Research Data Management Survey 2023 - Report on EPFL

This report and the Appendix with the results are based only on EPFL collector used to gather the results at EPFL. Graphs are available in the Appendix or on the online SurveyHero report: https://library-survey.epfl.ch/results/1579761/nbzon8oxtwl9yohhctkxh0h4ov93bpxl
For the full set of data, including other institutions collectors, refer to the dataset in Zenodo: https://doi.org/10.5281/zenodo.13836947

Context and Key Insights

Effective Research Data Management (RDM) is a critical aspect of the scientific process, ensuring data and code are well-organized, reproducible, and securely preserved. To better understand researchers' practices and needs, the EPFL Library conducts a biannual survey aimed at improving and developing tailored RDM services. In 2023, this survey was conducted in collaboration with DaSCH, Eawag, ETHZ, FHNW, and UNIL, reflecting a broader effort to harmonize RDM solutions across institutions.

The survey, conducted at EPFL between August and September 2023, sought to gain a comprehensive understanding of the active data management practices within the institution. Over this period, a broad spectrum of the EPFL community, amounting to 6.4% of the addressed population, interacted with the survey, but only a smaller fraction of 248 individuals (2.8%) provided responses. This report predominantly captured perspectives from EPFL's core research community, particularly doctoral students and postdoctoral researchers from engineering and basic sciences.

Data Storage and Open Practices

The survey highlighted a diverse range of data storage practices. While personal devices remain widely used, there is a clear shift towards institutional servers and cloud solutions. The increasing adoption of Git-based collaborative platforms signals a broader move toward open and reproducible research. This trend is further supported by a strong preference for open formats and open-source software. However, despite this shift towards openness, many researchers remain uncertain about open practices, pointing to gaps in awareness and clarity. In this context, data repositories and data/code journals are emerging as valuable tools, gaining traction as reliable platforms for research dissemination and preservation. Despite this progress, most data sharing remains internal, although an increasing number of researchers are engaging in external collaborations.

Documentation, Versioning, and Standardization

More researchers reported completing a Data Management Plan (DMP) in 2023 compared to previous years (2019 and 2021). However, significant knowledge gaps persist in areas such as the FAIR principles and data ownership. The survey also underscored the crucial role of README files in documentation, suggesting an opportunity to introduce standardized templates. Versioning practices are evolving, with Git-based platforms gaining popularity. However, a considerable number of researchers still rely on manual versioning methods, highlighting the need for modernization. Additionally, naming conventions for data organization vary widely, reinforcing the value of standardized guidelines—an effort already underway. One of the most striking findings was the widespread lack of familiarity with metadata standards, emphasizing the need for targeted educational initiatives.

Key Recommendations

The survey results suggest several areas for improvement:

1. **Enhanced RDM Infrastructure** – Integrated storage solutions should be expanded to accommodate the diverse needs of the EPFL research community.
2. **Targeted Educational Outreach** – Training programs should address knowledge gaps in metadata, licensing, and data ownership.
3. **Hybrid RDM Support** – A combination of online resources and in-person guidance can cater to different researcher preferences.
4. **Ethical Data Management** – Institutional guidelines should continue to emphasize ethical considerations, particularly for sensitive data.

Conclusion

EPFL researchers employ a wide range of data management practices, reflecting both autonomy and evolving institutional support. The survey findings outline a clear path toward a more structured and standardized RDM framework, ensuring improved efficiency, transparency, and collaboration across the research community.

A. Survey Targeting

- The EPFL collector has been activated on August 15, 2023, and closed on September 30, 2023, thus staying open **one month and a half** for collecting answers.
- We targeted our audience twice by email, with **theoretically 6879 people addressed**, plus news on our library's web page and screens.
- We have no clue how many people accessed the survey via email or news or screens (or how many even saw email or news or screens), but in total only **442 people clicked** on the collector link, i.e., 6.4% of those addressed.
- Of these, only **248 started replying**, i.e., 56% of those who clicked = 3.9% of those addressed.
    - Multiple language choice instead of just English (EN) has been introduced in this 2023 edition, and we observe that 39% of people who started answering preferred a language other than EN, in fact:
        - 151 people started answering in EN
        - 81 in FR
        - 11 in DE
        - 5 in IT
- Of the people who started replying, only **192 completed their response**, i.e., 77% of those who started to answer = 43% of those who clicked = 2.8% of those addressed.

