Quantitative assessment of research data management practice – 2019 Short Report

By Eliane Blumer and Francesco Varrato – EPFL Research Data Library Team

1 Background information

The objective of this survey is to collect information on academics' habits in terms of managing their research data, and in turn to identify their needs of services and support. This survey has been conducted in collaboration with three other institutes: TU Delft, the University of Cambridge and the University of Illinois-Champaign.

This is the second time that such a survey has been conducted, even if with slight changes, therefore allowing to analyze the evolution of practices and needs of researchers in RDM. The following report aims to analyze the EPFL results, and not the composed results from all the four institutions.

1.1 Methodology

The common core of the survey, developed in 2017 in collaboration with TU Delft, Cambridge University and the University of Illinois, was kept. A specific section containing specific questions for the EPFL, has been developed. In 2019, some questions were adapted for clarity. Differently from 2017, in 2019 the survey has not been conducted using Google Forms, but with SurveyHero, a Swiss-based and GDPR-compliant surveying platform.

In the 2019 survey, all questions have been set to require an answer, except: open-ended questions, probing questions (displayed following a display-logic as a result of previous answers), and personal questions about participants' contact information. A big change is the reduction of open-ended questions: in fact, in the 2017 survey, the presence of many open-ended questions implied a certain struggle to analyse their results, even though it's been useful to pave the way for the 2019 iteration.

Communication on the survey has been made on July 2, 2019 by email to the groups mainly concerned (scientific personnel, teachers, professors, PhD students, and lab responsible staff), as well as by social media as Facebook, LinkedIn, and Tweeter. The information was also relayed by the liaison librarians. A reminder by email has been sent on July 22. The survey has been closed approximately 7 weeks after launch.

Based on the EPFL Key Figures[1] for the segments of population targeted by the emails, about 6'060 people might have received the email communicating the 2019 survey, and we received 285 Total Responses.

Even if the estimated size of the targeted population is 7.6% smaller than the 6'530 estimated for 2017, the number of total responses in 2019 has grown by 20% (there were 237 responses in 2017). This participation growth may already imply a stronger resonance with the researchers of the themes linked to RDM, at least compared to 2017.

Unfortunately, not every participant completed the survey in each and every question, as a 76.8% of Completion Rate (= People who have participated and completed survey) has been measured.

2 Key results

In general, RDM (Research Data Management) practices and responsibilities among respondents remain similar to 2017, including writing of DMPs, automated back-ups and the choice of mainly general data repositories, and applying FAIR principles.

Respondents also continue to report uncertainty for topics such as the ownership of data, existence of DMP, data loss or data publication.

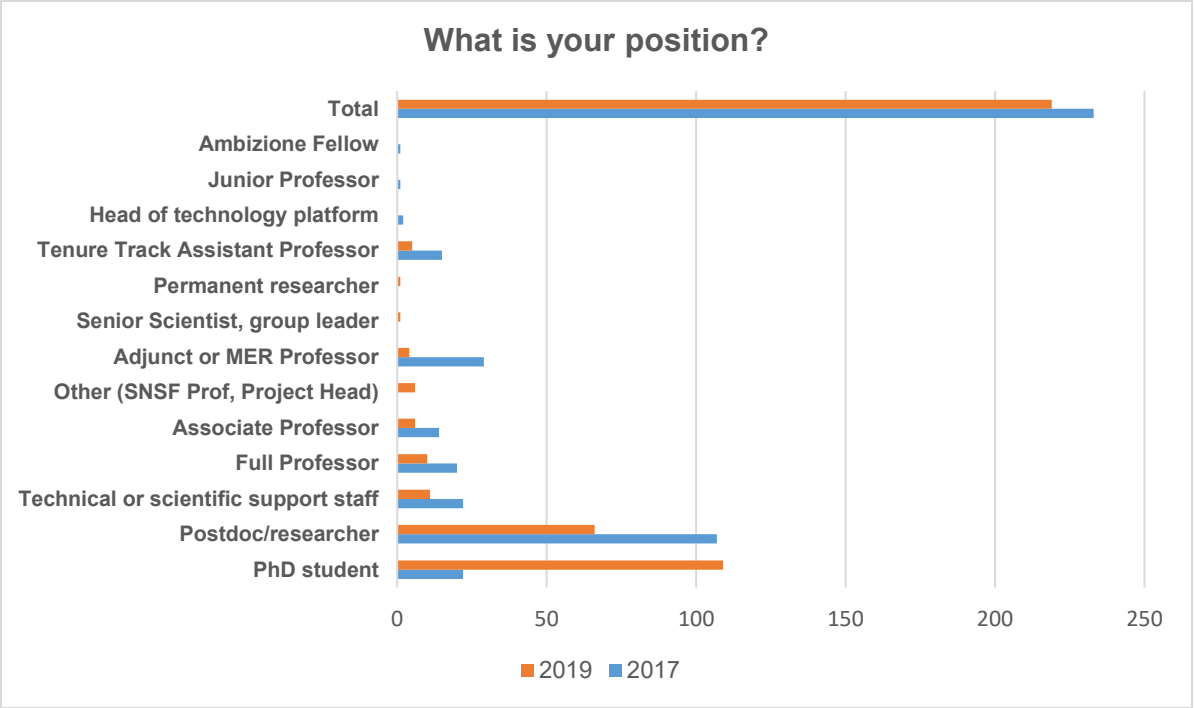
A relatively better situation in 2019 is registered regarding the versioning, the use of manual back-ups, and awareness around DMPs.

