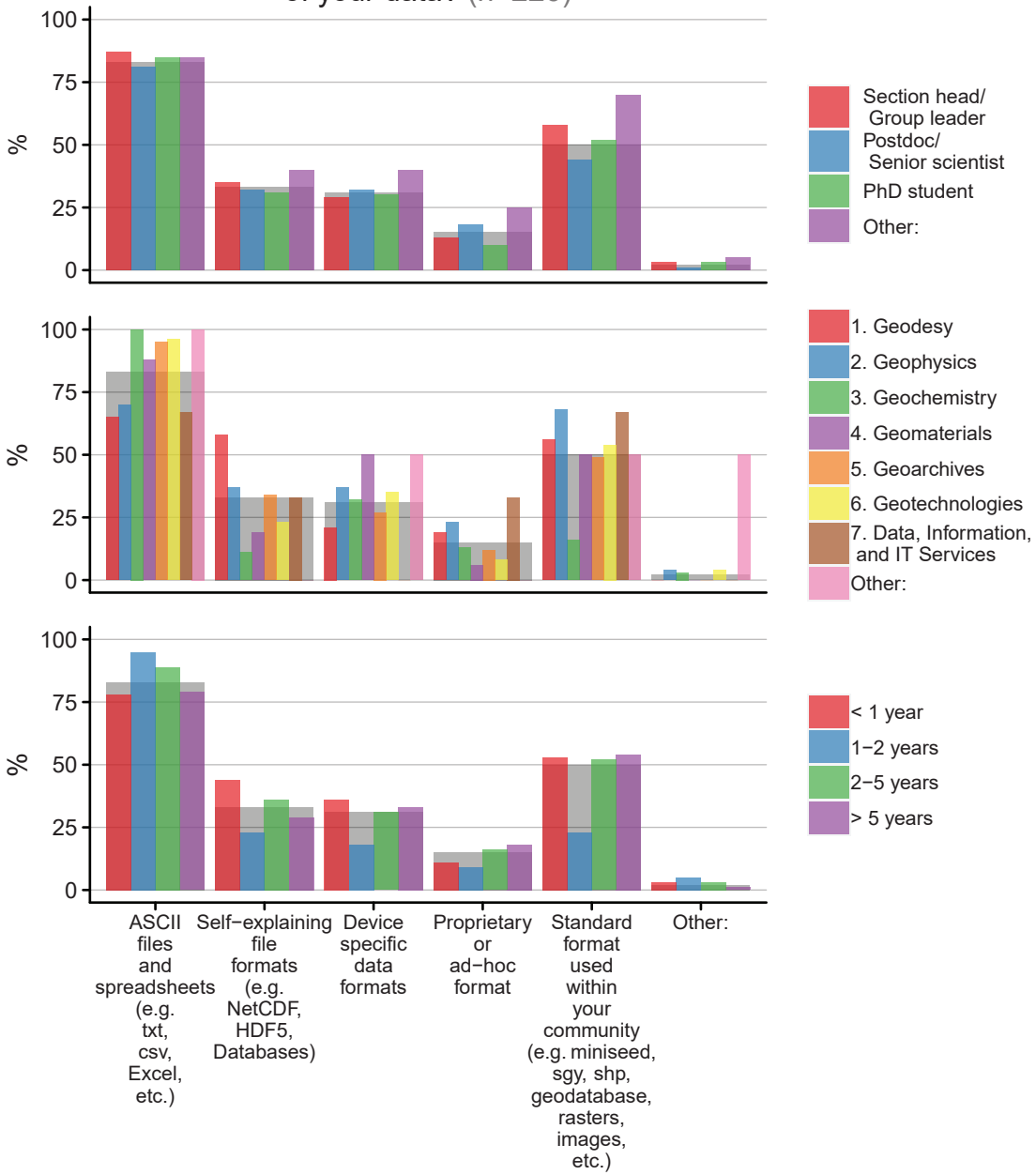
10. Where do you back−up your data? (n=221)



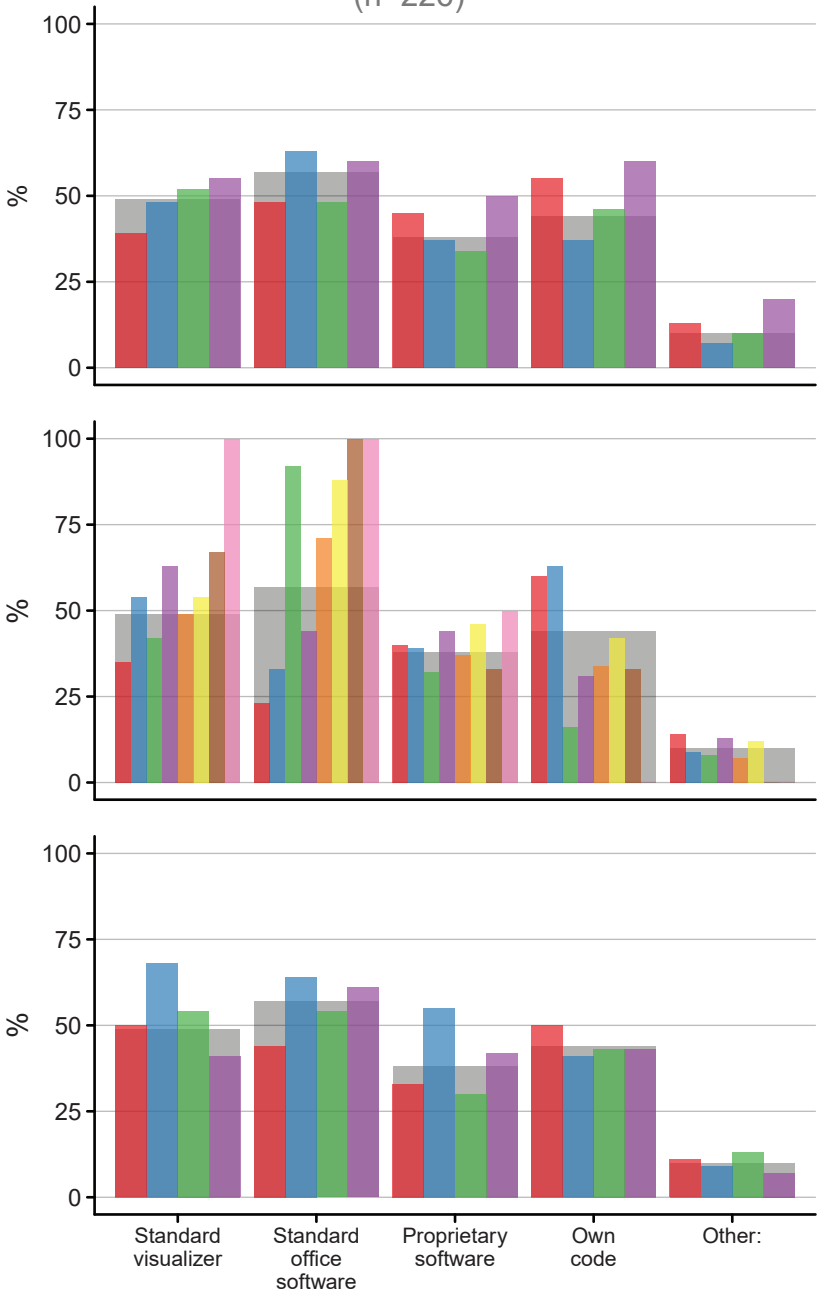
11. How do you obtain your research data? (n=226)



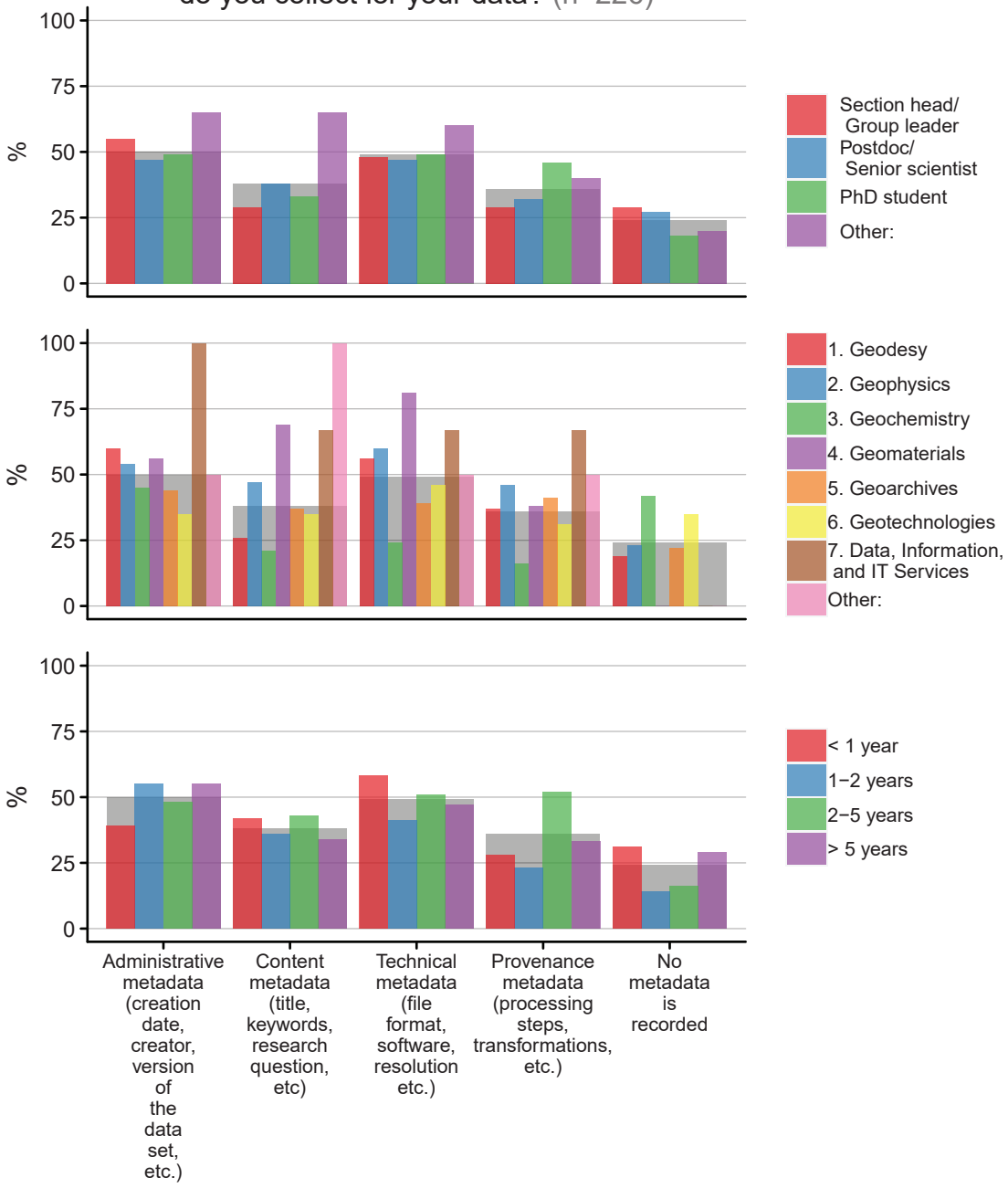
12. Very generally, how would you describe the format of your data? (n=226)