To compile this report, I mostly use the information gathered from the 248 respondents. Even if they didn't complete all the survey, their input can be important for the specific questions where they answered. I will highlight the cases in which an analysis should need to differentiate between complete or incomplete questionnaires.

Population Comparative Overview Across 2019, 2021, and 2023

The RDM surveys over the years have illustrated evolving engagement patterns and demographic shifts among the EPFL community.

- **Engagement Trends**:
    - **2019**: A participation rate of 3.6% was observed with a significant drop in the number accessing the survey link from prior estimates.
    - **2021**: Participation improved with a 6.25% response rate, despite a reduced estimated audience from more than 6,000 in 2019 to 4,000.
    - **2023**: The response rate has dipped again, to 3.9%, though the introduction of multiple languages saw 39% opting for non-English options.
- **Faculty Participation**:
    - **2019 & 2021**: Both years saw a balanced representation across faculties, with the School of Engineering and Basic Sciences leading. The distribution roughly mirrored the EPFL's faculty composition.
    - **2023**: The School of Engineering surges ahead with 36%, followed by Basic Sciences at 20% and Life Sciences at 14%.
- **Role Distribution Shifts**:
    - From **2017 to 2021**, there was a decline in Professor respondents from 34% to 14%, hinting at a growing interest in RDM topics among younger researchers.
    - The categories of Postdoc researchers and PhD students showed fluctuations but overall increased engagement with RDM topics, with Professors at 9%.

B. Respondents' Affiliation & Roles

Based on questions about affiliation, plus D2. to D4., this survey predominantly reflects the views of EPFL's doctoral students and postdoc researchers, and particularly from engineering and basic sciences. While there is a wide range of engagement RDM, a notable trend is the autonomy and self-reliance in setting data management practices. This provides valuable insights into the current state of Research Data Management (RDM) practices within the institution.

1. Institutional Affiliation:

- **EPFL** is the primary affiliation for most respondents, I.e., **96.37%** (239 out of 248).
- **ETHZ** and **other institutions** had a smaller representation with **0.81%** (2 out of 248) and **2.82%** (7 out of 248) respectively.

2. EPFL Affiliation:

A diverse range of departments within EPFL responded, with a notable concentration from the School of Engineering and the School of Basic Sciences.

- **STI - School of Engineering** had the highest representation with **35.56%** (85 out of 239).
- This was followed by **SB - School of Basic Sciences** with **20.08%** (48 out of 239) and **SV - School of Life Sciences** with **14.23%** (34 out of 239).
- The least represented were the **CDM - College of Management of Technology** and **EPFL Research Facility / Platform** with **0.84%** each.

3. Role at EPFL:

Most respondents are deeply immersed in research, with doctoral students and postdoc researchers leading the representation. This indicates that the survey primarily reached the heart of the academic research community.

- **Doctoral students / assistants** made up the largest group with **43.07%** (87 out of 202).
- **Scientific Collaborators / Postdoc Researchers** represented **23.76%** (48 out of 202).
- The least represented roles were **Data Manager / Data Scientist** with **1.98%** and **Administrative staff / Higher management** with **1.49%**.

4. Hours spent on managing data/code:

While most respondents spend a few hours weekly on RDM, there are dedicated individuals who allocate substantial time, suggesting varied levels of engagement with data/code management.

- Most respondents spend between **1 to 5 hours** weekly managing data/code.
- The **highest frequencies** were observed at 1 hour (**12.87%**), 4 hours (**7.43%**), and 5 hours (**7.43%**) per week.
- There are also significant numbers of respondents who spend **30 hours** and **42 hours** weekly, both accounting for **6.93%** and **2.97%** respectively.

5. Rules for managing data/code:

Autonomy is a key theme, with many respondents setting their own rules for data/code management, though a good portion also rely on institutional or group norms.

- **48.02%** (97 out of 202) set the rules for managing their data/code themselves.
- **13.37%** follow the rules set by their Principal Investigator (PI) or Supervisor.
- **12.38%** follow historical conventions without a specific person setting the rules.
- **8.42%** indicated that no rules are implemented for managing their data/code.

Nearly half of the respondents establish their own rules for managing data and code, reflecting a strong sense of autonomy and self-reliance among EPFL researchers. While this flexibility is valuable and often essential to accommodate diverse project needs, it can also result in inconsistencies in research data management (RDM) practices across the institution. Implementing standardized practices through training and integrated tools could help promote consistency, enhance data reproducibility, and strengthen collaboration, ultimately creating a more cohesive and efficient research ecosystem.