The main training need is for data documentation and organization, and support is mainly needed for data publication (choosing repositories and preparing data for publication).

2.1 Who answered the survey

In 2019, 50% of respondents are PhD-students, 5 times more than in 2017 (was 9%). On the flip side, post docs respondents are down from 46% to 30%. This might be linked to the sensitizing work done around RDM topics, reaching earlier career levels.

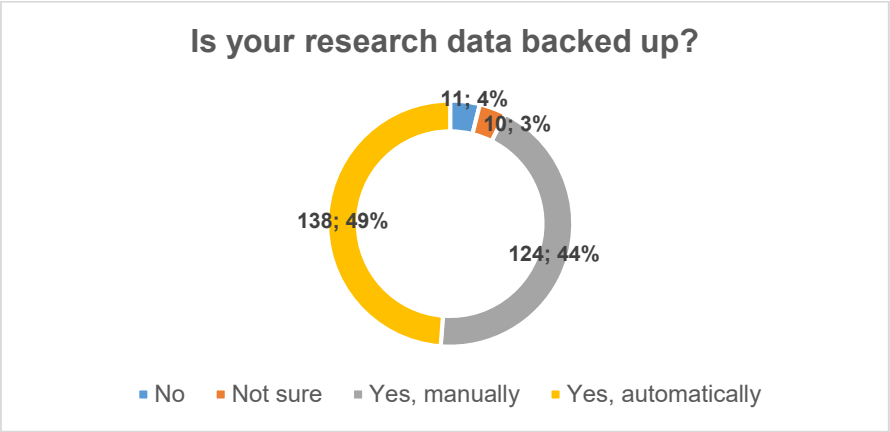
[1] EPFL Key Figures: www.epfl.ch/about/overview/figures. A total of 6'060 people: 5,713 scientific, administrative & technical staff (including PhD assistants), plus 347 professors.



Among the different faculties of the respondents, the highest participation has been registered from SB (30%), followed by STI (28%), which has also seen the largest increase, and ENAC (16%). We largest decrease in participation has been registered for SV (from 15% in 2017 to 10% in 2019).

2.2 Research data backup and data loss

The results indicate a similar practice of researchers towards data back-ups as in the previous 2017 survey: ~50% use automated back-up solutions. What's new is the fact that almost all the others seem to be using at least a manual back-up, which indicates a positive trend, even if perfectible.