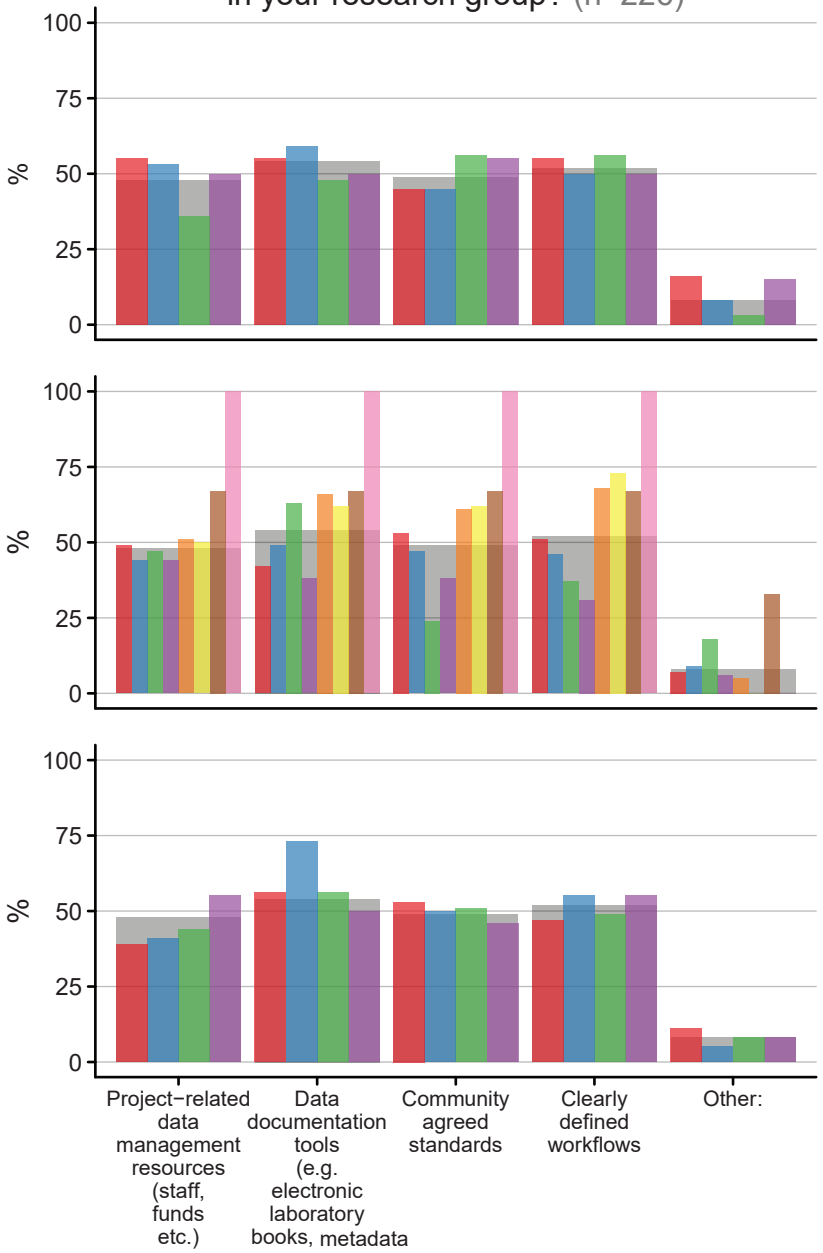
13. Which software is needed to read your research data? (n=226)



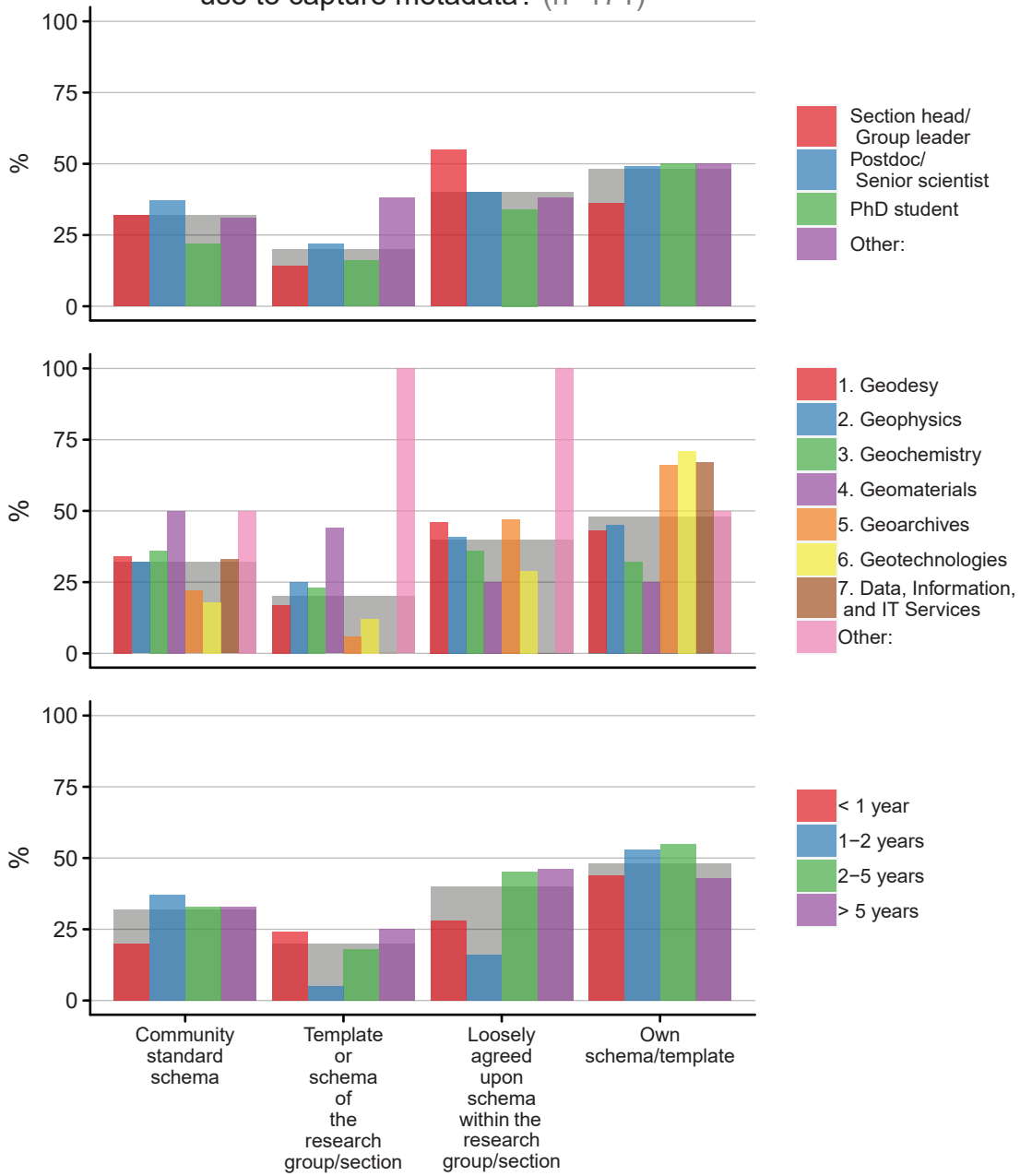
14. Which metadata (documentation of your data) do you collect for your data? (n=226)



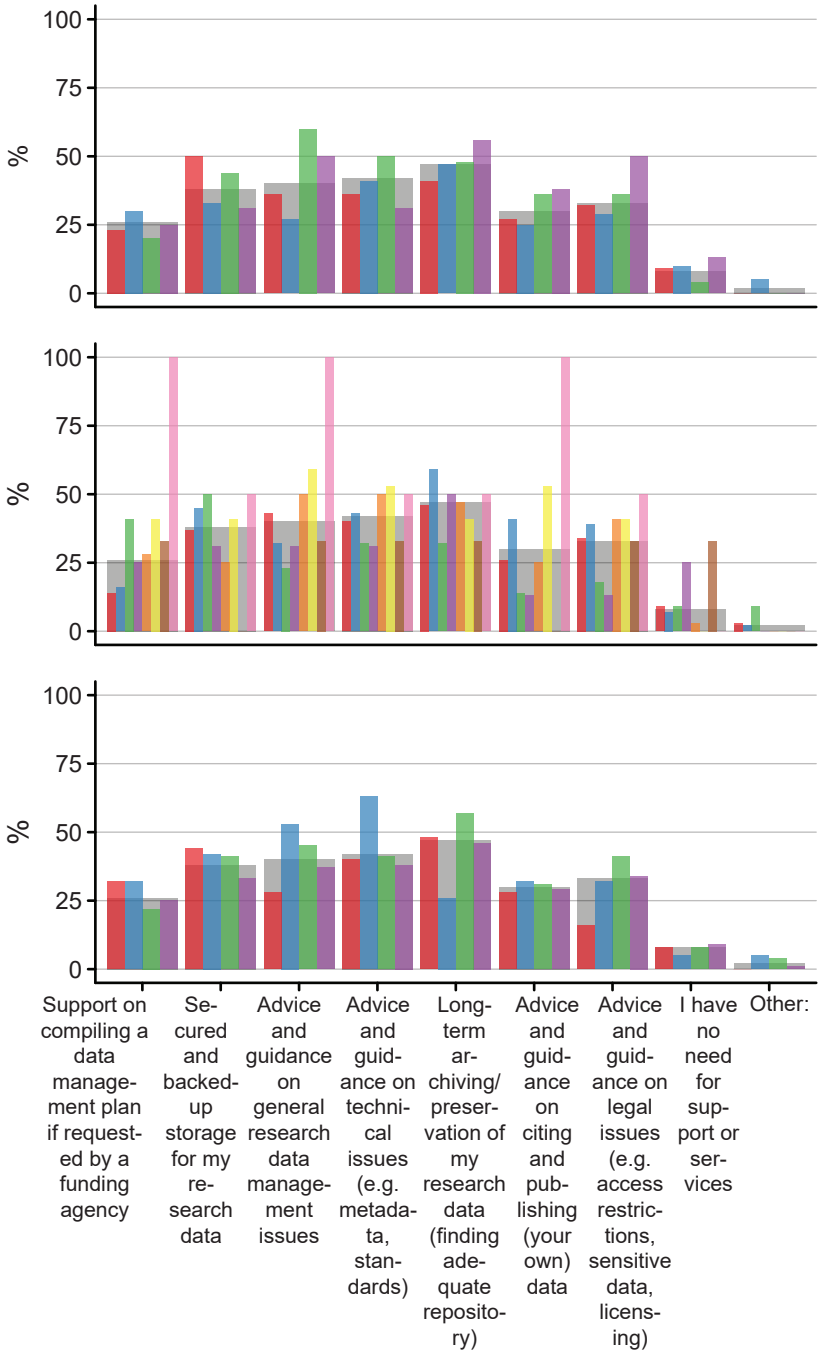
15. What would improve data documentation in your research group? (n=226)