C. Storage and sharing of active data

Based on questions A1. to A5., EPFL's research community relies on a mix of personal devices, institutional servers, and cloud solutions for data/code storage, with a growing trend towards collaborative platforms like GitHub/GitLab. While backup practices are widespread, there's a notable variation in backup frequency and method, underlining the need for heightened awareness and possibly more standardized backup practices. The combination of personal and institutional storage solutions, along with varied backup frequencies, emphasizes the importance of promoting best practices in data management and backup protocols.

1. Data/Code Storage Solutions

While respondents predominantly rely on personal devices like Laptops/Tablets for data storage, there's a significant leaning towards institutional servers and collaborative platforms like GitHub/GitLab, with cloud storage also being a popular choice. This is a clear progression over the span of the span of the span of the 2019, 2021, and 2023 surveys, from local or institutional storage to more collaborative, online platforms. While this increased awareness and adoption is beneficial for collaboration, it also underscores the need for enhancing training and communication on the use of online solutions.

- **Personal Devices**: 66.23% of respondents store data/code on Laptops/Tablets.
- **Institutional Infrastructure**: 59.31% use institutional servers.
- **Code Sharing Platforms**: Platforms like GitHub/GitLab are used by 61.47%.
- **Cloud Storage**: Nearly half, or 48.48%, use cloud storage solutions.

2. Collaborative Access for Sharing

Code sharing platforms emerge as the leading choice for collaborative data/code sharing, but institutional servers still hold substantial importance in the sharing ecosystem.

- **Code Sharing Platforms**: Preferred by 64.19% for collaborative access.
- **Institutional Servers**: Chosen by 52.84% for sharing data/code.

3. Backup Solutions

While institutional servers remain a primary backup destination, there's a noticeable trend of backing up on code sharing platforms, external drives, and cloud solutions.

- **Institutional Servers**: A leading backup choice with 46.72% relying on them.
- **Code Sharing Platforms**: Used by 45.41% for backups.
- **External Drives**: 37.12% backup on external HDDs/SSDs.
- **Cloud Solutions**: Opted by 33.62% for backups.

4. Backup Methods

The community exhibits a blend of backup methods with many combining automatic and manual approaches, though a significant portion still relies solely on manual methods.

- **Mixed Method**: 40.53% use both automatic and manual backup methods.
- **Manual Only**: 36.12% rely solely on manual backups.
- **Automatic Only**: 15.86% exclusively use automatic backups.

5. Backup Frequency

Backup routines among respondents vary widely, from continuous to yearly schedules, suggesting diverse data needs and possibly different perceptions of data value or volatility.

- **Continuous/Hourly**: Chosen by 16.81% of respondents.
- **Daily**: Opted by 19.91%.
- **Weekly and Monthly**: Both frequencies are preferred by 19.03%.
- **Yearly**: 11.5% perform backups only on a yearly basis.

D. Data/code organization and documentation

The answers to questions B1. to B6. highlight the prevalent use of README files for documentation, the dominance of Git for versioning, and a varied approach to naming conventions. The responses also underscore a knowledge gap regarding metadata standards, with many not familiar with them.

1. Documentation for Data/Code

The majority leans towards using README files for data/code documentation, with JupyterLab Notebooks or similar scripts also being a popular choice. However, a noticeable portion does not engage in any form of documentation.

- **README Files**: A major portion, 68.98%, use README files for documentation.
- **Data Processing Scripts**: 37.5% use JupyterLab Notebooks or similar scripts.
- **Lack of Documentation**: 11.57% don't document their data or code.

2. Version Management

Git-based platforms dominate versioning practices. Yet, many still resort to manual methods, emphasizing the mix of modern and traditional approaches in managing data/code versions.

- **Git Usage**: 67.13% rely on Git or Git-based platforms for managing versions.
- **Manual Management**: 49.54% manually change file/folder names for versioning (e.g., date, version number).
- **Built-in Tools**: 15.28% use tools built into storage/sharing solutions for versioning.

3. Naming Convention for Files/Folders

Almost half of the respondents adhere to some form of naming convention, showcasing an awareness of organization. Still, there's a segment that does not follow any structured naming approach.

- **Some Convention**: 47.68% follow a naming convention, whether personal, shared with collaborators, or standard in their field.
- **No Convention**: 14.35% don't adhere to any naming convention.