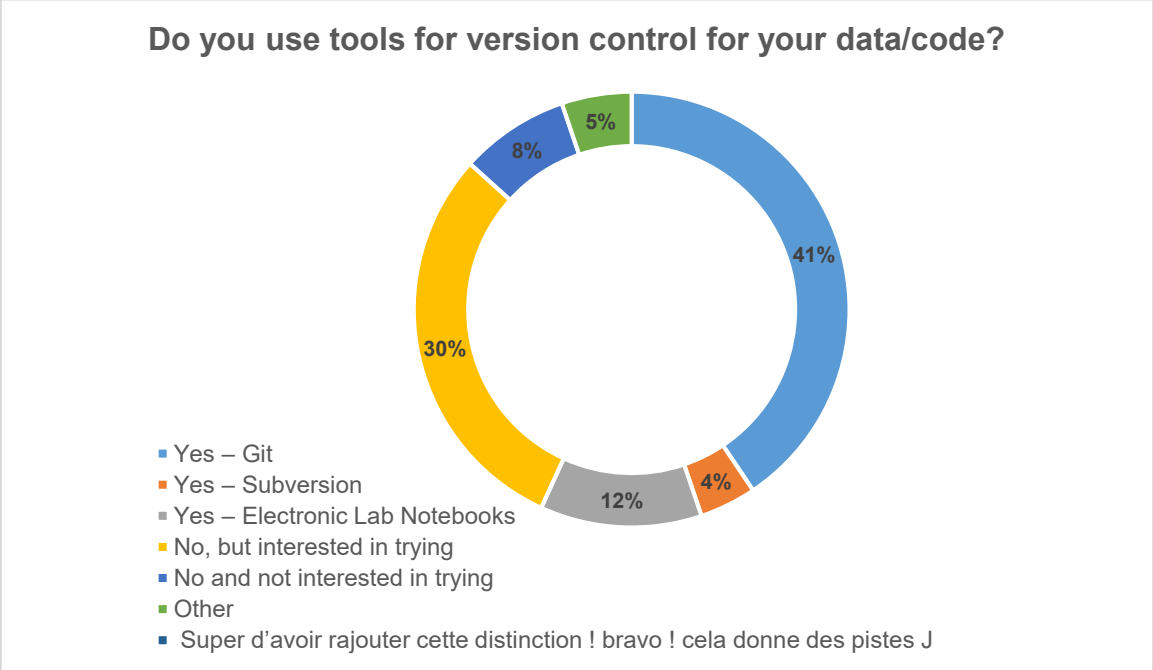


This positive result is strengthen by the high percentage of responses reporting no data loss (81.27%). On the other hand, ~10% of respondents report to be unsure or to have had some data loss: this percentage is not small and points out to a strong need for improving awareness around the risks of data loss.

Among those who report to have lost data, a whopping 58% perceive / recall a data loss bigger than 7 day of work.

2.3 Version control for data and code

We remark quite a polarized situation: 42% of respondents do not use any version control tool, while 53% use some. The most mentioned tool is GIT (45%). This shows a certain awareness and practice of specialized solutions for research data management, especially for version control.



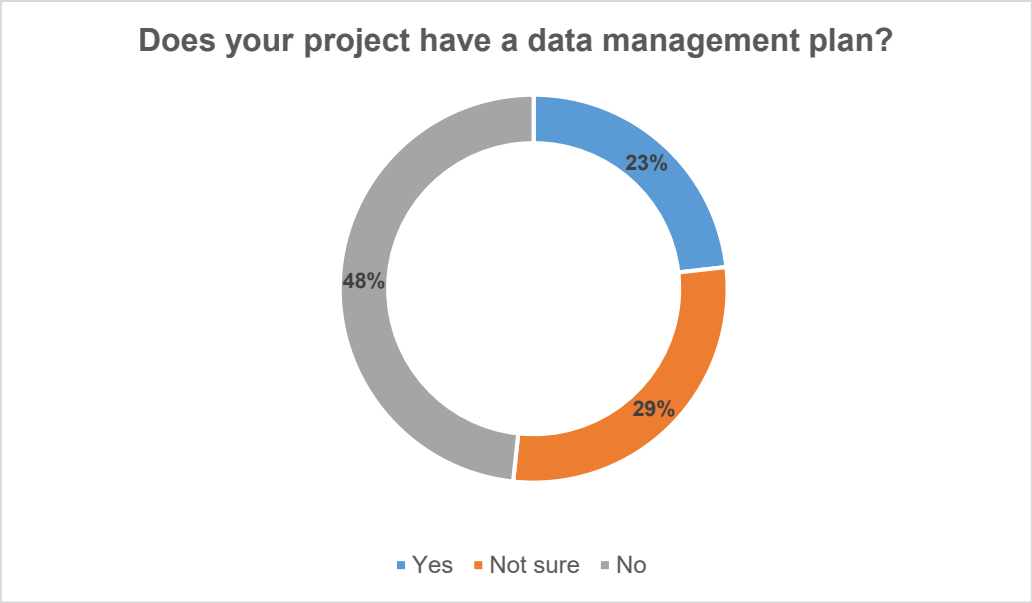
The survey identifies an important pain point: 33% of respondents, i.e. one third, do not use a version control, but are interested in trying.
The comparison with 2017 is not so direct, as the more general term of "research data management" was used, and the versioning only mentioned as an example.

2.4 Data management plan

The answers concerning the practices around the DMP denote a substandard habit: only 23%of the researchers are aware of having a DMP, while 77% are not sure or don't have one. This is a slight progress compared to 2017, as the sum of researchers not sure or not having a DMP was 84%; moreover, in 2019 "only" 48% reported of not having a DMP, instead of 65% in 2017.