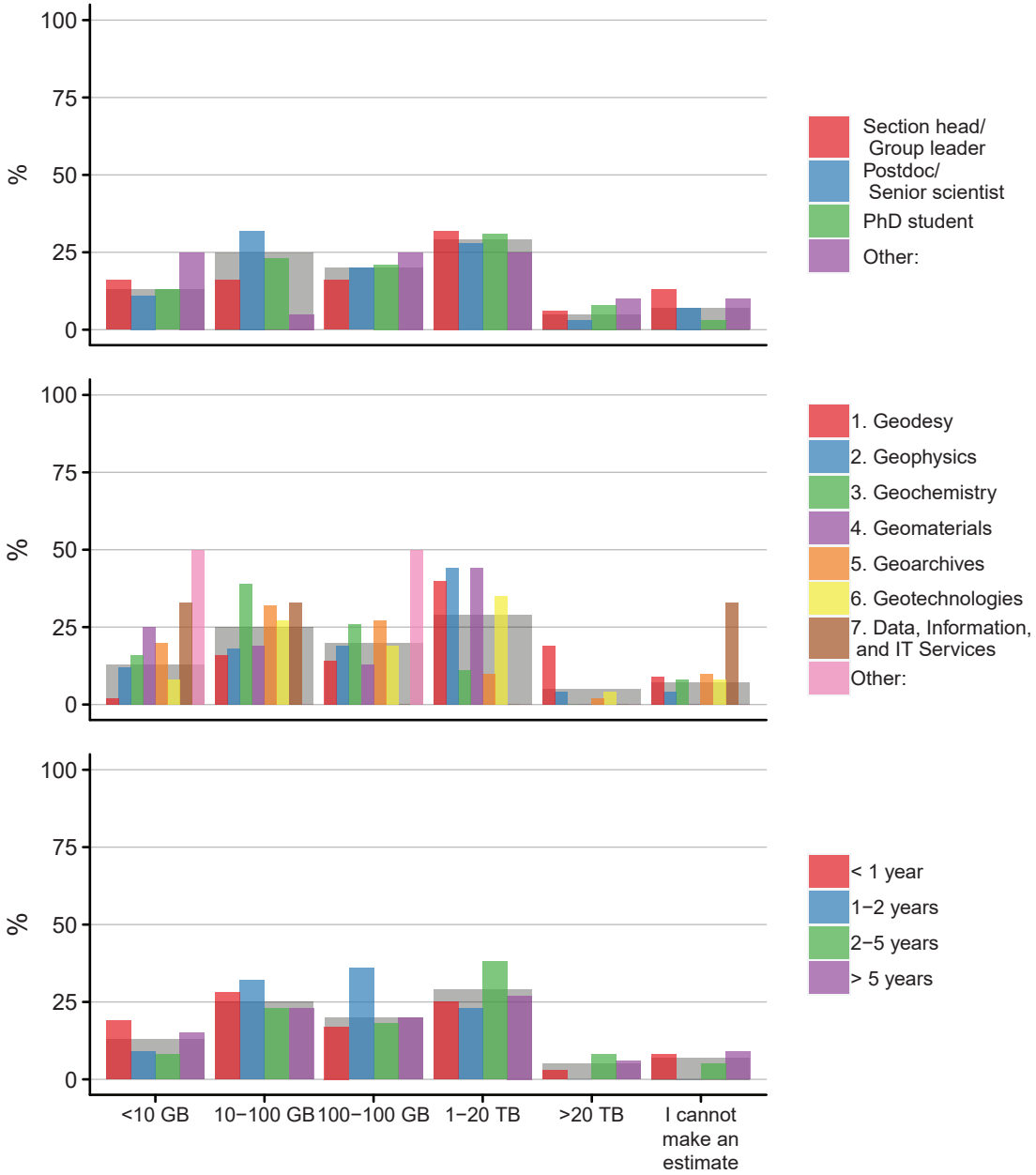
16. Which metadata scheme or standard do you use to capture metadata? (n=171)



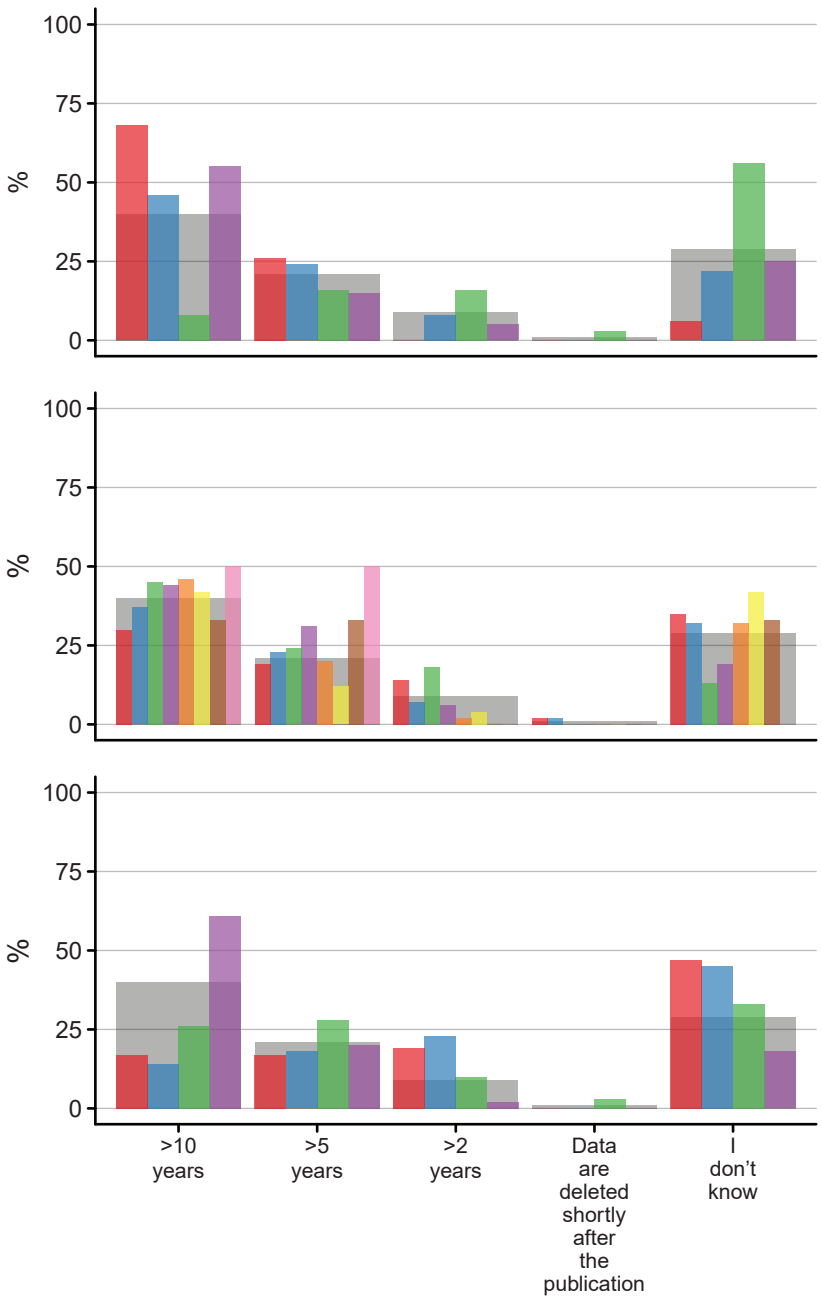
17. Where do you see the greatest need for support in your research data management process? (n=171)



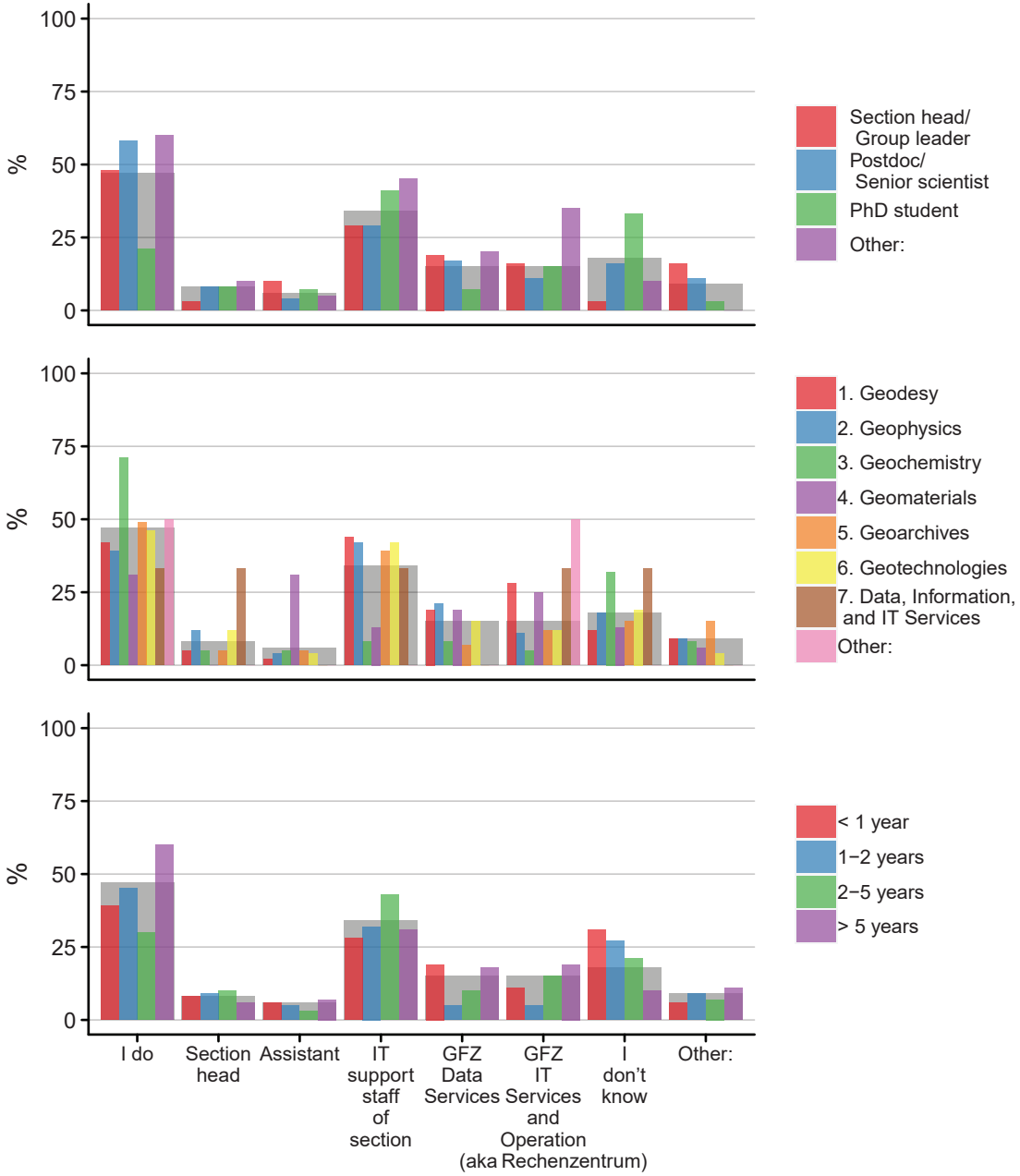
18. How much data did you generate in the last five years, excluding backups? (n=226)