4. Knowledge of Metadata Standards

A significant number of respondents are either not familiar with or unsure about metadata standards, highlighting a potential area for educational outreach.

- **Aware**: 16.74% are familiar with metadata standards.
- **Unaware**: 54.42% don't know any metadata standards.
- **Uncertain**: 28.84% aren't sure what metadata standards entail.

5. Usage of Metadata Standards

Among those aware of metadata standards, there's a split between users and non-users. The variety of mentioned standards indicates diverse needs and practices in the community.

- **Users**: Among those who know metadata standards, 38.89% actually use them.
- **Non-Users**: In the same group, 61.11% don't use any metadata standards.
- **Variety**: Mentioned standards include OME, BIDS, NWB, among others.

E. Opening tools and data

Based on questions C1. to C3. And D1., the community displays a mixed approach towards openness. While a substantial number prefer open or mostly open formats and software, there's still a segment relying on proprietary or mostly proprietary options. Notably, many respondents indicated uncertainty about their practices, highlighting potential gaps in awareness or clarity.

1. Use/Production of Open Format Files or Open Source Code/Software

The EPFL research community shows a preference for open practices, but there's a notable presence of uncertainty regarding the adoption of open or proprietary formats and software.

- **Open Formats**: 28.43% exclusively use open formats, while 32.35% mainly utilize open formats.
- **Proprietary Formats**: Just 0.49% exclusively use proprietary formats, but 8.82% lean towards mostly proprietary formats.
- **Open Code/Software**: 23.53% exclusively use open code/software, and 46.08% mainly utilize open code/software.
- **Proprietary Code/Software**: Only 1.47% exclusively rely on proprietary code/software, while 10.29% lean mostly towards proprietary solutions.
- **Uncertainty**: A significant 23.04% are unsure about their practices concerning using open formats, and 12.75% are uncertain about using open code/software.

2. Software/Platform Usage for Data/Code

Collaborative platforms, computing environments, and IDEs dominate the software landscape, emphasizing the blend of tools researchers use for their data and code needs.

- **Collaborative Platforms**: 74.51% engage with platforms like GitHub or GitLab.
- **Computing Environments**: 66.18% use environments like Matlab or Jupyter Notebooks.
- **Integrated Development Environments (IDEs)**: 56.86% utilize IDEs like VisualStudio and RStudio.

3. Tools for Writing DMPs (Data Management Plans)

While a significant portion doesn't prioritize DMPs, those who do are split between basic tool needs and a desire for more specialized options, revealing an awareness gap in available tools.

- **DMP Relevance**: 57.84% don't see the need for a DMP.
- **Basic Tools**: Among DMP writers, 17.65% find a word processor sufficient.
- **Tool Awareness**: 16.18% desire more specialized tools but are unaware of their options.

4. Data/Code Sharing

Data and code sharing remains predominantly internal within research groups, but there's a growing trend towards external collaborations and public sharing, highlighting a broadening horizon of collaboration.

- **Internal Sharing**: Over half, 51.3%, share data/code only within their research group.
- **External Collaborative Sharing**: 48.19% share with academic partners in other institutions.
- **Public Sharing**: A small portion are committed to public sharing, especially upon research publication.

F. Legal issues and Cold data

Based on questions D5. and from E1. to E3., at EPFL, while a minority navigate the complexities of handling sensitive data, those who do prioritize ethical considerations. Platforms like GitHub and GitLab have emerged as the dominant avenues for disseminating and preserving research. In the realm of licensing, the MIT and GPL/LGPL licenses are the preferred choices for software. However, when it comes to broader content and data, the Creative Commons licenses, particularly CC-BY and CC-BY-NC, stand out. Notably, there's a discernible knowledge gap among researchers about licensing and best practices. This underscores the importance of targeted institutional guidance and further education, especially concerning the nuances of licensing and its broader impact.

1. Working with Personal/Sensitive Data and Ethics Review

While a minority of respondents work with sensitive data, there's a clear adherence to ethical considerations among those who do, with many undergoing or planning to undergo ethics review. Yet, a significant majority do not engage with personal or sensitive data in their research.

- **Sensitive Data Involvement**: 18.82% work or plan to work with personal or sensitive data.
- **Ethics Review**: 11.39% have their projects in the ethics review process or already approved, while 2.97% don't request any review.
- **No Sensitive Data**: A dominant 71.78% do not work with personal/sensitive data.