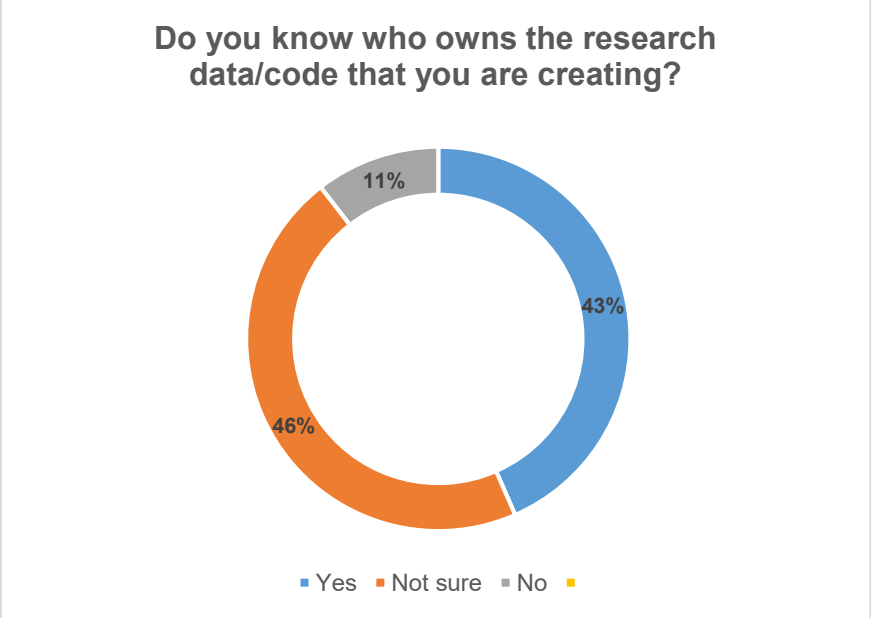
The large majority of those who stated to have a DMP (68% of 23%), name the funder's requirement as main reason for having it, while the 15% name their supervisor.

Another remarkable result is the fact that only about one third of the ones who said to have a DMP (32% of 23%), also said to have updated it. This further highlights the scarce interest for the DMP as a worktool, indicating the necessity to make its adoption easier and wider.



2.5 Ownership of data and code

While 43% of respondents stated to know who the owner is, the negative answer is stable at about the 10% as in 2017.



Researchers are generally confused about the ownership of data or code, as almost one half of the respondents (46%) is unsure about who owner might be. This growth in incertitude (from 36% in 2017), highlights a certain difficulty surrounding the ownership of digital works, but might as well indicate an increased self-awareness of such difficulties.

The majority (73%) of the ones stating to know who owns their data or code, name the EPFL as the owner. On the other hand, the 41% also think of being the owner of the data or code. In this multi-choice question, this implies an overlap of ~14% of the answers, confirming the difficulties around the ownership of the data/code, and possibly about concepts such as authorship, ownership, use rights, etc.

2.6 Responsibility for data/code management

As in the 2017 survey, the majority of answers state a self-responsibility towards the management of data (plus code, in the 2019 version): the 86% states such a self-responsibility. This also highlights the lack of specialized support, e.g. Data Stewards at the faculty level, or Data Manager at the lab level.

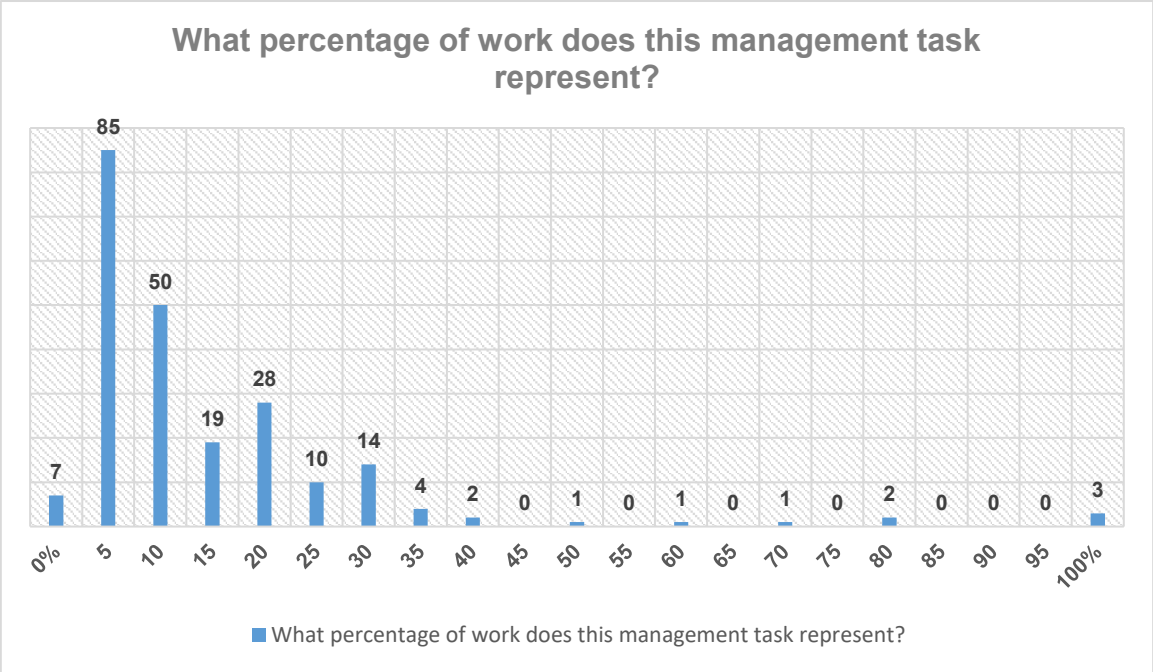
We also remark a high uncertainty: 27% think the supervisor is responsible, 10% name the unit's IT manager, 6% name the EPFL, 8% think it is the project partner, and 12% are not sure. The sum of percentages is higher than 100%, as the question allowed for multiple choices.

