Bibliothèque de l'EPFL

EPFL Library Research Data Team EPFL

October 2022

Quantitative assessment of research data management practice – Report 2021

By Francesco Varrato, Chiara Gabella, and Eliane Blumer

Contents

Background information

This work is a continuation of the Data curation services project, which aims to evaluate the appropriateness and feasibility of setting up data curation / data stewardship services for EPFL researchers. While the objective of this survey was and is to collect information on researchers' habits in terms of managing their research data, as well as to identify their needs for data curation services/support, for this edition of the survey a particular focus has been given to the ways in which they disseminate data and code. The previous two editions of such a survey were carried out in 2017 and 2019, in collaboration with TU Delft, Cambridge University and Illinois University: the two years regularity seems to be the best fit to keep track of trends as well short-term evolution of RDM practices, and a good timeframe to reduce the burden on both the researchers and the Library in deploying the survey and integrate their results into specific academic services.

Methodology

Realization of the questionnaire

In its past two editions, the survey consisted of a common part with partner universities, plus a specific part for EPFL, as the latter contained questions on data curation service needs. The survey, internally adjusted and validated (Research Data Library team and SPI team of EPFL Library) for its 2021 edition, has seen a slight departure: while the two parts have been retained and expanded, their changes from previous editions have not been reviewed by the previous partner universities. This is due to two main factors: on one side, the partner universities didn't carry out the 2021 survey, and on the other side, EPFL Library has been working in much stricter contact with other EPFL services and used this survey to align the efforts.

While in the 2017 edition Google Form was used as the survey tool, for the 2019 and 2021 editions SurveyHero has been chosen, for its advanced features as well as for its respect for privacy laws (SurveyHero's company is based in Switzerland and complies with GDPR).

In the 2021 survey, all questions required a response, except for the ones implying free text writing and the ones about participants' personal information (affiliation and contact). Moreover, thanks to the skip logic of certain questions, not all questions were to be answered by all participants. These methodological choices reduce the overall completion time for some participants, thus trying to improve their retention.

Survey launch, follow-up, and closing

The survey has been named "Quantitative assessment of RDM practices 2021", and its collector "Assessment of RDM practices 2021". The communication about the survey was done on 9 June 2021via email (personnel-scientifique.epfl@epfl.ch; mer.epfl@epfl.ch; professeurs.epfl@epfl.ch; doctorants.epfl@epfl.ch) by the Research Data Library team; it has also been posted on the Library's webpage, but no information has been communicated on Facebook, Linkedin, or Tweeter. Information was also relayed by the Research Data Library team to the EPFL Data Champions community, while it was arguably not relayed by liaison librarians (with a few exceptions) to their faculties of reference. A follow-up email and was sent out three weeks after the launch, on June 30, 2021. The survey has been closed 4 weeks after launch, on July 10, 2021.

Summary of the collected information

1. Participation

For the 2021 survey and based on the EPFL statistics[1] for the segments of the **population targeted** by the emails, about **4'000 people** (as FTE[2]) have supposedly received the email communicating the survey. Previous surveys' reports estimated a much larger number of receivers, estimated as more than 6'000, i.e. the entire EPFL staff.

The surveying platform provided a **Participation Rate of 62.0%** (= People who have accessed the survey and started answering), up from 40.3% in 2019 (+49%). We thus estimate that about 403 people have accessed the survey link (= Total Responses / Participation Rate), down -43% from the 2019 estimation of 707 people[3]. This implies that the 403/4000 = ~10% of potential respondents is at least curious or interested enough in the subject of RDM: it's impossible to say whether the remaining ~90% has been poorly targeted or is somehow not interested.

We received **250 Total Responses**, a response rate 6.25% (= People who started answering the survey divided by the total targeted population), i.e. a representative sample of the targeted population with a confidence level of 90%**.** The number of total responses was 237 in 2017 and 285 in 2019, which implies an increase with respect to the first edition (+5%) and a decrease compared to the second one (-12%).

[1] A total of about 4'000 people: ~3'611 corps intérmediaire (PT, MER, post-docs, PhDs students, …) and ~350 professors. Data source: www.epfl.ch/about/overview

[2] The FTE measure is more precise as better documented, and the number of persons corresponding to the FTEs is statistically almost 1:1 (6'369:5'925 to be precise). Data source: /www.epfl.ch/about/overview/fr/statistiques-institutionnelles/statistiques-personnel

[3] In turn, this implies a rate of general interest of 10% (= People who accessed the survey link / Population targeted = 707 / 6'060), instead of the 12% estimated for 6'060 figure of Population targeted in 2019.

As per all previous surveys and for surveys in general, not everyone completed the survey in each question. A **Completion Rate of 81.6%** has been measured (= People who have participated and completed the survey), +6% up from 76.8% in 2019.

With an Average Completion Time (Trimmed[4]) of **08:44 min**, we observe a +54% increase in the time dedicated to fill the survey by the respondents, up from 05:40 min in 2019.