2. Data/Code Publication Locations

Code sharing platforms like GitHub and GitLab stand out as the predominant choice for publishing, with data repositories also playing a significant role. However, there's a segment of the population that remains uncertain about publication or does not intend to publish at all.

- **Code Sharing Platforms**: A significant 61.88% publish or plan to publish on platforms like GitHub, GitLab, or Bitbucket.
- **Data Repositories**: 30.2% opt for data repositories such as Zenodo or MaterialsCloud.
- **Unsure or Never**: 15.84% are unsure about where to publish, and 11.88% have never published and don't plan to.

3. Licensing for Data/Code Publication

The community shows a preference for both the MIT License and GPL/LGPL for software, but Creative Commons licenses, especially CC-BY and CC-BY-NC, are also prominently chosen, particularly for data and content dissemination. A significant segment, however, is unfamiliar with licensing concepts, underscoring the need for education on this front.

- **Creative Commons**: Specifically, CC-BY and CC-BY-NC are prominently chosen, together accounting for 35.96% of responses, highlighting their significance in data and content dissemination.
- **MIT License**: 29.78% publish or plan to use the MIT License.
- **GPL/LGPL**: 23.03% use or plan to use the GPL or LGPL licenses.
- **Lack of Knowledge**: A notable 31.46% are unaware of what licenses are.

4. Data/Code Archiving Locations

Code sharing platforms and institutional servers are the top choices for long-term archiving, ensuring accessibility and preservation of research outputs. However, as with publication, there's an element of uncertainty, highlighting the need for clearer guidelines or resources on archiving practices.

- **Code Sharing Platforms**: 55.94% archive or plan to archive on platforms like GitHub, GitLab, or Bitbucket.
- **Institutional/Faculty Servers**: 29.21% opt for institutional or faculty servers for archiving.
- **Data Repositories**: 25.25% choose data repositories such as Zenodo or MaterialsCloud for long-term archiving.

G. Services and support

While most EPFL researchers navigate RDM topics autonomously, many rely on peer insights, especially for code sharing and versioning. Online resources and documentation would be the preferred support methods, followed by in-person group consultations. However, many researchers are unaware or unsure about the array of services and tools available to them. Suggestions lean towards more hands-on, tailored training, clear guidelines, and streamlining the available resources.

1. EPFL Service/Support Consultation for Various Data/Code Management Topics

While most researchers handle data and code management topics independently, colleagues serve as a significant source of information. Formal EPFL structures, however, are less frequently consulted.

- **No Need for Support**: For many tasks, a significant number of researchers feel no need to consult, with 67.82% handling Data Management Plans on their own.
- **Code Sharing/Versioning**: Code sharing and versioning stands out as a task where colleagues are frequently consulted, with 40.91% turning to their peers.
- **Library RDM Team**: The Library RDM team is notably consulted for licensing and copyright issues by 5.13% of respondents.
- **General Uncertainty**: A considerable number, 35.71% for example in publishing data/code, turning to colleagues for guidance.
- **Formal Structures Underutilized**: Very few consult the EPFL Data Champions or the Research Office (ReO) across most tasks.

2. Preferred Support Type

Online resources are popular, but there's a significant preference for in-person and interactive support methods.

- **Web-based:** 59.9% prefer online documentation.
- **In-person:** 33.85% favor group/lab consulting, while 30.21% like one-on-one consultations.
- **Training:** 27.6% are receptive to on-site training, and 23.96% utilize e-learning platforms.

3. Suggestions for RDM Support or Training

While some respondents are satisfied with the existing EPFL services, others express a need for more practical guidance. There is a strong demand for enhanced infrastructure, particularly for sharing large datasets. Many respondents also seek ready-to-use templates and clearer institutional guidelines on various aspects of research data management. Additionally, there is a clear need for comprehensive RDM training, both at the start of research projects and on an ongoing basis.

Training & Awareness:

- Introduce comprehensive RDM training sessions on data protection and ORD benefits.
- Develop mandatory RDM onboarding for new PhD students.
- Offer a mix of in-person and online training modules.

Tool & Infrastructure Needs:

- Enhance data sharing infrastructure, especially for large datasets.
- Improve reliability and communication regarding platforms like EPFL's GitLab.
- Educate researchers on available tools and platforms for better utilization.
- Provide clarity on platforms and conditions for free data/code sharing.

Simplification & Tailored Solutions:

- Design faculty-specific RDM solutions, moving away from a generic approach.
- Streamline data management processes, avoiding over-complication.
- Strive to provide more specific, actionable advice rather than generic guidance.