2.7 Percentage of work used for data management

The answers result in an average estimation of 15% for the work percentage used in RDM activities.

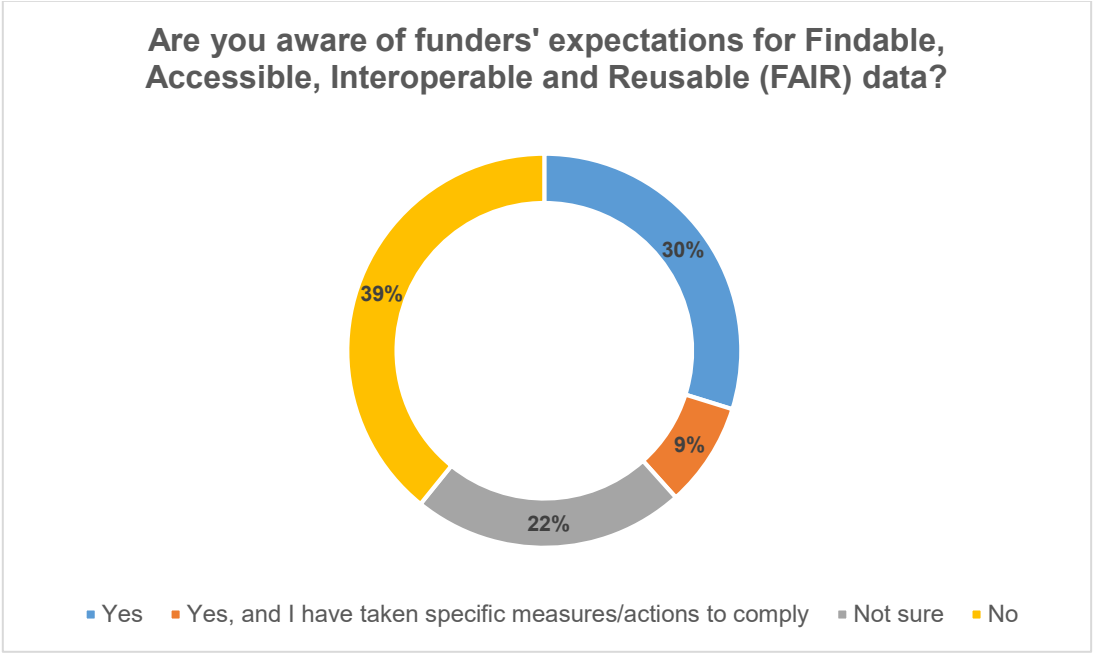
The large standard deviation (± 16% of worktime) reflects a large variety of practices: while a relative majority of participants (63%) estimate a percentage lower or equal to 10%, many participants estimate nonetheless a higher percentage.

In this scattered landscape, by weighting the participants answers with their worktime percentage dedicated to RDM, the 12% of respondents who declared a 20% worktime represent the population segment with the biggest impact on EPFL with regard to RDM activities.



2.8 FAIR data principles

The results are well comparable to the 2017 survey, meaning that awareness about FAIR dataset principles is not augmented: the awareness about FAIR dataset principles is weak, with a large percentage of participants (62%) being unsure or unaware of them (was 61% of 2017).