As for the respondents themselves, we extract some insights from questions 15. to 17.: these are not mandatory questions, so not all respondents provided an answer. In the following table one can observe that the are no big changes in distribution of responses by faculty:

	2021	2019	2017
SB - Basic Sciences	67	65	66
STI - Engineering	60	62	76
SV - Life Sciences	31	22	35
ENAC - Arch., Civil and Envir. Engineering	25	35	33
IC - Computer and Communication Sciences	10	16	19
CDH - College of Humanities	6	4	4
CDM - College of Management	-	5	-
Other EPFL affiliation	6	5	2
(NOT at EPFL)	-	5	1
EPFL Middle East	1	-	-
Total	206	219	236

This year, we added a question about campus location, which tell us that about 10% of respondents is not located at the main EPFL campus of Lausanne:

Lausanne (main campus)	181	89.16%
EPFL Neuchâtel	10	4.93%
EPFL Geneva	8	3.94%
EPFL Fribourg	1	0.49%
EPFL Valais Wallis	9	4.43%
Other ...	5	2.46%

Of the 250 respondents, we observe that a great variety of roles:

	2021	**2019**	**2017**
Postdoc / Scientific collaborator	82	66	107
PhD student	76	109	22
Full Professor / Titular professor / Emeritus	20	10	20
Tenure Track Assistant Professor	6	5	15
Technical or Scientific staff	7	11	22
Administrative staff	3	-	-
Permanent res. / Senior res. / Research assistant	2	1	-
Associate Professor	1	6	14
Data manager / IT manager	1	-	-
Group leader	1	1	-
Head of technology platform	1	-	2

[4] See for instance: en.wikipedia.org/wiki/Truncated_mean

Other …	2	6	
Adjunct or MER Professor	-	4	29
Junior Professor	-	-	1
Ambizione Fellow	-	-	1
Total	**202**	**219**	**233**

If we perform a less granular categorization, we can distinguish 4 macro categories:

- PhD students
- Postdoc researchers (of any order)
- Professors (of any order)
- Others (admins, staff, etc.)

and therefore, we can transform the previous table in

	2021	**2019**	**2017**
PhD student	76	109	22
Postdoc / Scientific collaborator	82	66	107
Permanent res. / Senior res. / Research assistant	2	1	-
Ambizione Fellow	-	-	1
Full Professor / Titular professor / Emeritus	20	10	20
Tenure Track Assistant Professor	6	5	15
Associate Professor	1	6	14
Group leader	1	1	-
Adjunct or MER Professor	-	4	29
Junior Professor	-	-	1
Technical or Scientific staff	7	11	22
Data manager / IT manager	1	-	-
Administrative staff	3	-	-
Head of technology platform	1	-	2
Other …	2	6	
Total	**202**	**219**	**233**

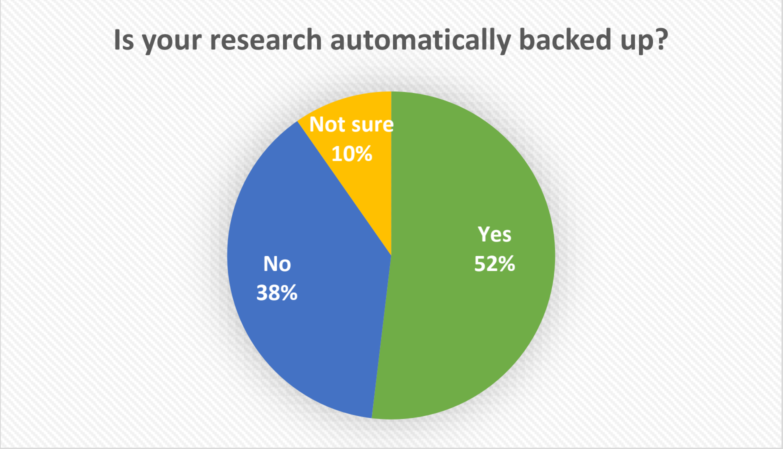
and in turn, we observe the trend of these macro categories:

	2021	**2019**	**2017**
PhD student	76	109	22
Postdoc researchers	84	67	108
Professors	28	26	79
Others (admins, staff, etc.)	7	11	22
Total	**202**	**219**	**233**

As a percentage of the actual population, the Professors represent a big chunk of the respondents when compared with their distribution in the targeted population: in fact, while 14% of the 2021 survey respondents can be classified as Professors, only ~9% of the targeted population is composed by Professors. Nonetheless, their weight is slightly higher than in 2019 (12%), but greatly reduced compared to the 2017 survey (34%): this reduction is indicative of Postdoc researchers and PhD students being more sensitive to the topics of RDM, a general trend observed also independently with more early researchers being involved in such topics (ex. participating to training or support requests to the Research Data Library team).

2. Research Data Management habits

Automatic data backup:

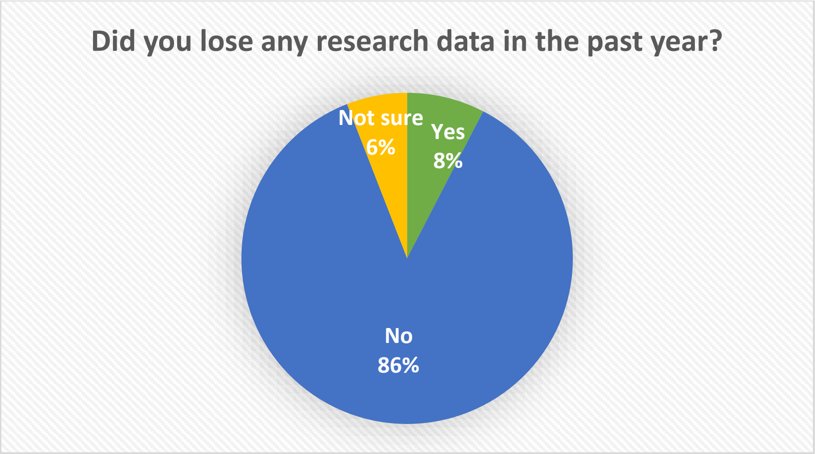


Participants stating that their research data is saved automatically were then asked to specify what processes were used. The responses were quite heterogeneous and included:

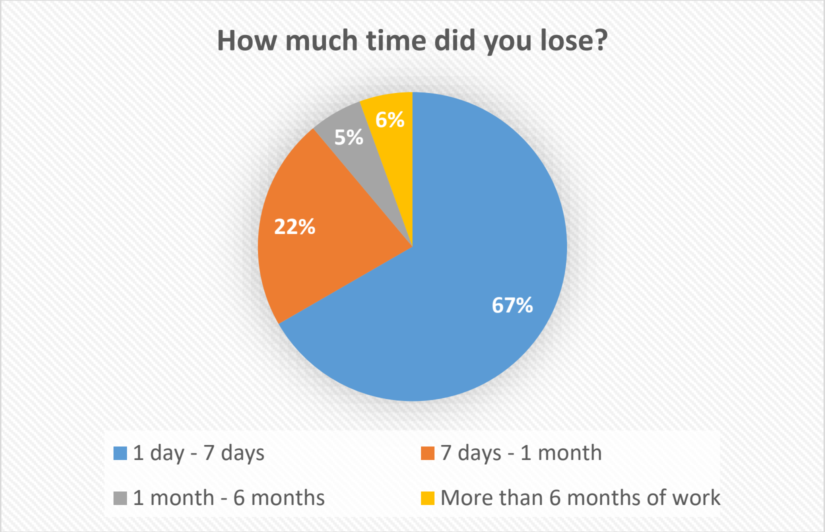
EPFL or lab server	38
external hard drive	12
Dropbox	11
Google drive	8
Git	8
OneDrive and others cloud solutions	6
Crash Plan (cloud solution)	4
Time machine (Mac)	4
Switchdrive	4
Vital-IT	2
SLIMS	1
own backup server	1

The relatively frequent use of DropBox, Google drive or other cloud solutions underlines a need to communicate on the security of the data "saved" in this way and their pros (e.g., convenience) and cons (e.g., access rights management).

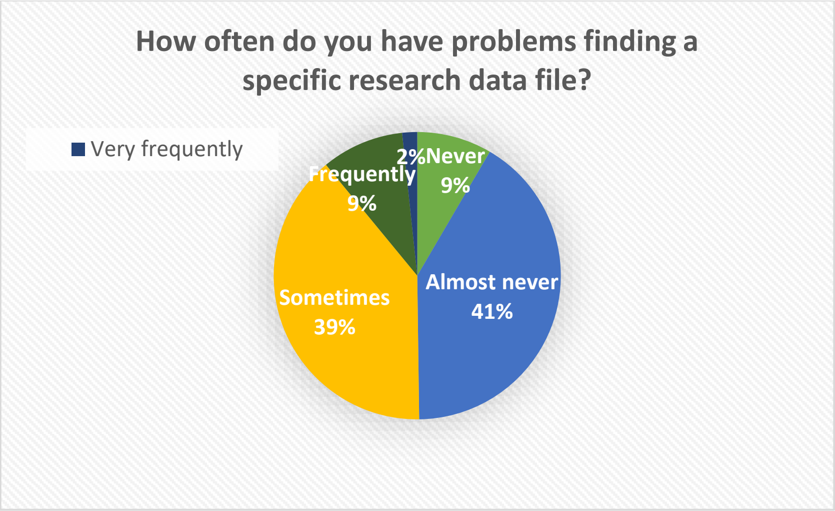
Loss of research data:



18 participants admit to having lost research data in the past year. For 12 of them, it took 1 to 7 days to either find or reproduce it. Four participants, meanwhile, lost 7 days to 1 month. One researcher lost 1 to 6 months, and another one lost more than 6 months.

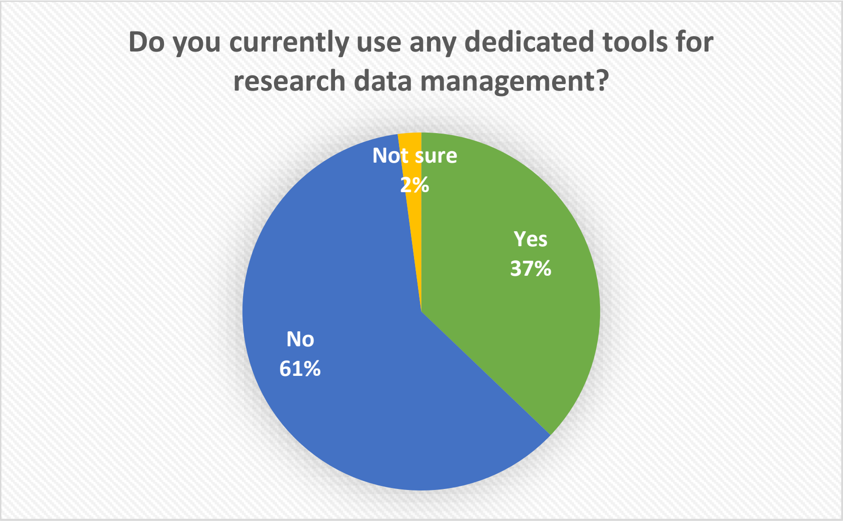


Difficulties in retrieving research data:



Half of the participants have difficulties in finding their data files: it seems more difficult for 39% (sometimes difficulties), 9% (frequently) and 2% (very frequently) of the cases. Perhaps this underlines a need to communicate and train the researchers on data organization and documentation.