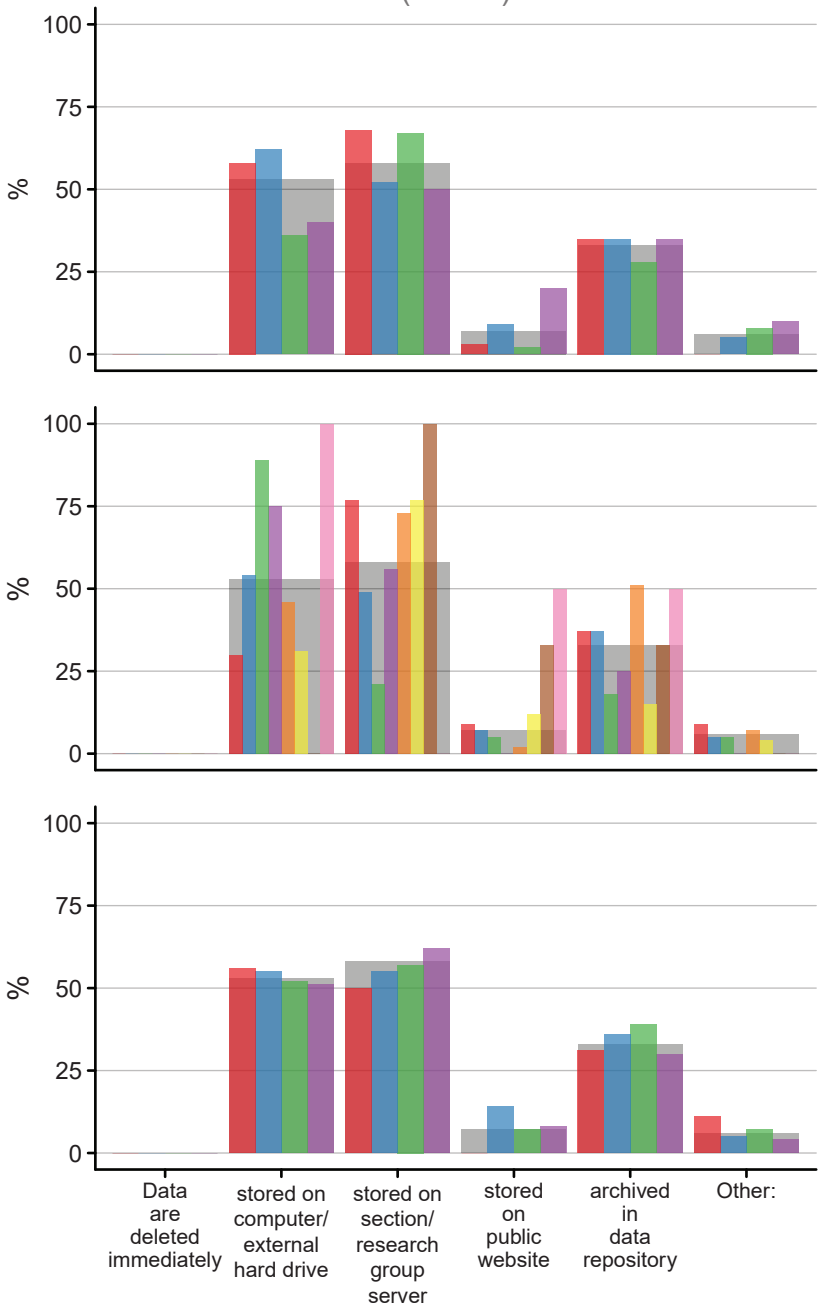
19. How long are research data that were the basis for your scientific publications securely stored? (n=226)



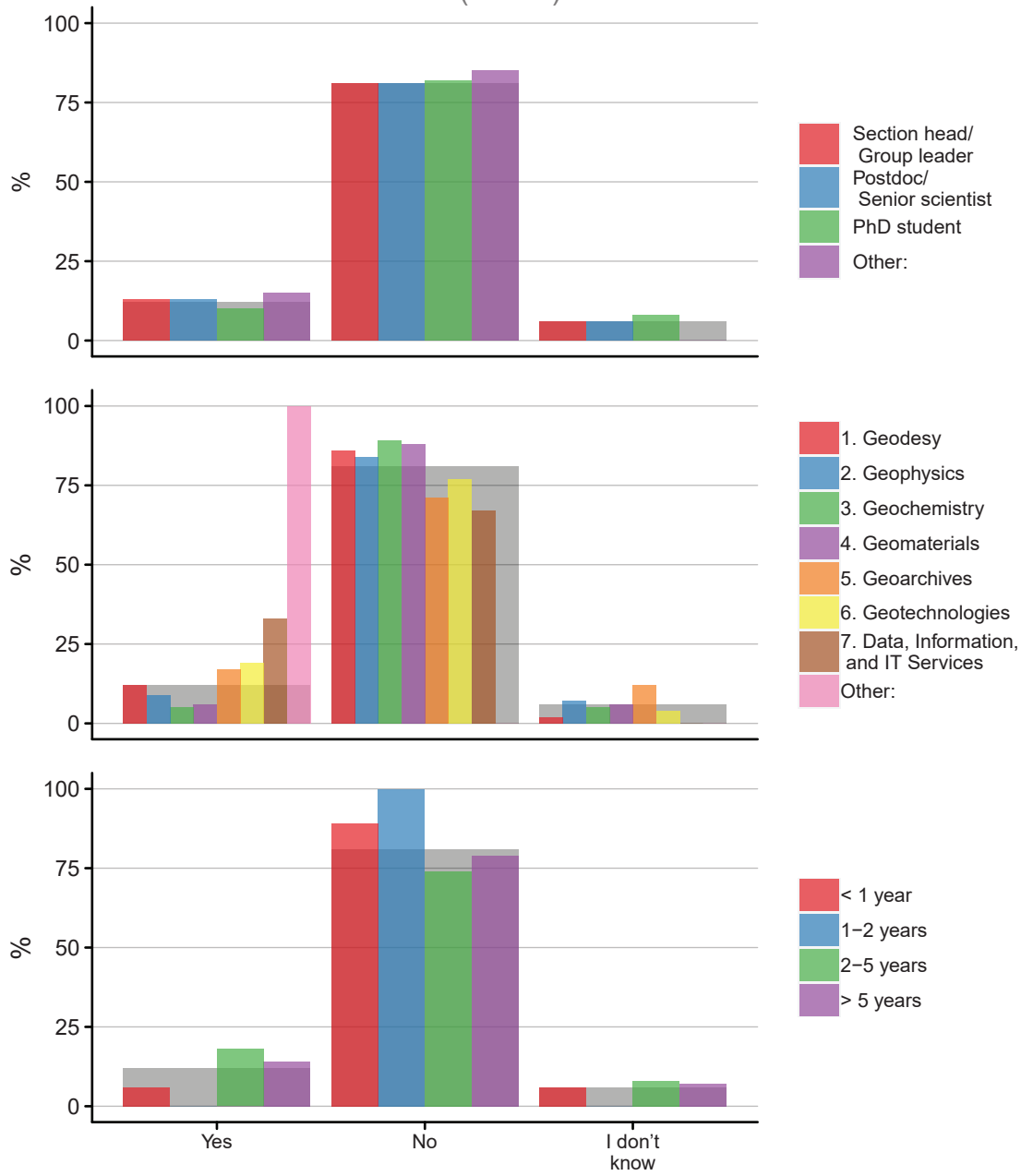
20. Who takes care of long−term research data archiving in your group? (n=226)



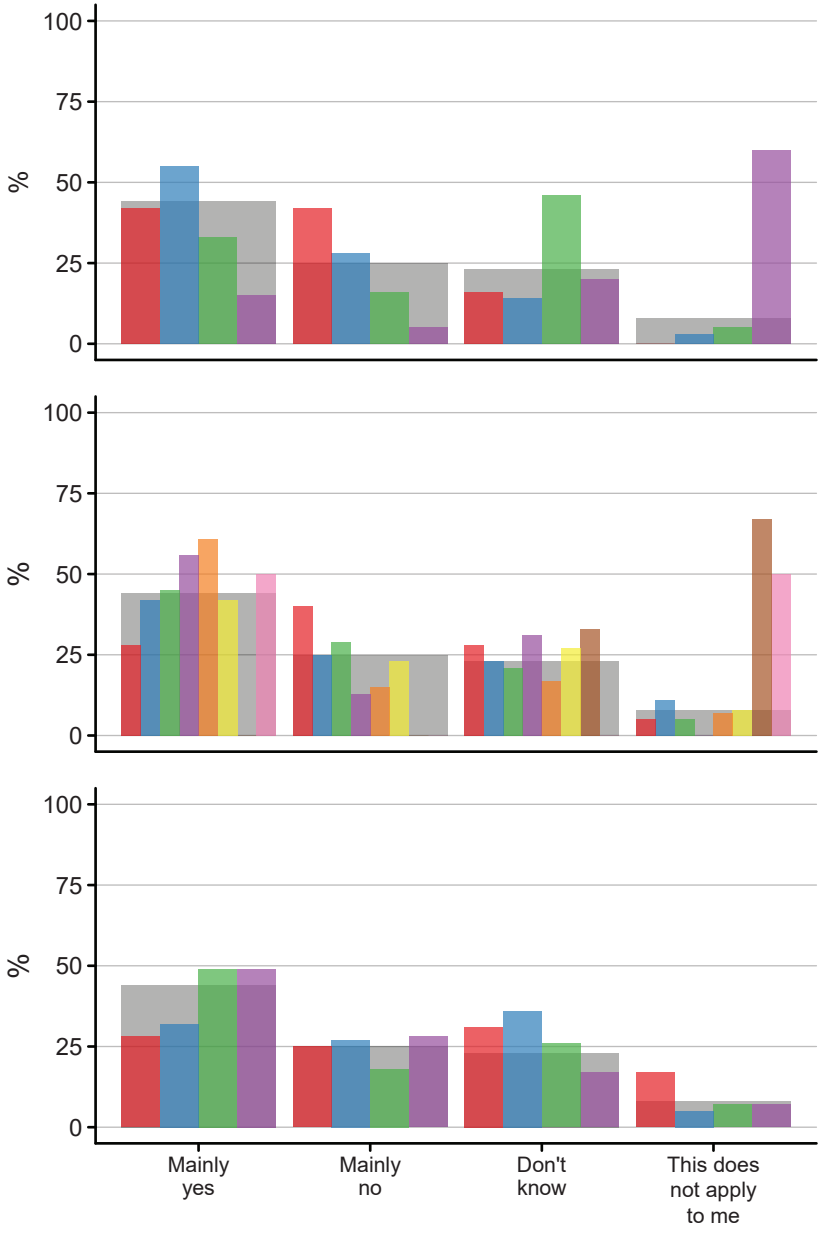
21. What happens to data from completed projects? (n=226)