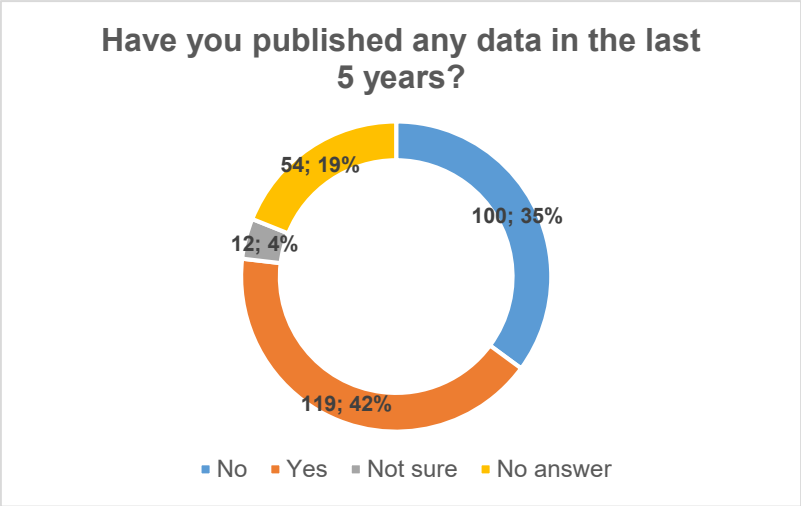
On the other side, almost 9% of respondents state to have taken specific measures/actions to comply with funders' expectations on FAIR data: this is far from irrelevant and beyond the simple awareness. We argue that part of the respondents might have answered negatively because it feels not directly concerned by the "*funders*" part of the question: in fact, 50% of the respondents are PhD students.

A simple text analysis of the open answers of those who have taken specific measures/actions highlights that data is way more cited than code (11% vs. 2%). Moreover, topics linked to

publishing (repository, paper, publication, etc.) and research reproducibility represent the major drive to put in practice the FAIR principles. This is probably a consequence of the implicit link to data publication expected by the funders, as well as to the usual association of the FAIR principles with Open Science.

2.9 Data publication (last 5 years)

We notice a polarized situation: 41% of participants did publish data in the last five years, while 35% did not.



The “Not sure” answer (4%) is interesting, as one may wonder whether respondents are not sure what data publication is, or whether their data has been published without them knowing.

A substantial 19% of the respondents did not answer to this question. We argue that not answering could reflect the ignorance about data publication as for the “Not sure” answer.

2.10 Research data repositories

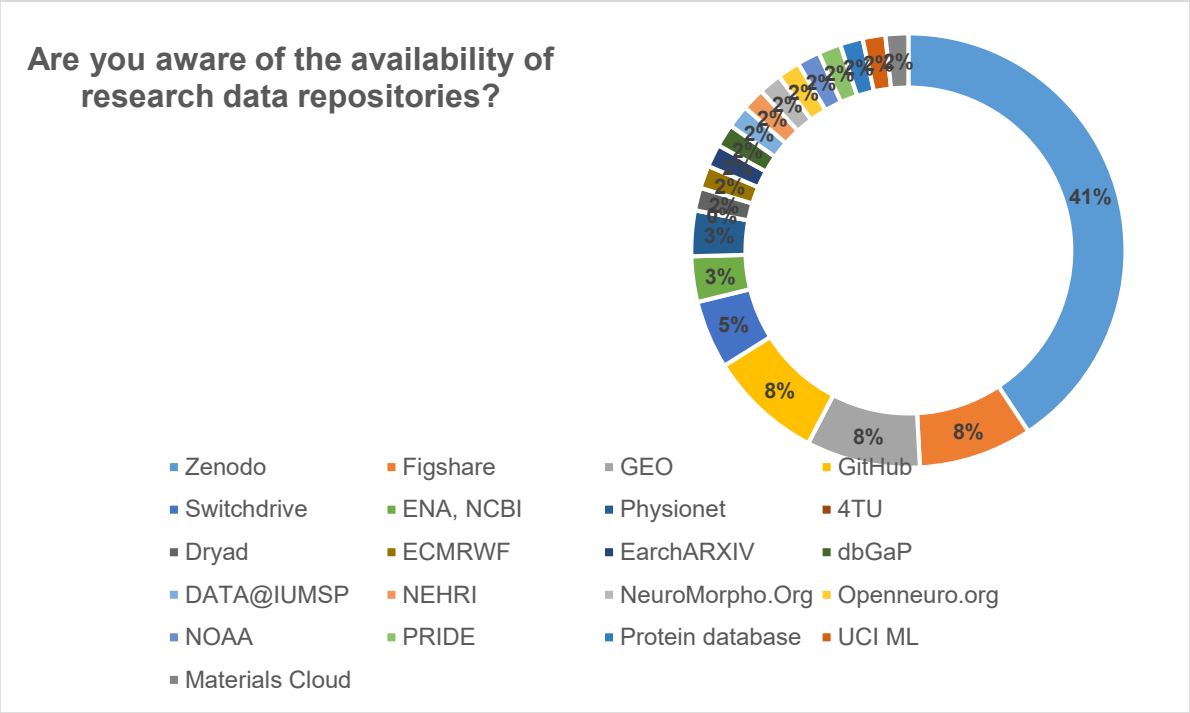
The participants aware of research data repositories, but not using them, is 12% lower in 2019 than in 2017 (44% instead of 50%). One might wonder why they are not using repositories.