Research data management tools used:



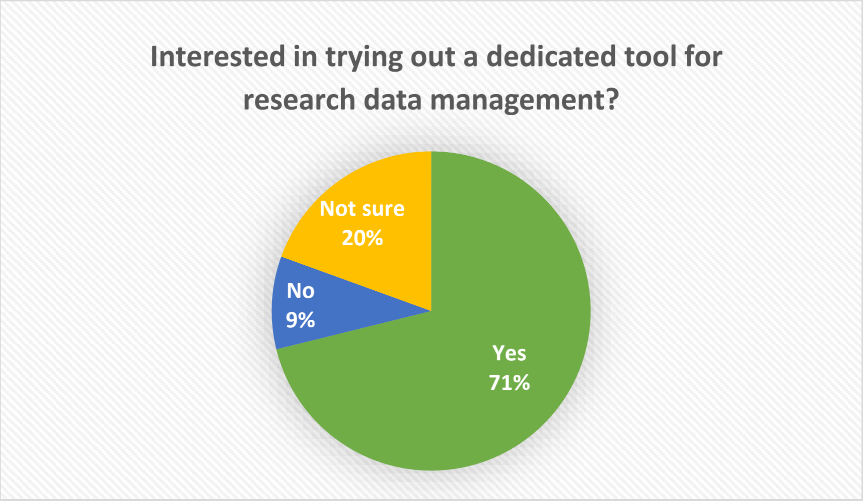
Among the tools used by the 88 participants who answered "yes" to the question are:

- Git, which is widely represented
- Electronic lab books (SLIMS, Benchling, E-Notebook, openBis, other ELNs)
- DropBox and Google Drive are also mentioned.

Git/GitHub/GitLab	55
SLIMS	9
DropBox	8
SVN	6
Google drive	6

c4science	4
MySQL	3
Evernote	3
Slack	3
AiiDA	2
Pyrat	2
E-Notebook	2
other ELN	2
openBis	1
Benchling	1
Jupyter	1
oneNote	1
Zotero	1

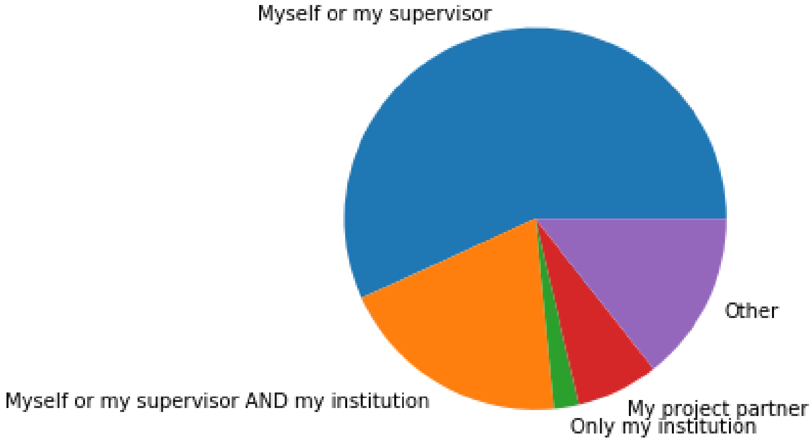
Among participants who do not use specific tools, 71% (106 people) would be interested in trying solutions to help them manage their research data:



Responsibility for managing your research data:

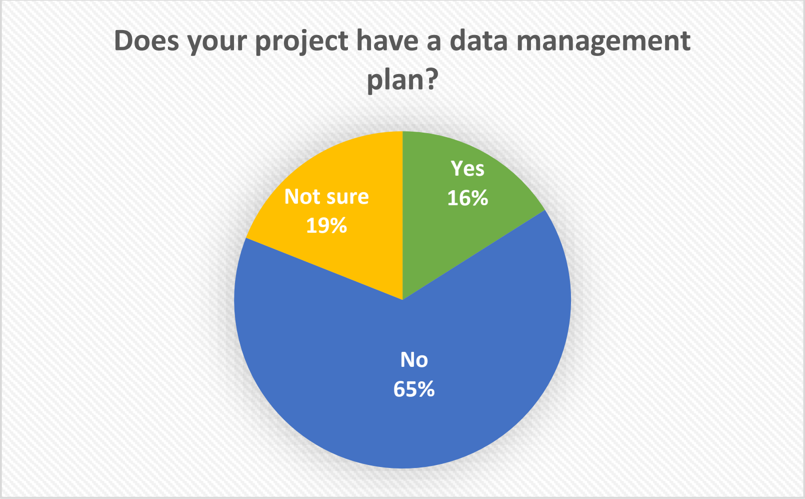
To the question "Who is responsible for managing your data?", a large majority of participants include themselves as the responsible persons, which is rather positive as personal involvement, but highlights the lack of professional dedicated roles (unless the respondents overlap with this role).
The category "Other" contains answers such as "My funding agency", "Not sure", or combinations that do not contain "Myself" or "My supervisor", answers that seem to illustrate a lack of understanding of responsibility for data management.

Who do you think is responsible for the stewardship of research data resulting from your project?



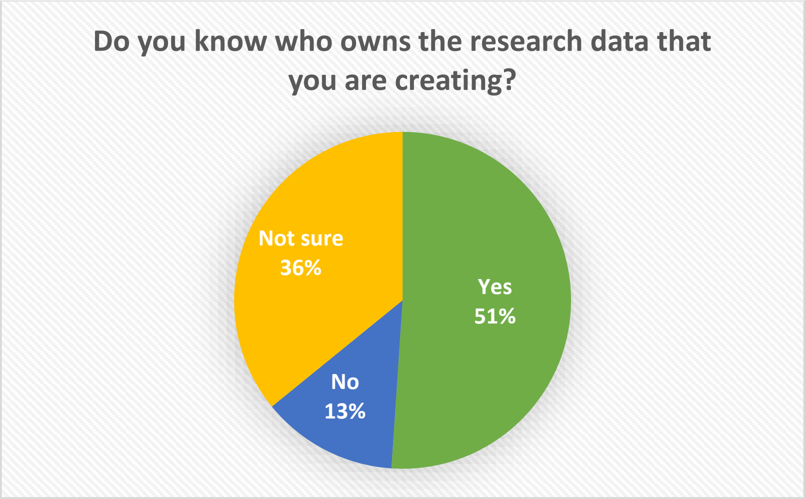
3. Knowledge of research data management

Data Management Plan



Only 16% of participants (38) have completed a DMP for their project, and 19% (45) do not know if they have.