Templates & Guidance:

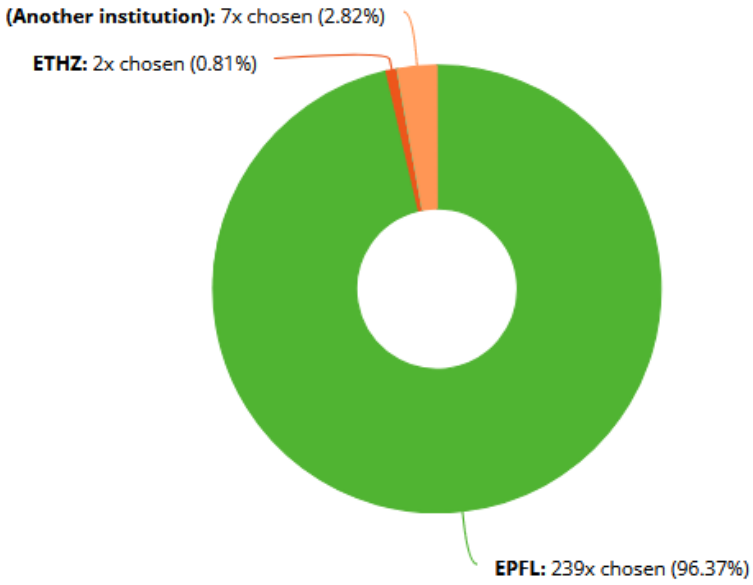
- Provide ready-to-use templates, particularly for Data Management Plans (DMPs).
- Offer clear institutional guidelines on data sharing, licensing, and anonymity.
- Continuously gather feedback on current services to ensure they meet researchers' needs.
- Establish comprehensive online documentation with an option for expert consultation.

Appendix – Results overview and graphs

What is your main institutional affiliation?

Number of responses: 248

Answer	Times Chosen	Percentage
Eawag	0	0%
EPFL	239	96.37%
ETHZ	2	0.81%
FHNW	0	0%
UNIL / CHUV	0	0%
(Another institution)	7	2.82%



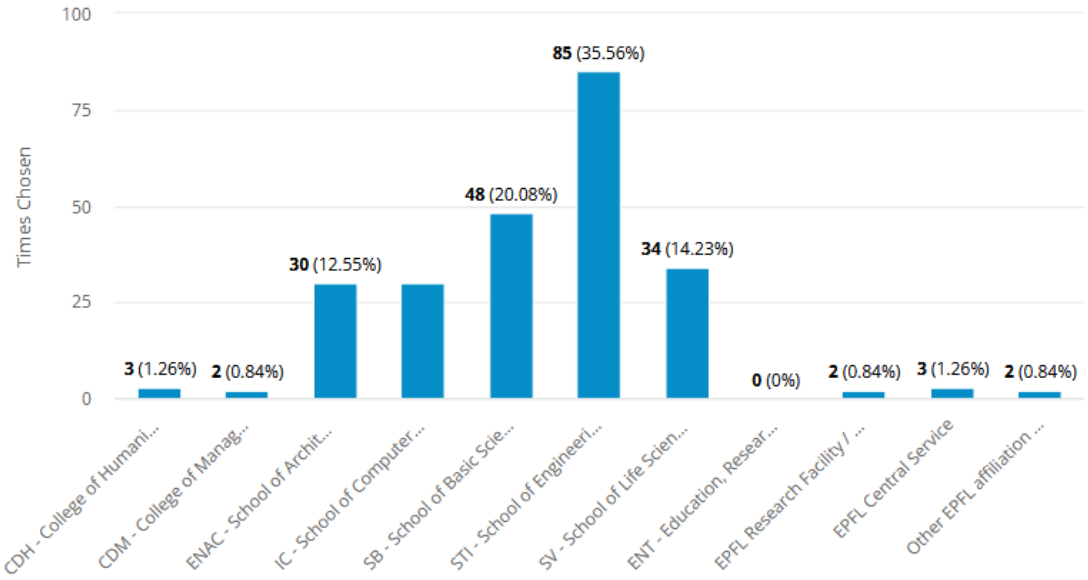
Please, indicate your institutional affiliation

Number of responses: 7

What is your main EPFL affiliation?