When asked which repositories they use, the surveyed population completely mixes up data repositories, code sharing platforms, and cloud solutions of any sort.

Generic data repositories are mentioned the most, with Zenodo being the most mentioned (41% in 2019 versus 28% in 2017). this highlights well a known paradox of data repositories and their bad usage, as the deposited content is oftentimes rather low curated and very heterogeneous. Other generic data repositories, such as Dryad or Figshare, are also mentioned.

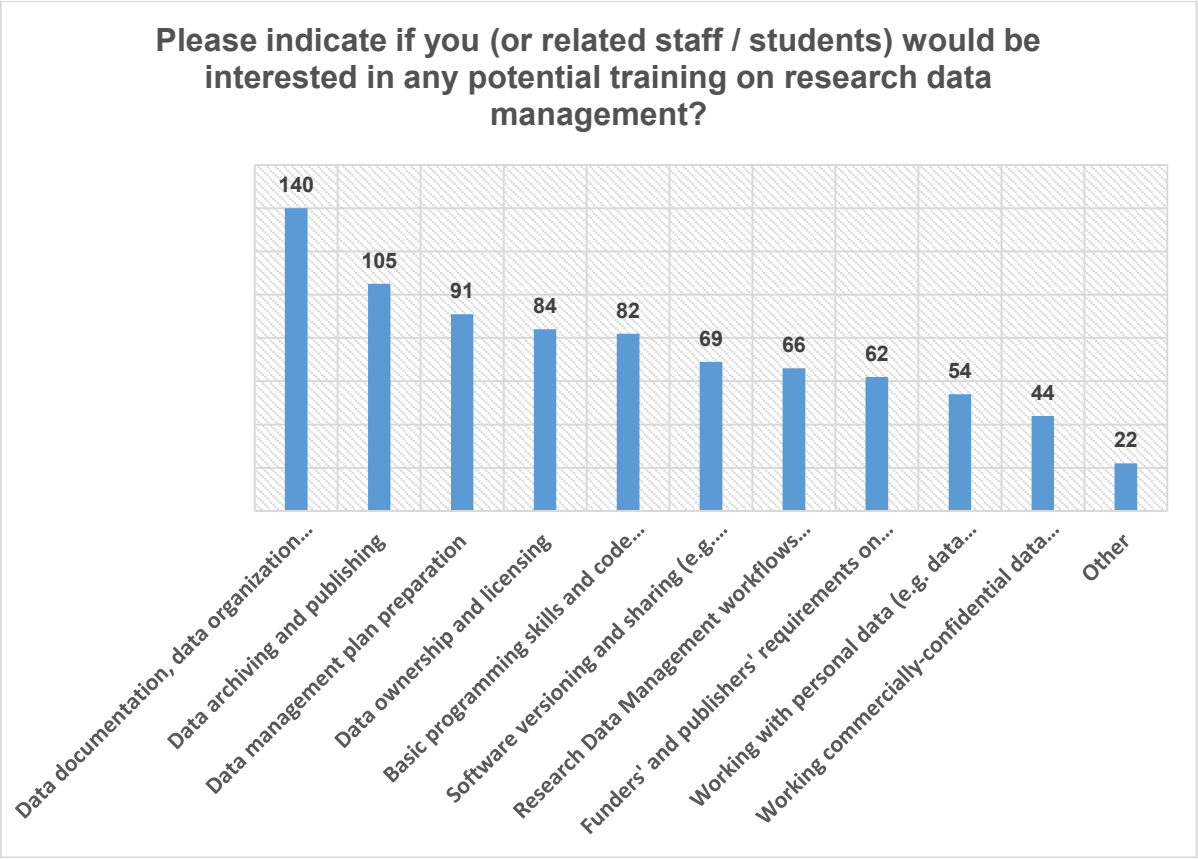
Respondents also mention disciplinary-specific data repositories, whereas many of them are in the Life Sciences field.

For code, GitHub seems to be the favorite repository.



2.11 RDM training

Demand for RDM training is not too clear cut, as high request percentages concern many topics.