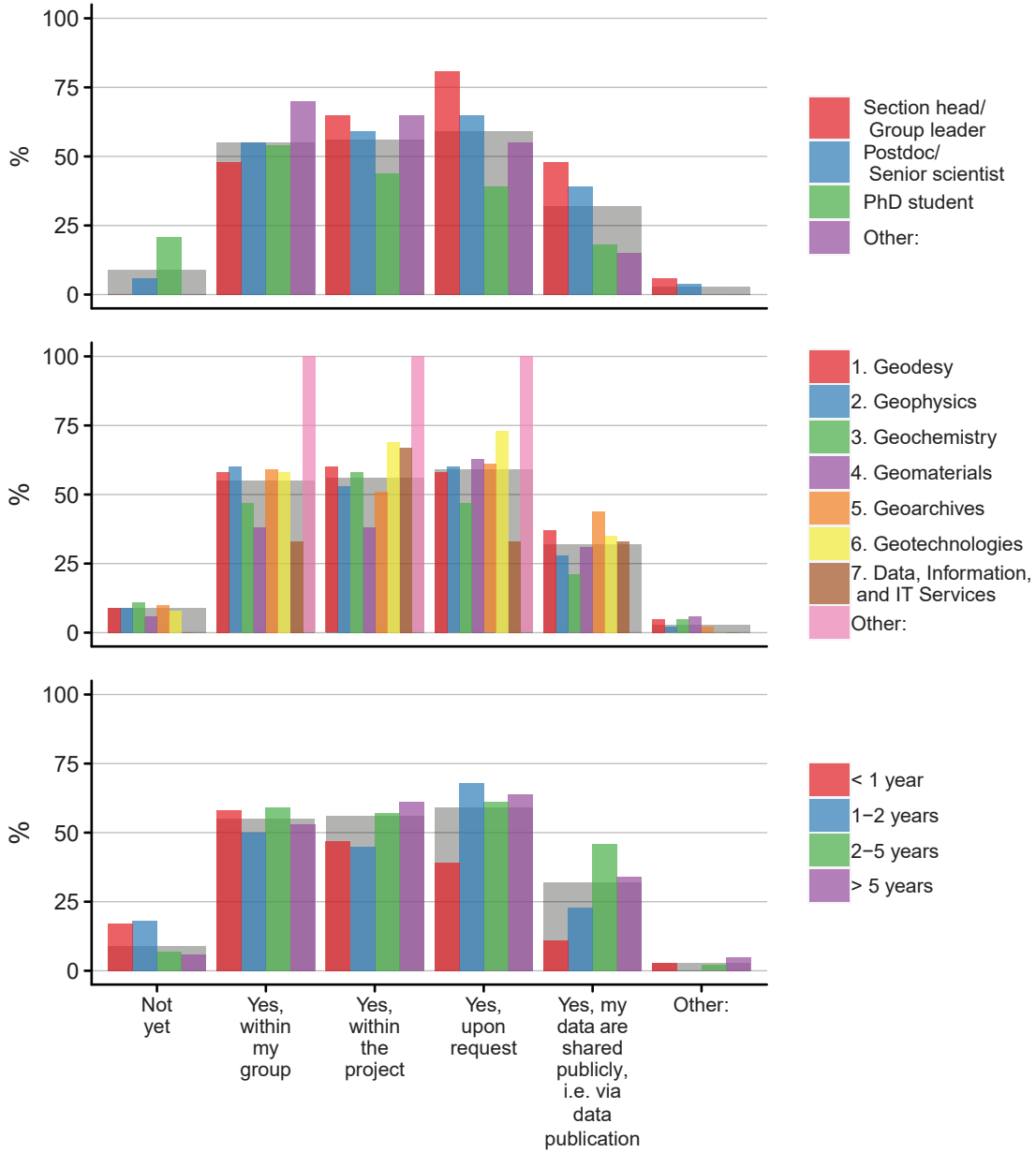
22. Do your research data contain sensitive personal information? (n=226)



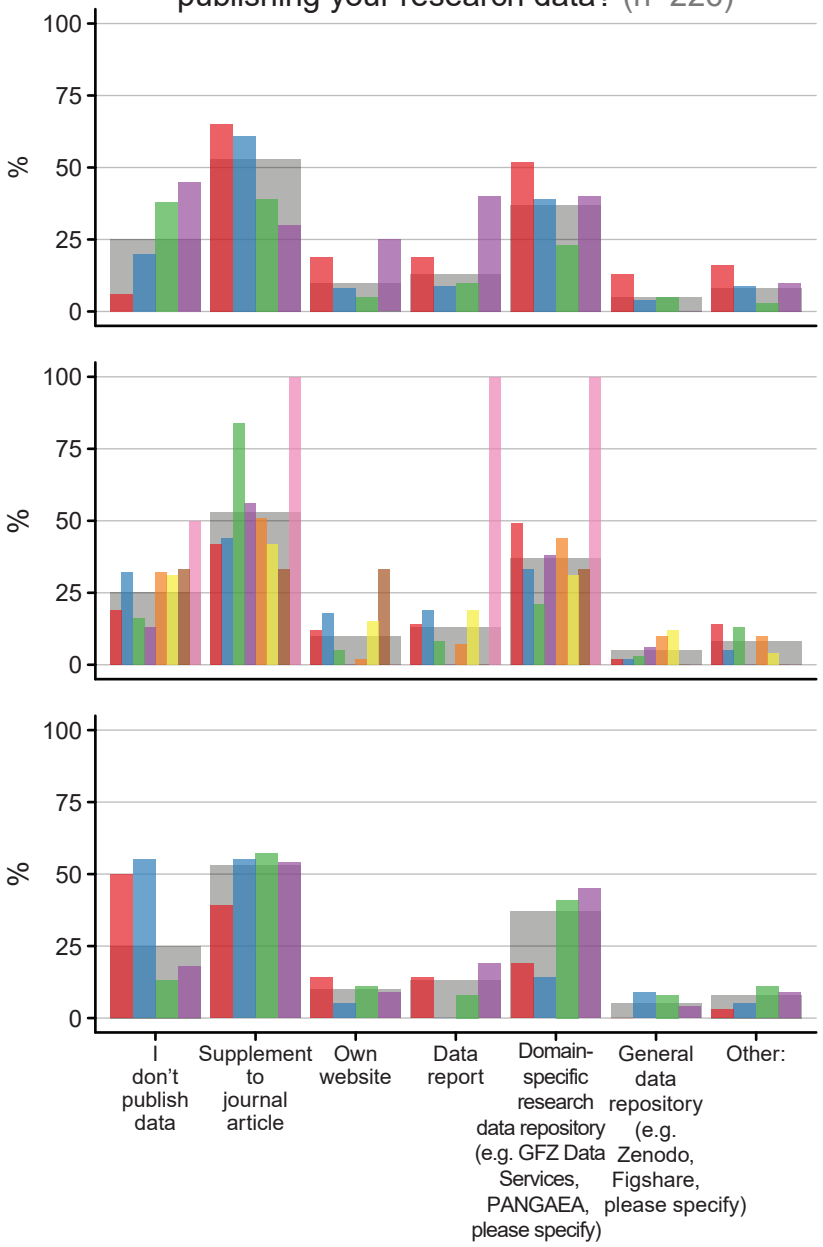
23, Do the journals in which you publish have data sharing/ accessibility requirements? (n=226)



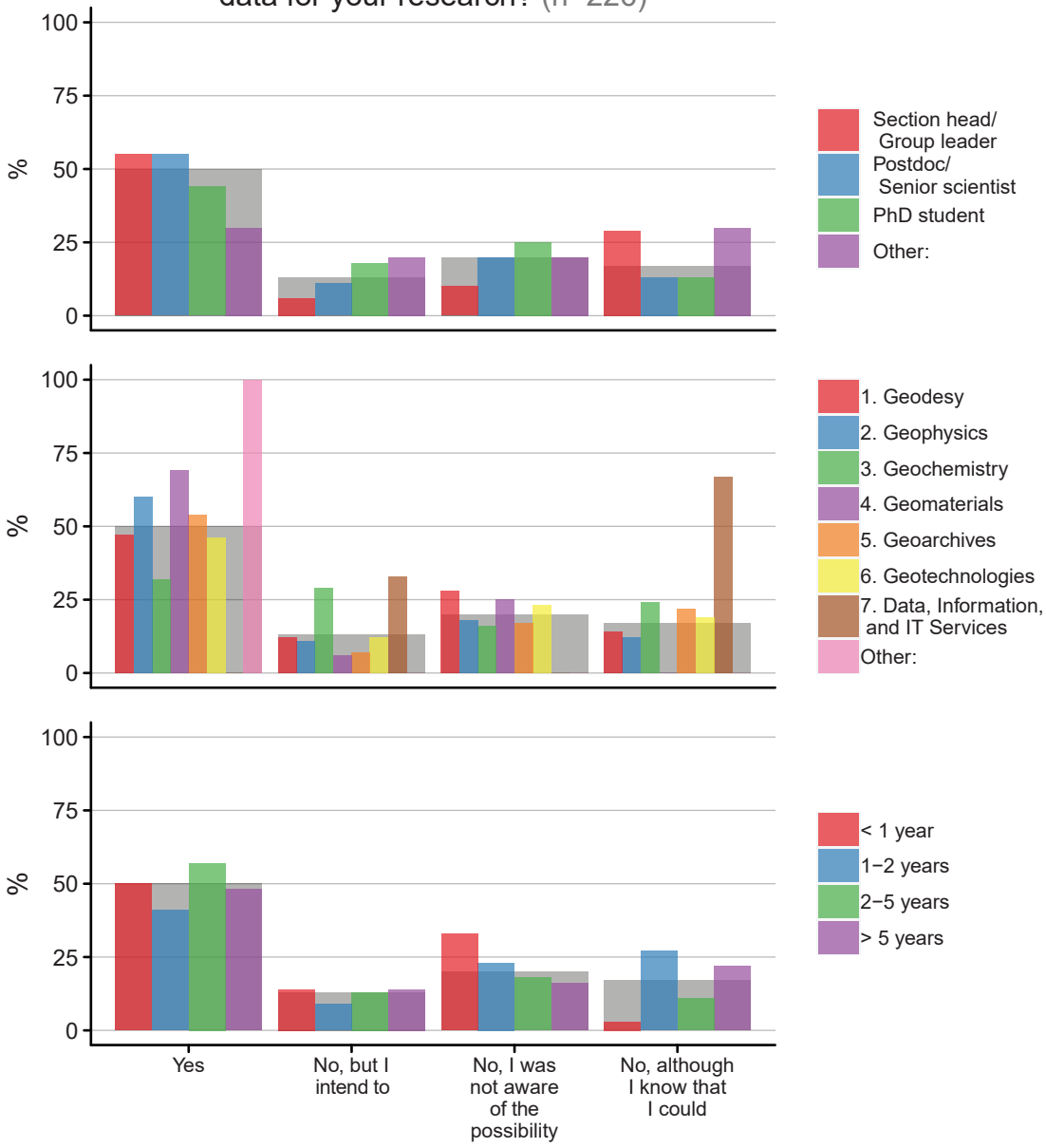
24. Do you share your data? (n=226)