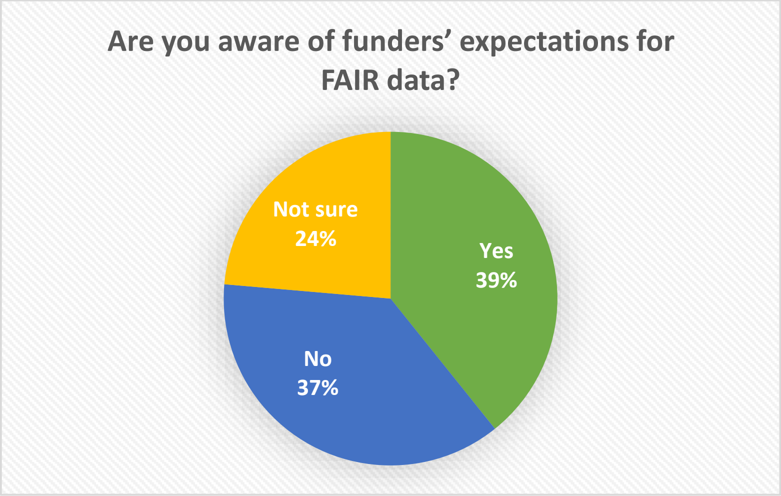
Ownership of research data



For this particular question, half of the participants does not know or is unsure about the ownership of the research data created. Regarding the other half, many get the answer wrong and, for example, believe they are the sole owners of their data, while in fact EPFL is the owner:

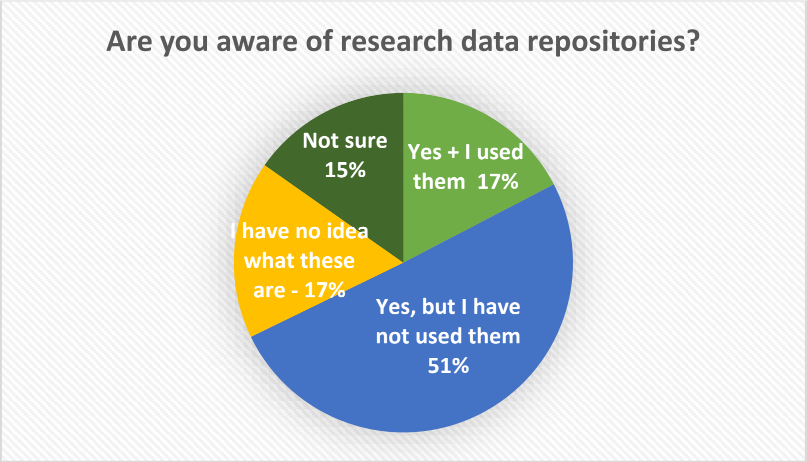
EPFL (or industrial partners)	56
PI/group/project	31
Myself	17
Funders	5
Public domain	5

Awareness of FAIR principles



About 60% of the participants are not aware of the FAIR principles. If we relate this figure to the 6530 scientists at EPFL, there are potentially 3,900 researchers to be trained or made aware on this subject.

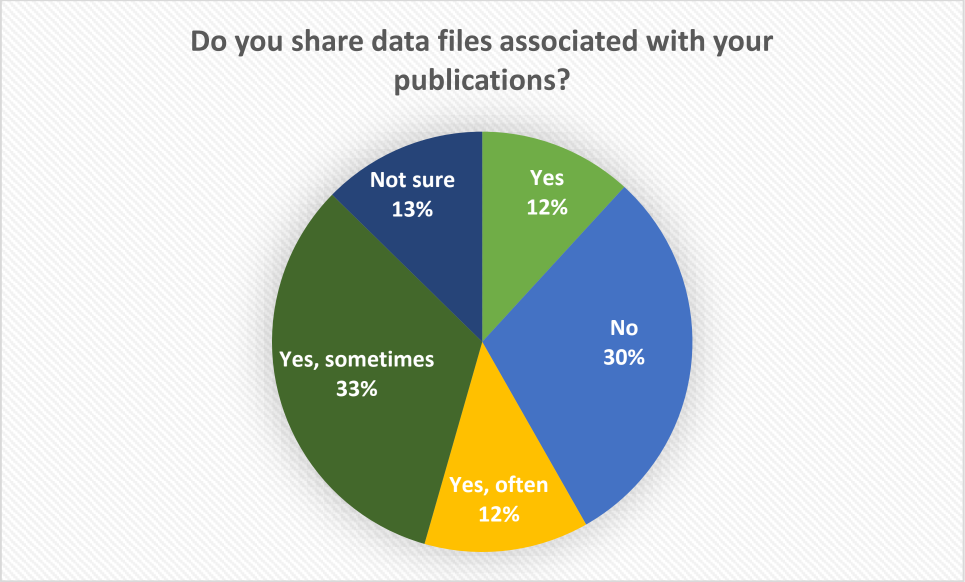
Awareness and use of data repositories



A portion of the data repositories mentioned by participants who responded that they were already using these tools are listed below. Zenodo is the most used, as well as GEO (Gene Expression Omnibus) for genomic data. Note that Google Drive and Git are also mentioned as repositories for research data, which again illustrates some unfamiliarity with the topic.