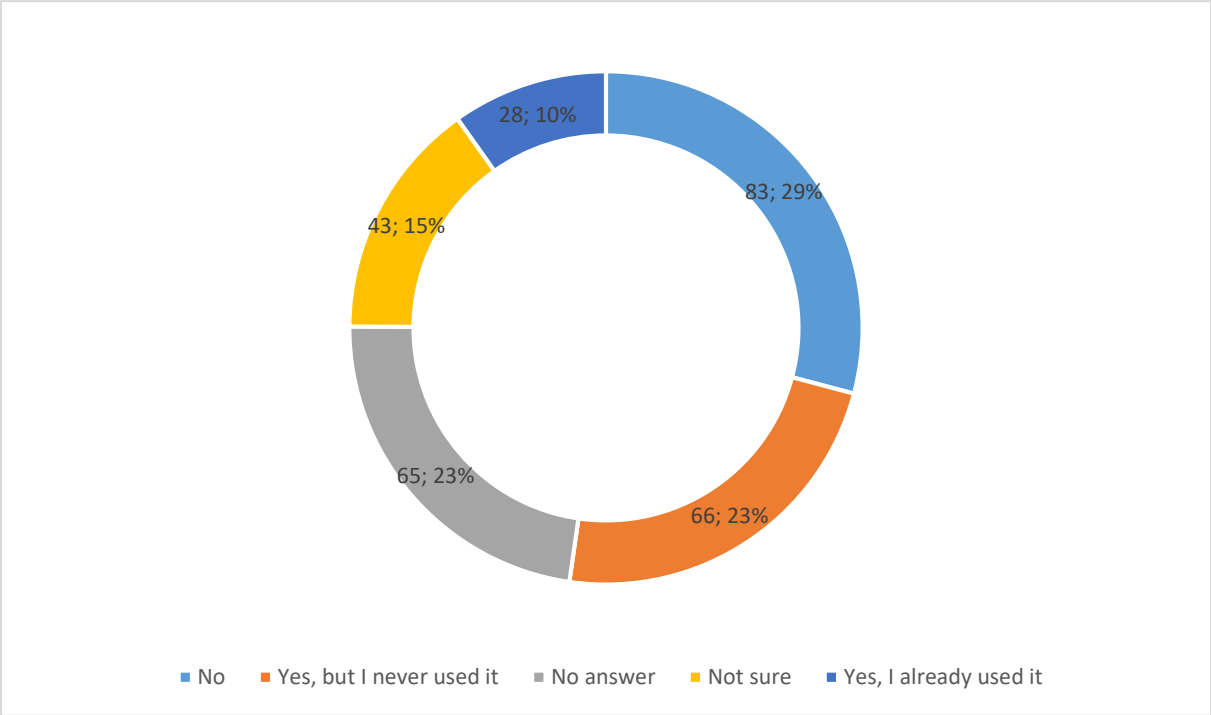


The majority of demands concerns Data documentation and data organization and storage (61%), while demand for training on Data archiving and publishing comes in second place

(45%) with almost the same percentage than for DMP preparation (40%), Data ownership and licensing (36%) and Basic programming skills (36%). The sum of percentages is higher than 100%, as the question allowed for multiple choices.

2.12 Awareness of RDM support

The results show a decrease in all answer categories compared to 2017. Only 12% of participants have already used the RDM support service, and the majority (35%) has never heard about the RDM support.



Next up are respondents who have heard about the support, but never used it (28%).

It is interesting that 18% are not sure whether they've heard of the RDM support, which might be linked to the scattered services landscape at EPFL. This confusion or ignorance might also explain the high rate of participants (27%) who did not answer.

2.13 RDM support

The heterogenous interest for RDM training (see 2.12) well matches the needs for RDM support, as only the request for Guidance for data anonymization has received significantly less interest than the other options, despite this option has seen in 2019 the highest increase in total choices (10 times more).

The three options around a data repository are the most chosen in 2019 (ranking 1st, 2nd and 3rd), gaining in importance compared with the 2017 survey.

While Storage and security issues remains as important as in 2017, the need for RDM recommendations in the laboratory has registered the larger decrease (2nd to 9th), possibly due to the change of the wording, but also to a real low need of such services.

Need for RDM support *(& changes compared to 2017)*	**2019 rank**	**2017 rank**
Guidance 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) *(has been detailed)*	**1** (↑4)	5
Availability of an institutional repository *(same question)*	**2** (↓1)	1
Guidance for choosing a repository *(same question)*	**3** (↓1)	2
Advice for storage and security issues *(same question)*	**4** (-)	4
Support for data management plan implementation *(unclear abbreviations have been removed)*	**5** (↑2)	7
Guidance for choosing a license to apply to your data *(same question)*	**6** (↑3)	9
Advice for verifying data integrity and checking data obsolescence *(same question)*	**7** (↑1)	8
Guidance for cleaning and managing data in the laboratory *(was "lacking "cleaning", added as part of ELN/LIMS service reflection")*	**8** (↓2)	6
RDM state-of-the art in the laboratory & related recommendations *(replaced the question about "data curation")*	**9** (↓7)	2
Guidance on finding datasets and how to re-use them *(has been detailed)*	**10** (-)	10
Guidance for data anonymization in case of sensitive data, commercial data or data licensed by third parties *(has been detailed)*	**11** (-)	11