Number of responses: 239

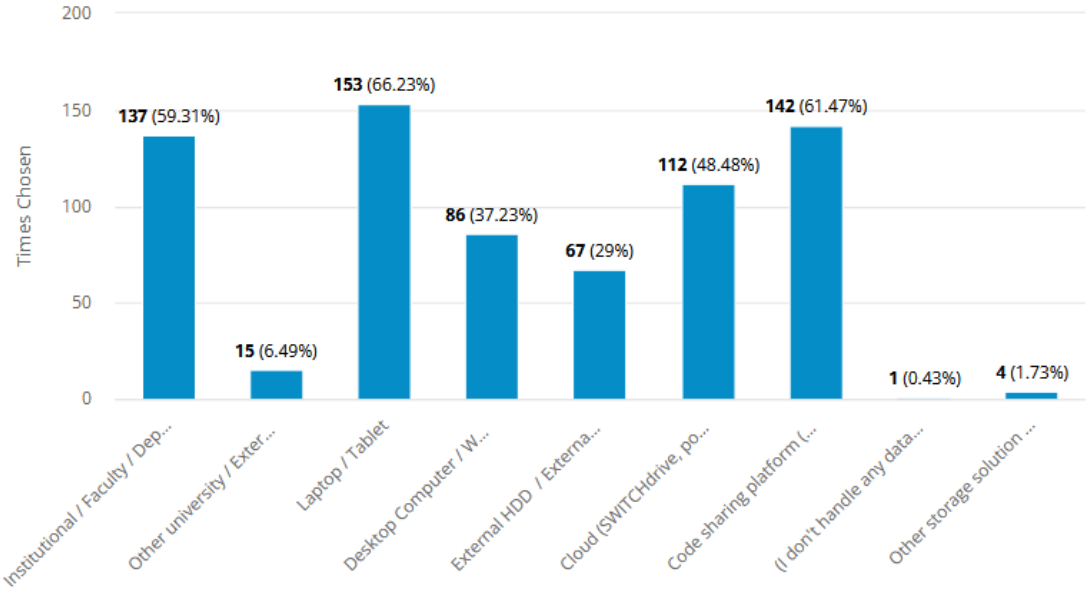
Answer	Times Chosen	Percentage
CDH - College of Humanities	3	1.26%
CDM - College of Management of Technology	2	0.84%
ENAC - School of Architecture, Civil and Environmental Engineering	30	12.55%
IC - School of Computer and Communication Sciences	30	12.55%
SB - School of Basic Sciences	48	20.08%
STI - School of Engineering	85	35.56%
SV - School of Life Sciences	34	14.23%
ENT - Education, Research, Innovation and other Centers	0	0%
EPFL Research Facility / Platform	2	0.84%
EPFL Central Service	3	1.26%
Other EPFL affiliation ...	2	0.84%



A1. Where do you store your research data / code, even temporarily?

Number of responses: 231

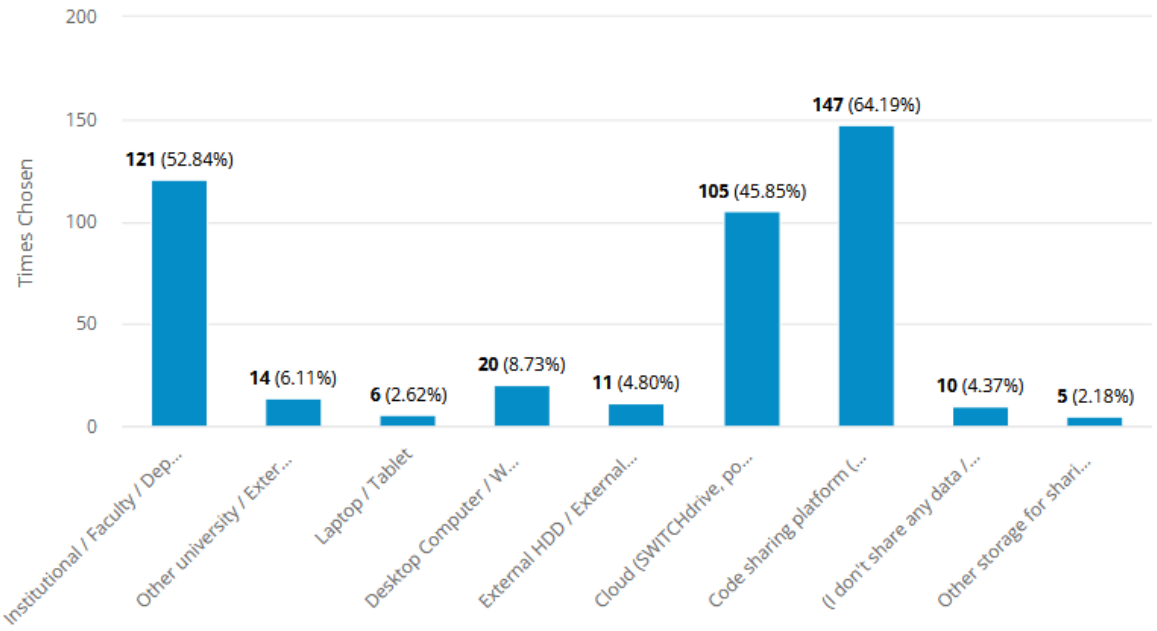
Answer	Times Chosen	Percentage
Institutional / Faculty / Department servers	137	59.31%
Other university / External facility servers	15	6.49%
Laptop / Tablet	153	66.23%
Desktop Computer / Workstation	86	37.23%
External HDD / External SSD	67	29%
Cloud (SWITCHdrive, polybox, Google Drive, OneDrive, ...)	112	48.48%
Code sharing platform (GitHub, GitLab, Bitbucket, ...)	142	61.47%
(I don't handle any data / code)	1	0.43%
Other storage solution ...	4	1.73%