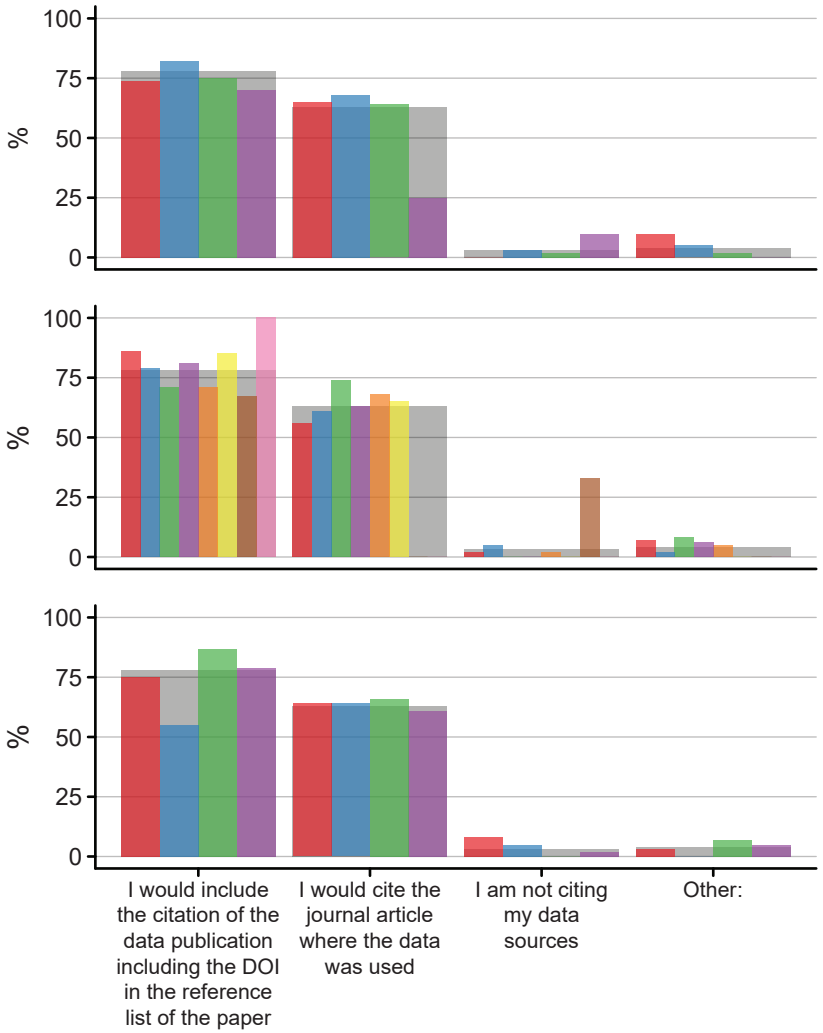
25. What is your preferred platform for publishing your research data? (n=226)



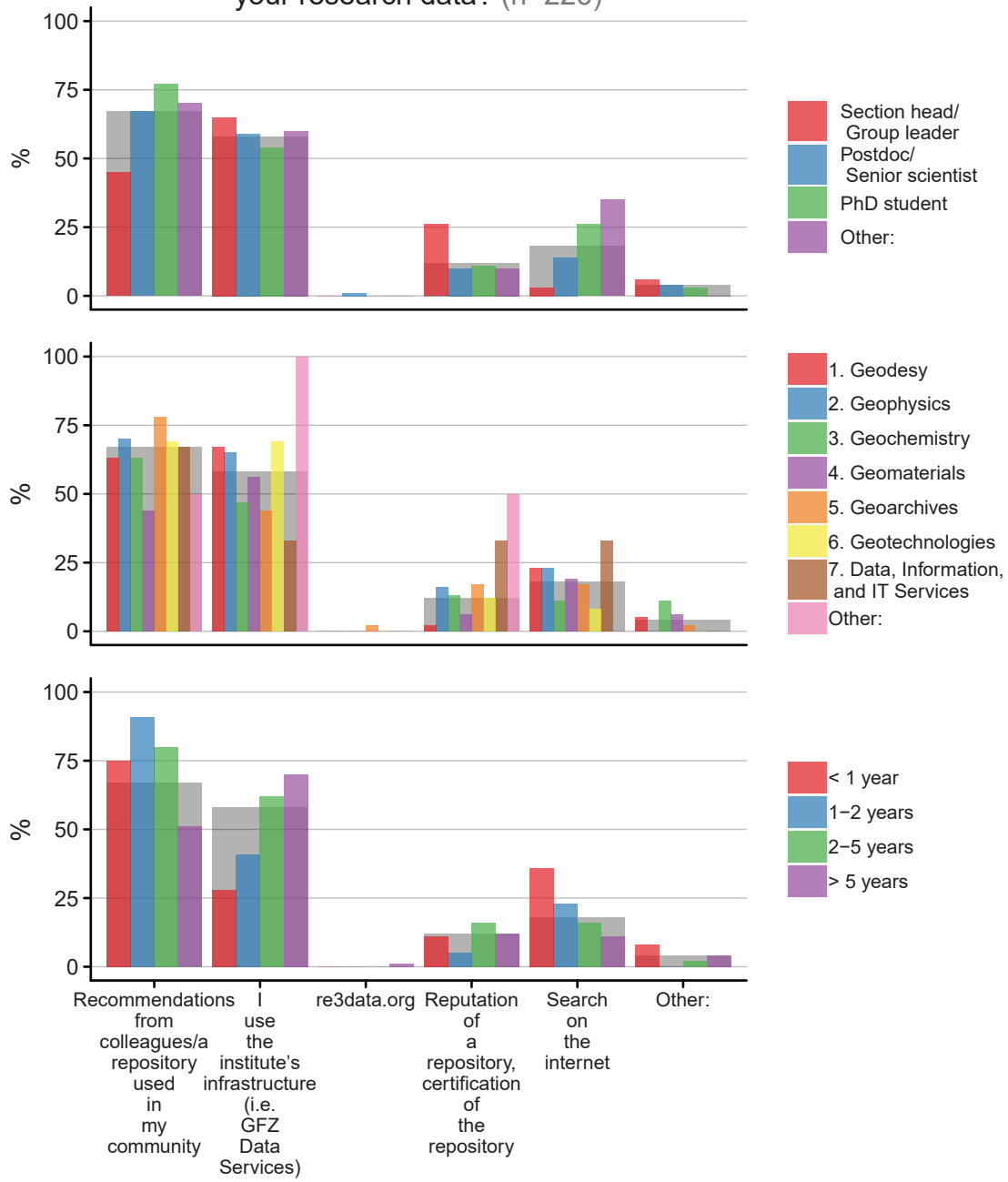
26. Have you ever used a data repository to obtain data for your research? (n=226)