Zenodo	8
GEO	6
git	4
ArXiv	3
Figshare	2
NCBI	2
Pangea	2
Dryad	1
Google Drive	1

Data sharing



57% of participants say they share their data, more or less frequently, and in a variety of ways that include Zenodo, GEO, Git, c4science, as well as cloud solutions:

Google drive	11
GEO	9
Git	8
Zenodo	7
website	6
EPFL/Lab server	6
email	6
Dropbox	4
c4science	4
Publishers	4
arXiv	3
Figshare	3
SLIMS	2
SWITCHdrive	2

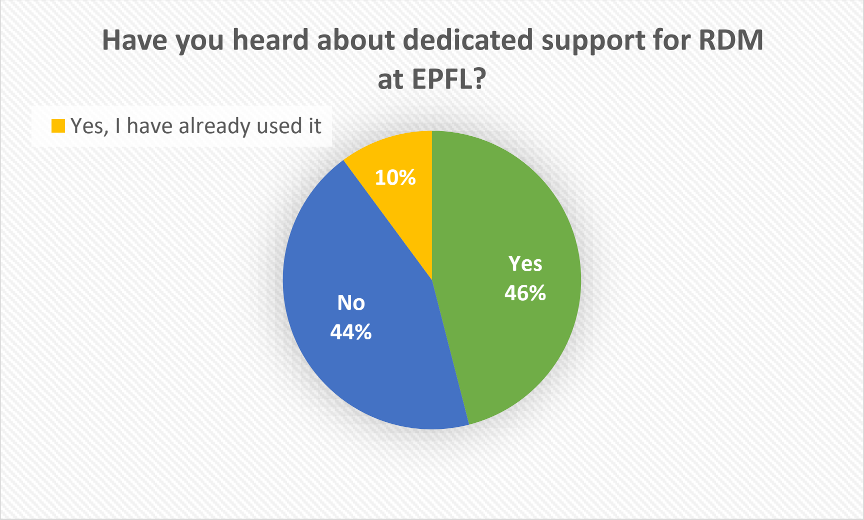
Scientists who do not want to share their data cite the following reasons:

It is not required by the journal or the funding agency	26
Other	11
Data is sensitive	8
I do not find it useful	8
I plan to publish subsequent studies on the same dataset	7
I do not know where and how to share the data	7
Datasets are large	4

“Other” reasons provided by the participants include, for example, "the benefit is not worth the effort it would require"; "I am afraid my data will be plagiarized, reused in a publication without giving me credit"; "not in a shareable form".

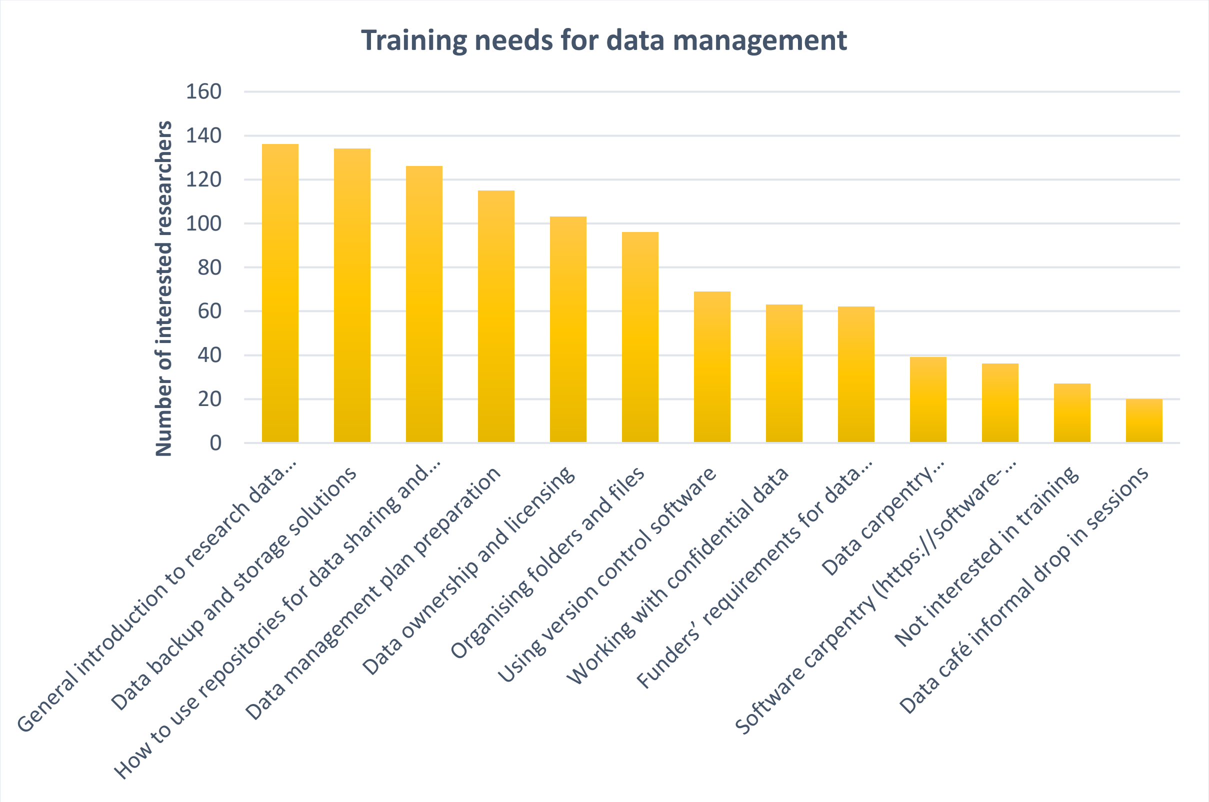
4. Knowledge of available services and training and support needs

Knowledge of the EPFL RDM support service



24 researchers have already received support from the Research Data Library team, but 104 do not know that they can be helped with this topic.