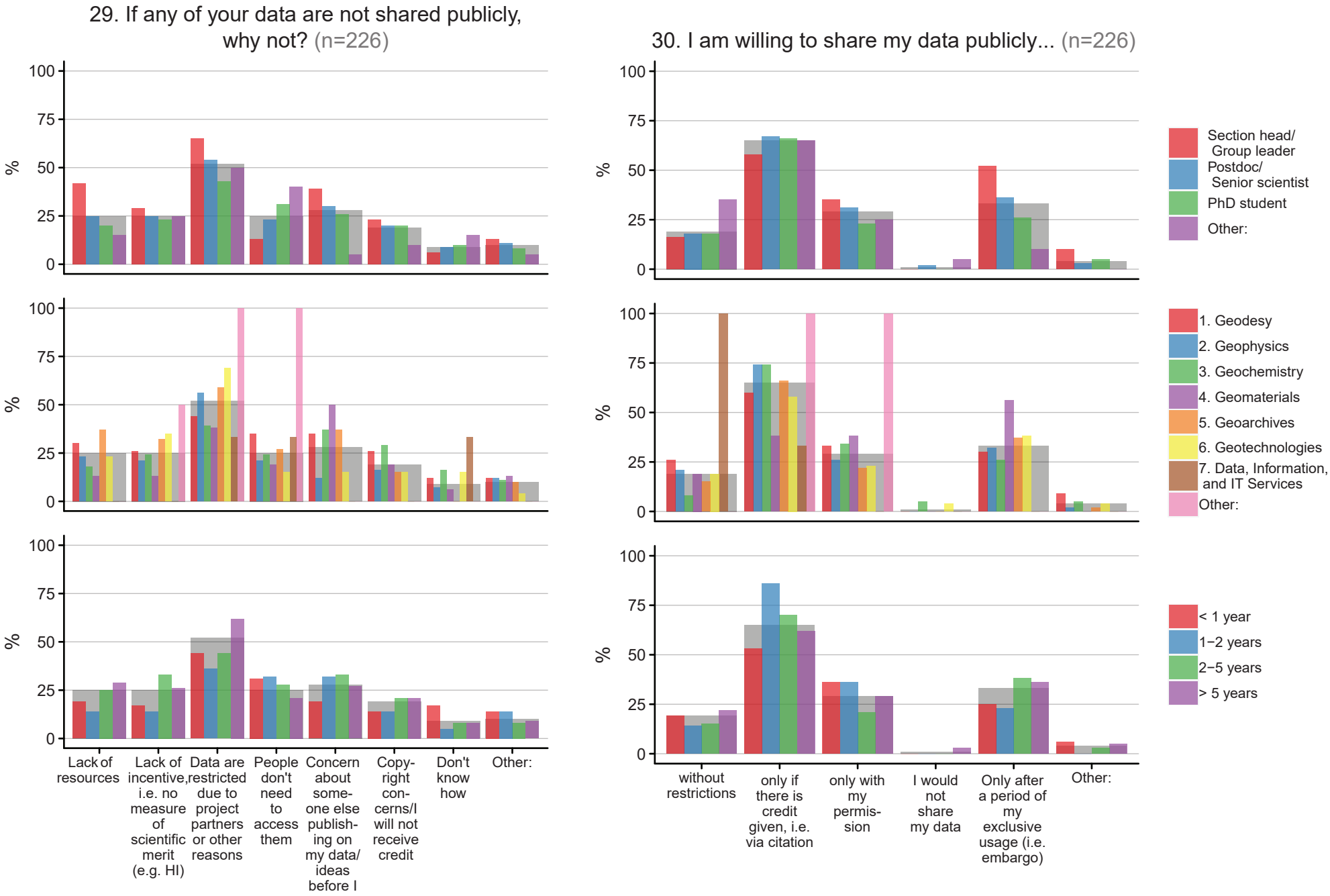


27. How would you cite data from other researchers if you are using them in a paper? (n=226)

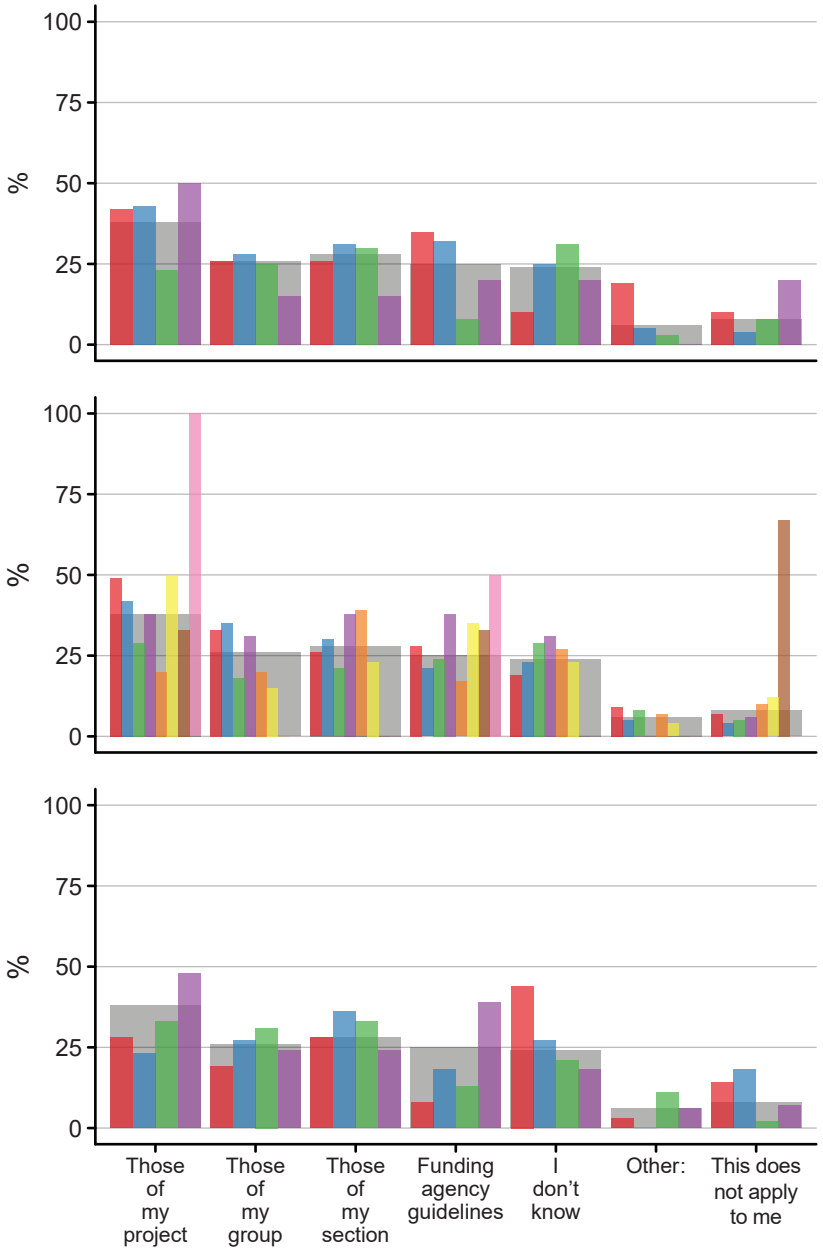


28. How would you select a suitable repository for your research data? (n=226)

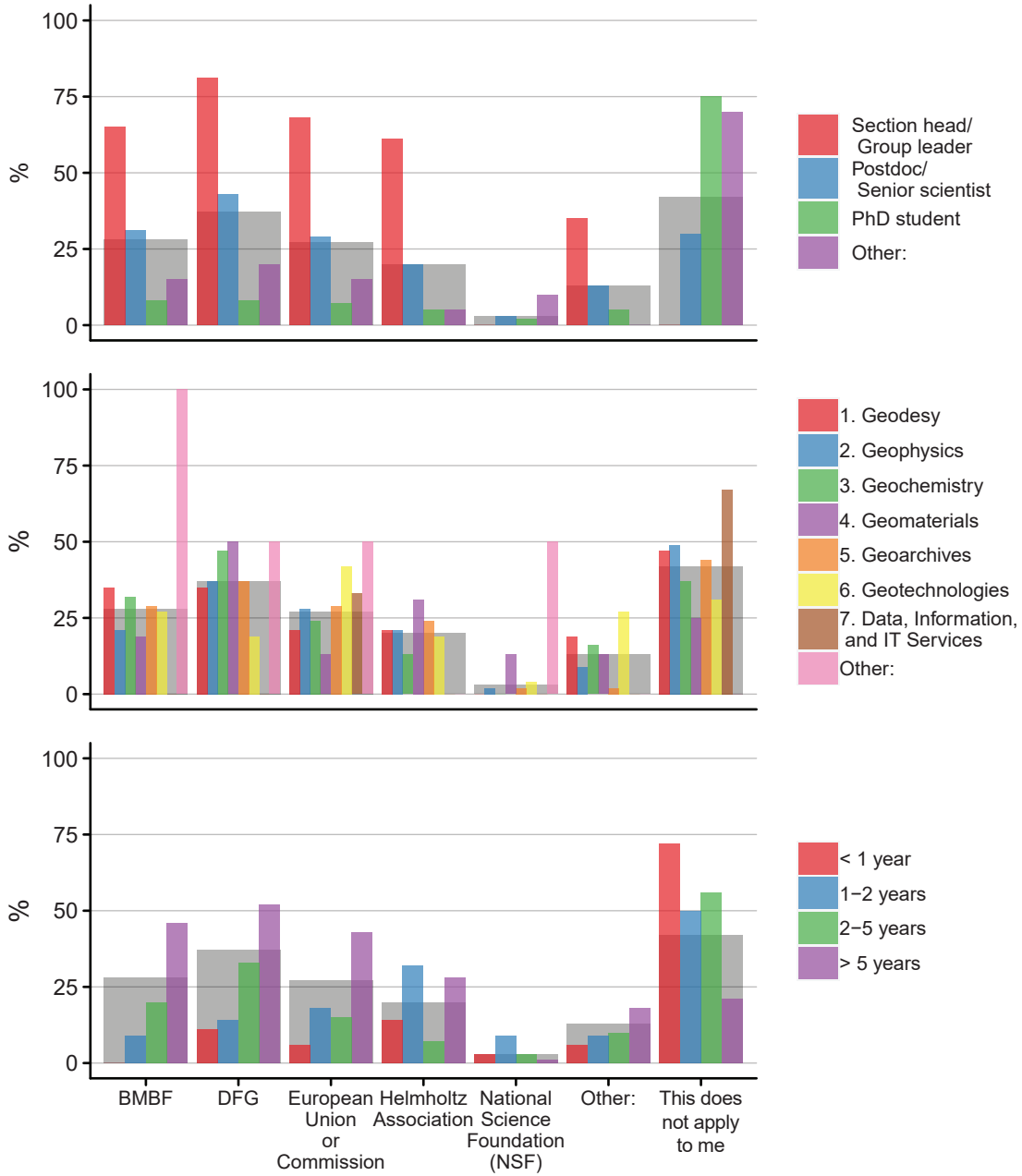



29. If any of your data are not shared publicly, why not? (n=226)
30. I am willing to share my data publicly... (n=226)
%
100
75
50
25
0
Lack of resources
Lack of incentive, i.e. no measure of scientific merit (e.g. HI)
Data are restricted due to project partners or other reasons
People don't need to access them
Concern about some-one else publish-ing on my data/ ideas before I
Copy-right con-cerns/I will not receive credit
Don't know how
Other:
without restrictions
only if there is credit given, i.e. via citation
only with my permis-sion
I would not share my data
Only after a period of my exclusive usage (i.e. embargo)
Other:
Section head/ Group leader
Postdoc/ Senior scientist
PhD student
Other:
1. Geodesy
2. Geophysics
3. Geochemistry
4. Geomaterials
5. Geoarchives
6. Geotechnologies
7. Data, Information, and IT Services
Other:
< 1 year
1−2 years
2−5 years
> 5 years

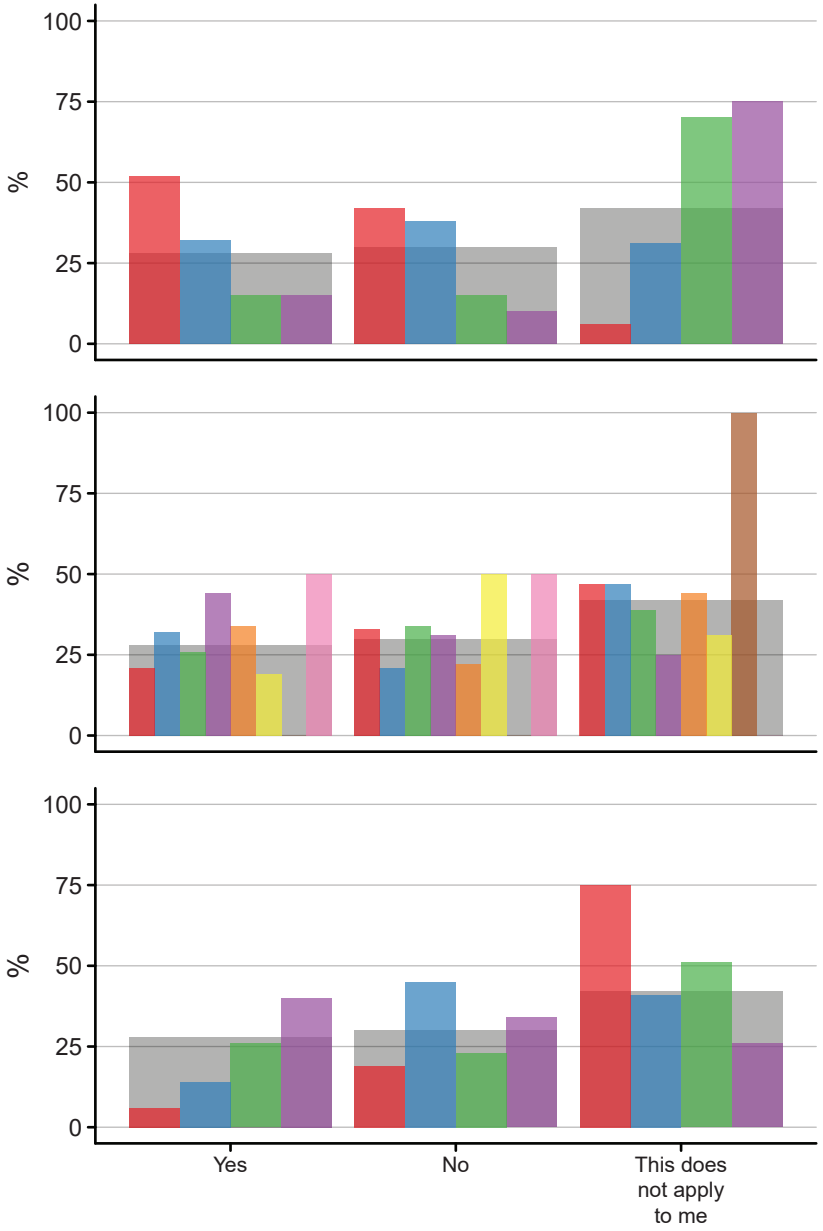
31. Are there formal guidelines, procedures or workflows that specify your handling of data? (n=226)