Requests for training in Research Data Management

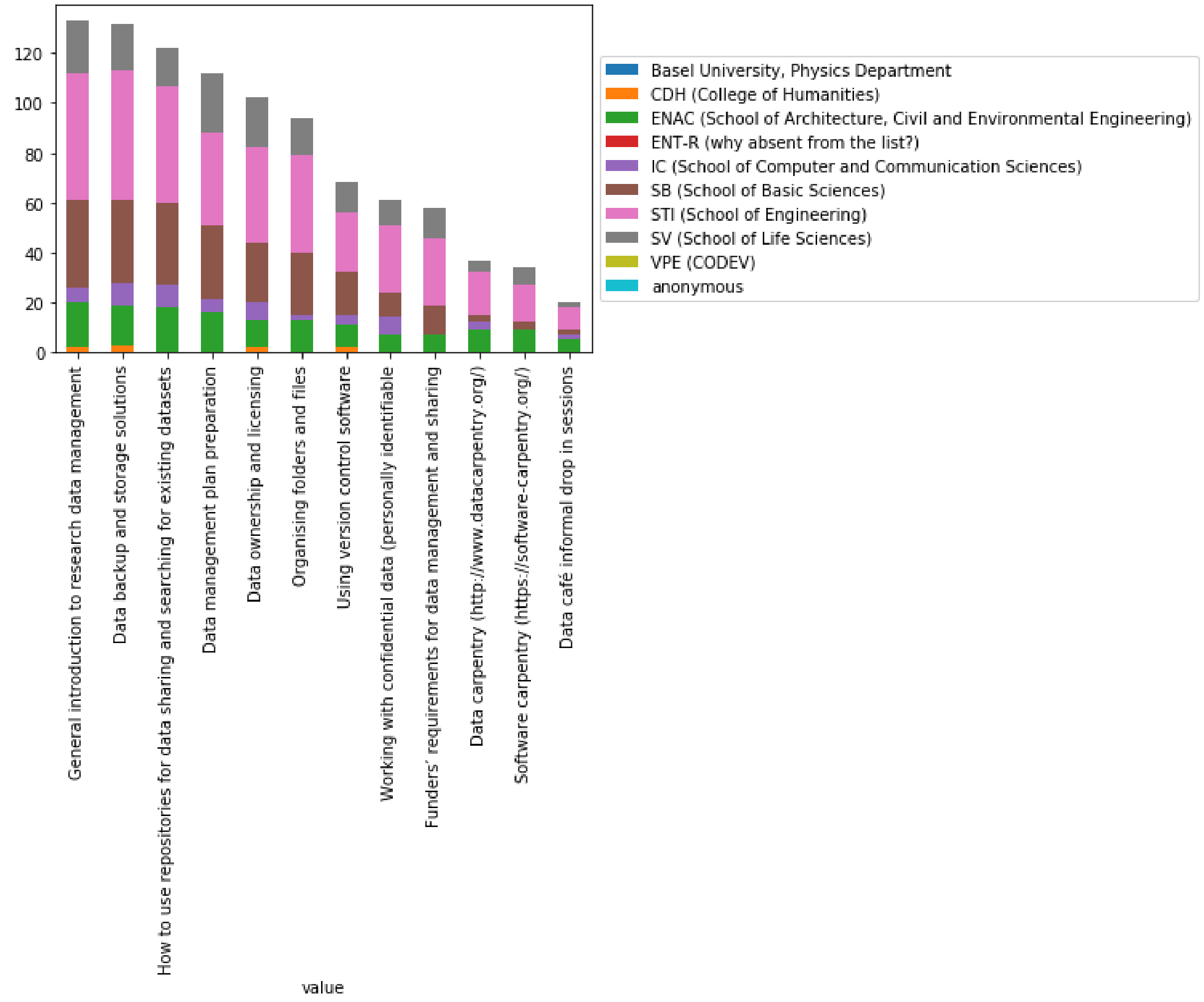


For more clarity, the titles of the different trainings are listed below:

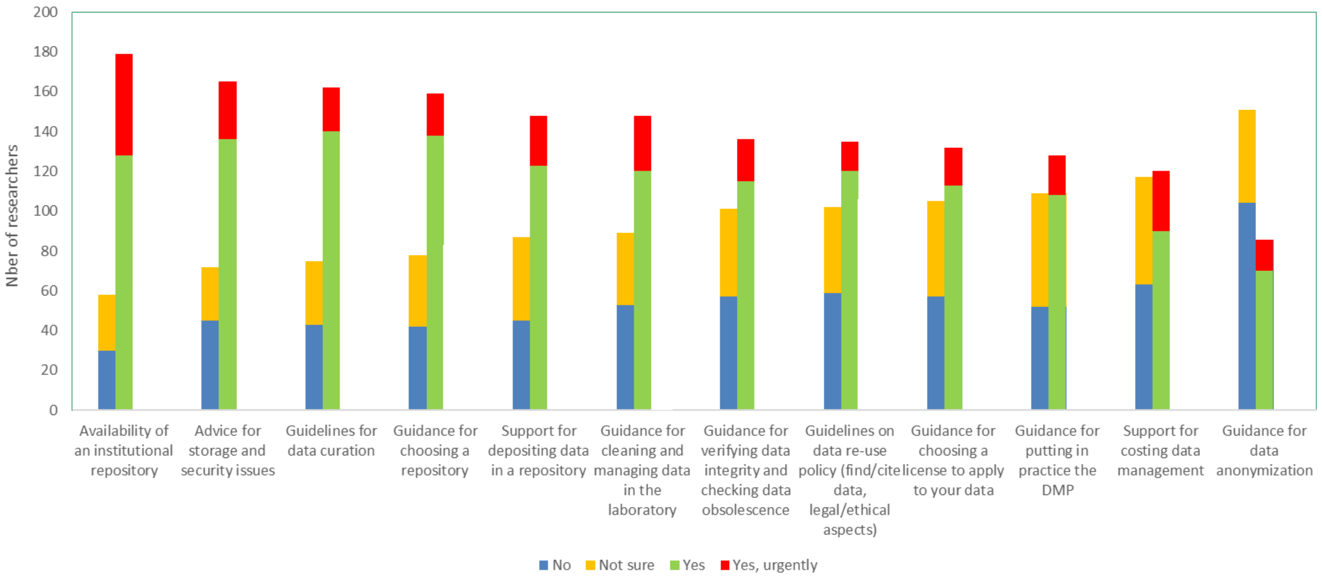
Training Name	**Counts**
General introduction to research data management	136
Data backup and storage solutions	134
How to use repositories for data sharing and searching for existing datasets	126
Data management plan preparation	115
Data ownership and licensing	103
Organizing folders and files	96
Using version control software	69
Working with confidential data (personally identifiable, commercially sensitive etc.)	63
Funders' requirements for data management and sharing	62
Data carpentry (http://www.datacarpentry.org/)	39
Software carpentry (https://software-carpentry.org/)	36
Not interested in training	27
Data café informal drop in sessions	20

A total of **999 requests for training** were identified by this questionnaire, for 210 participants. Only 27 participants were not interested.

It is possible to analyze the requests for trainings in more detail, by type of faculty. The graph below shows, for example, that SB researchers are less interested in training on confidential data.



Support requests and curation services for research data



For greater clarity, the titles of the different means are listed below:

Support	Yes, urgently	Yes	No	Not sure
Availability of an institutional repository	51	128	30	28
Advice for storage and security issues	29	136	45	27
Guidelines for data curation (what data to preserve, what metadata and documentation to link to data)	22	140	43	32
Guidance for choosing a repository	21	138	42	36
Support for depositing data in a repository (includes choice of datasets to be deposited, choice of metadata, data preparation for depositing, link between datasets and publications)	25	123	45	42
Guidance for cleaning and managing data in the laboratory (evaluation of the data life-cycle management in the laboratory, analysis of data workflows, guidelines on data format and interoperability of datasets)	28	120	53	36
Guidance for verifying data integrity and checking data obsolescence	21	115	57	44
Guidelines on data re-use policy (how to find data, how to cite data, legal and ethical aspects)	15	120	59	43
Guidance for choosing a license to apply to your data	19	113	57	48
Guidance for putting in practice the DMP	20	108	52	57
Support for costing data management	30	90	63	54
Guidance for data anonymization in case of sensitive data	16	70	104	47

Overall, half of the participants are interested in services or support for data curation, and a total of **1'698 requests for support** were hereby counted (multiple choice). **The most urgent request is for the provision of an institutional data repository**. The service that receives the least interest is anonymization of sensitive data, which is explained by the fact a relatively smaller percentage of researchers handle (or is aware of) this type of data.

Other comments from participants

Participants had the opportunity to add a comment on the RDM support at EPFL.
Among the 44 written comments, 13 insist on the need for an institutional data repository, 9 mention storage and its price, 5 would like clear recommendations from EPFL (for data or metadata), 5 say that the RDM topic is very important while 2 would like to avoid that data management becomes a burden.

5. Other remarks

- There are still multiple possibilities for analysis at other levels, ex. by sorting: by faculty for each question; by position; by faculty and position; by participants who have already made a DMP; by those who already use a data repository (not Google Drive); by those who have already used our service; by those who are aware of the FAIR principles (to see if they adhere to them and how); etc.

- It is also conceivable to categorize the needs and extrapolate the figures to the entire university in order to estimate the population affected by these categories: the figures are in fact categorized and the question is also multiple choice:

	Survey participants (237)	Extrapolation to all targeted population (4000)
Lack of knowledge of FAIR principles	142	2396
Willingness to discover RDM tools	106	1789
Lack of knowledge about data ownership	116	1958
Unfamiliarity with data repositories	76	1283
Unawareness of the existence of the library's RDM service	104	1755
Training requests	999	4000 (max)
Support requests	1698	4000 (max)

Conclusion

This survey highlights the importance of increasing awareness of RDM best practices among EPFL researchers. The results of this study have already been used, along with other initiatives, to gather information on data dissemination platforms actually adopted by the EPFL population: one of the outputs is the comparative table on go.epfl.ch/datarepo.

Data curation services are also of great interest: more than half of the participants would like to benefit from 11 of the 12 services mentioned. Only a data anonymization service seems not to capture the interest of respondents, probably due to a low fraction of researchers handling personal data.

The survey also shows the need to refocus training and support services via continued education in FAIR principles and the basics of RDM best practices. Indeed, half of the participants requested training on RDM (data storage and back-up, data repositories, writing DMPs, tools, …). In light of the respondents' roles, and in the current context of Open Science and funder demands, it appears fundamental that best practices in RDM be taught to researchers from the beginning of their career.
Efforts in this sense are already in place: along its current offer of training, the EPFL Library offers well-established services, e.g., information on data storage and collaborative solutions, guidance on data dissemination platforms, support and review of DMPs, etc.
However, more work needs to be done so that RDM best practices become an integral part of researchers' daily routines. In the future, it will be essential that RDM training be strengthened and offered at multiple levels, in the doctoral school curricula and along researchers' career paths.