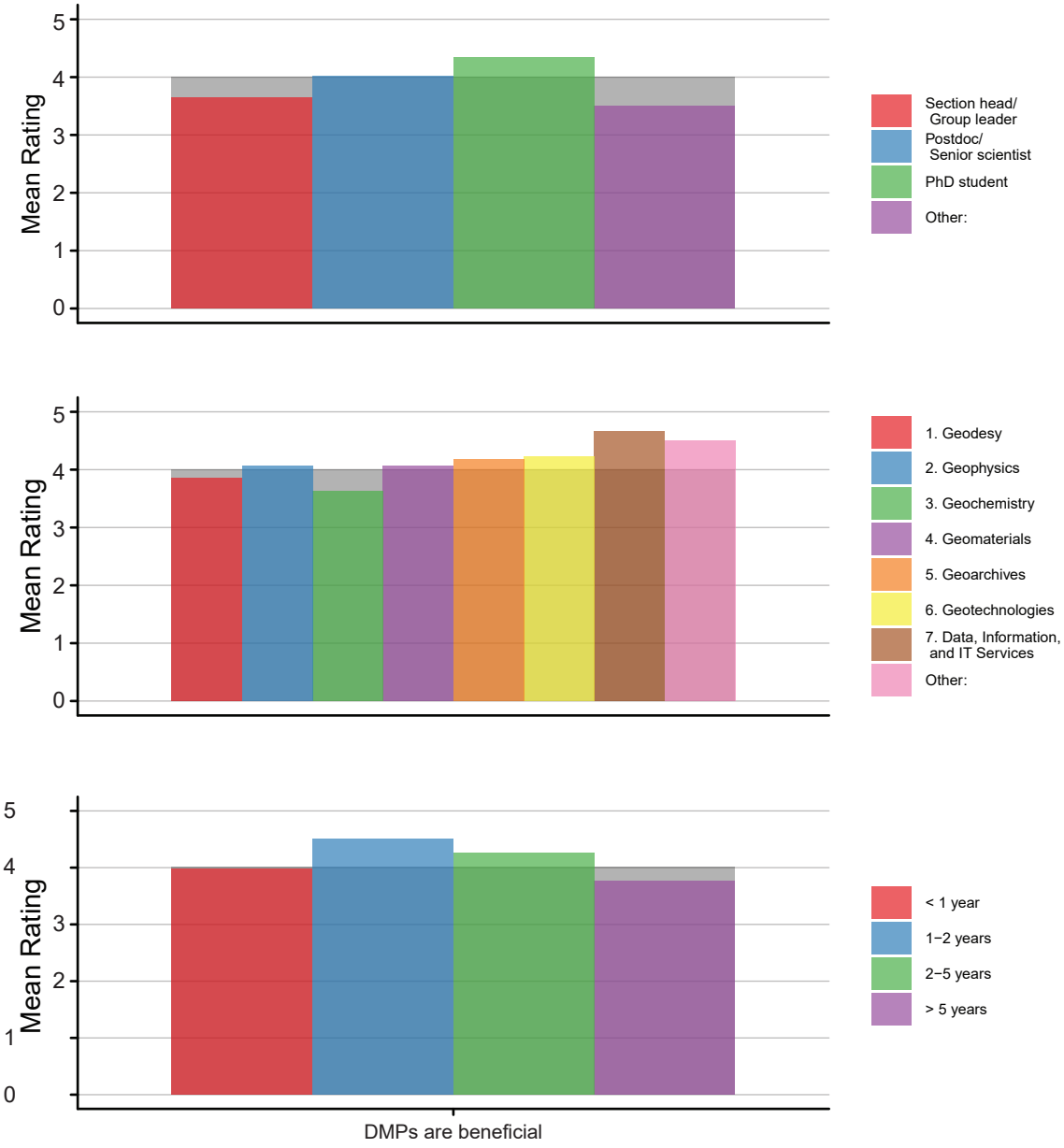
32. To which of the following funding agencies have you applied for project grants in the past five years? (n=226)



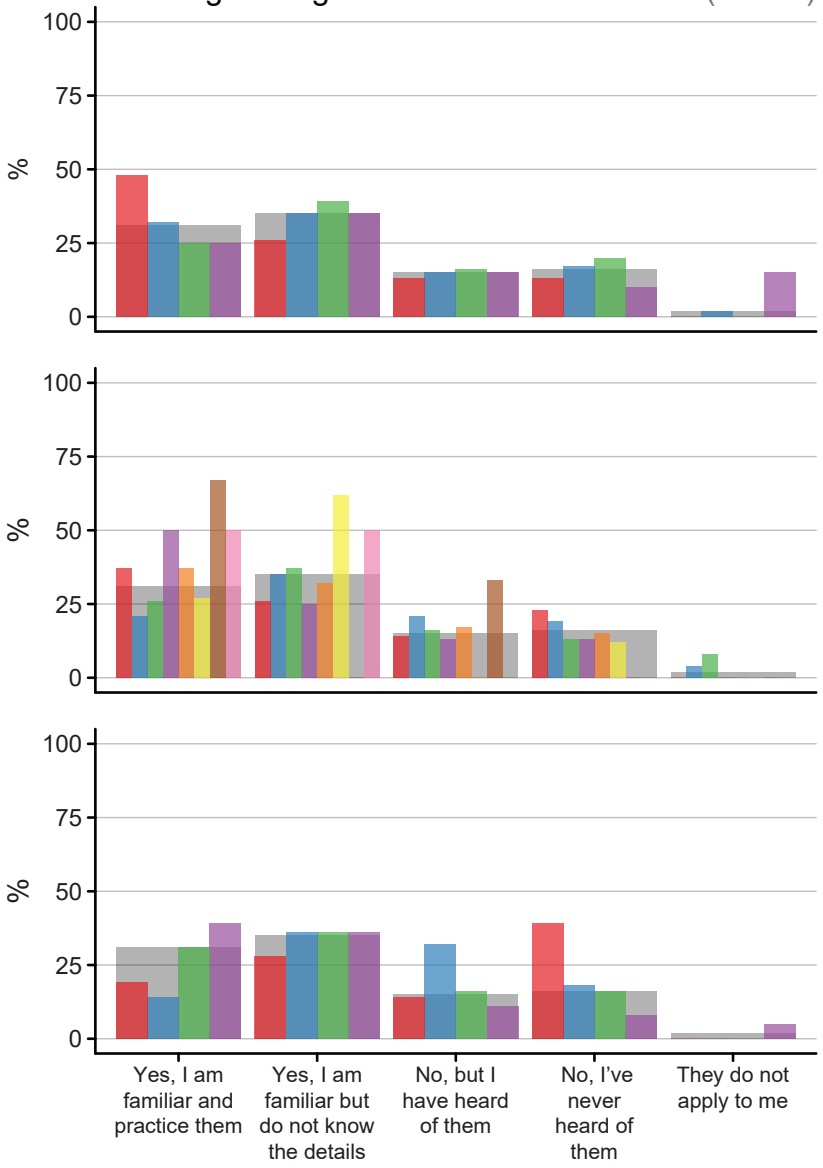
33. Did the funding agency/agencies require a data management plan for your project? (n=226)



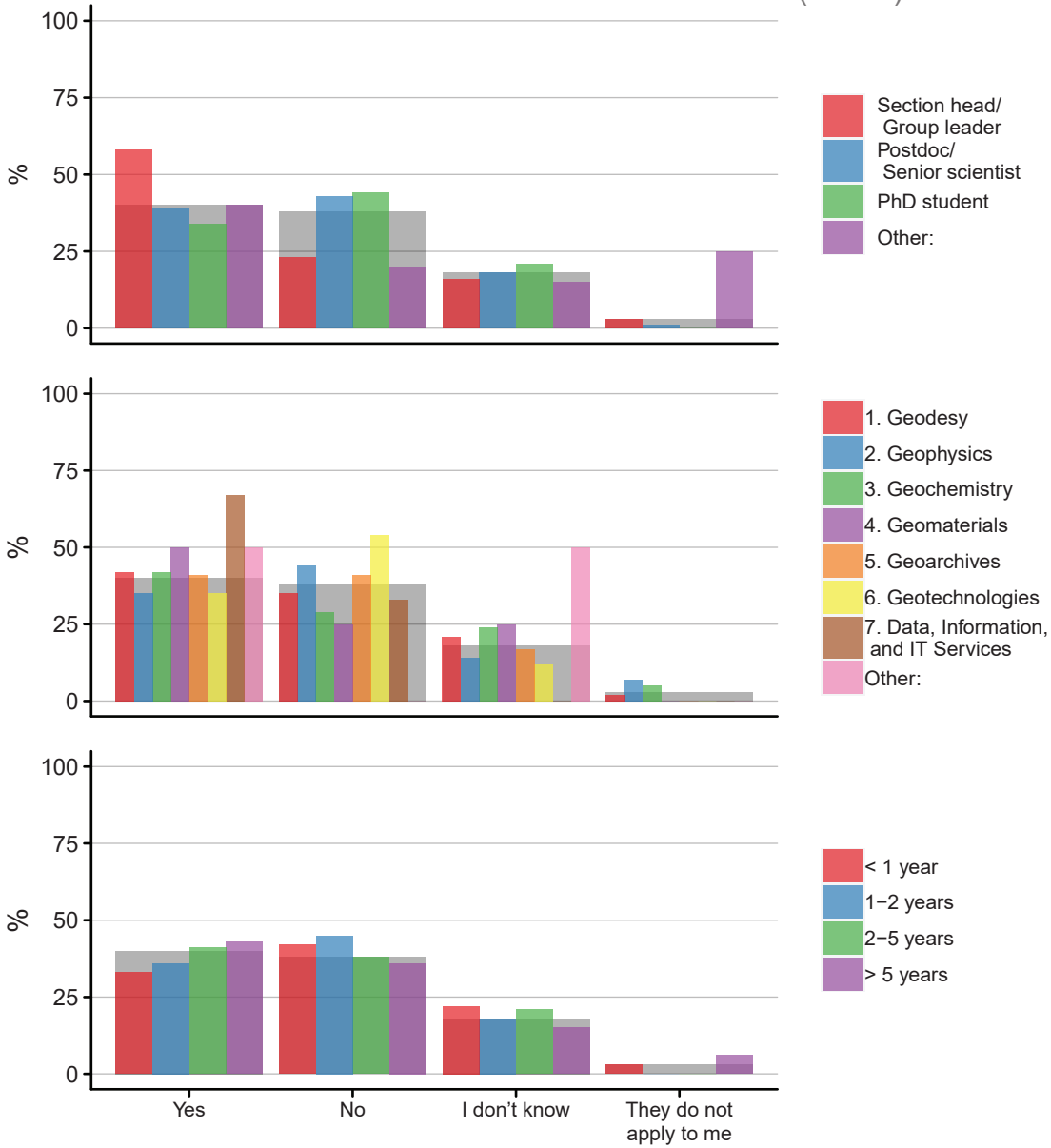
34. Please rank from low to high (1-5) the benefit of creating and implementing data management plans for all research projects (e.g. PhD project) or laboratories... (n=226)



35. Are you familiar with the GFZ−adopted Guidelines for Safeguarding Good Scientific Practice? (n=226)



36. Are you familiar with the Guidelines on Research Data at the GFZ German Research Centre for Geosciences? (n=226)



ISSN 2190-7110