A2. On which storage solution(s) do you have a collaborative access with others to share data / code?

Number of responses: 229

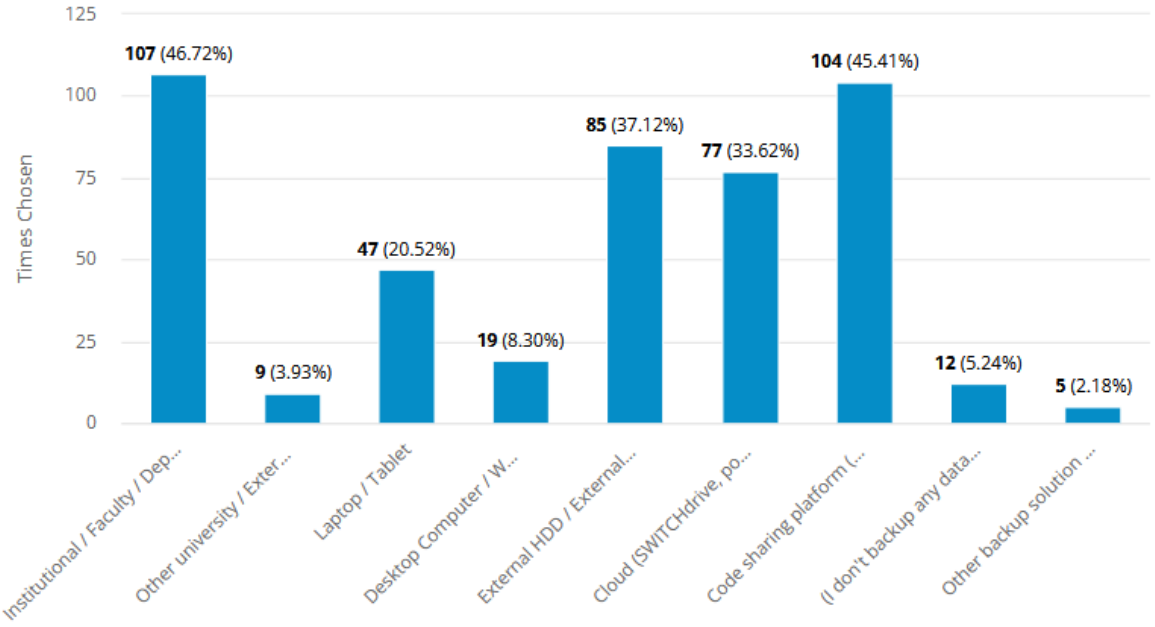
Answer	Times Chosen	Percentage
Institutional / Faculty / Department servers	121	52.84%
Other university / External facility servers	14	6.11%
Laptop / Tablet	6	2.62%
Desktop Computer / Workstation	20	8.73%
External HDD / External SSD	11	4.8%
Cloud (SWITCHdrive, polybox, Google Drive, OneDrive, ...)	105	45.85%
Code sharing platform (GitHub, GitLab, Bitbucket, ...)	147	64.19%
(I don't share any data / code)	10	4.37%
Other storage for sharing ...	5	2.18%



A3. Where do you backup your research data / code?

Number of responses: 229

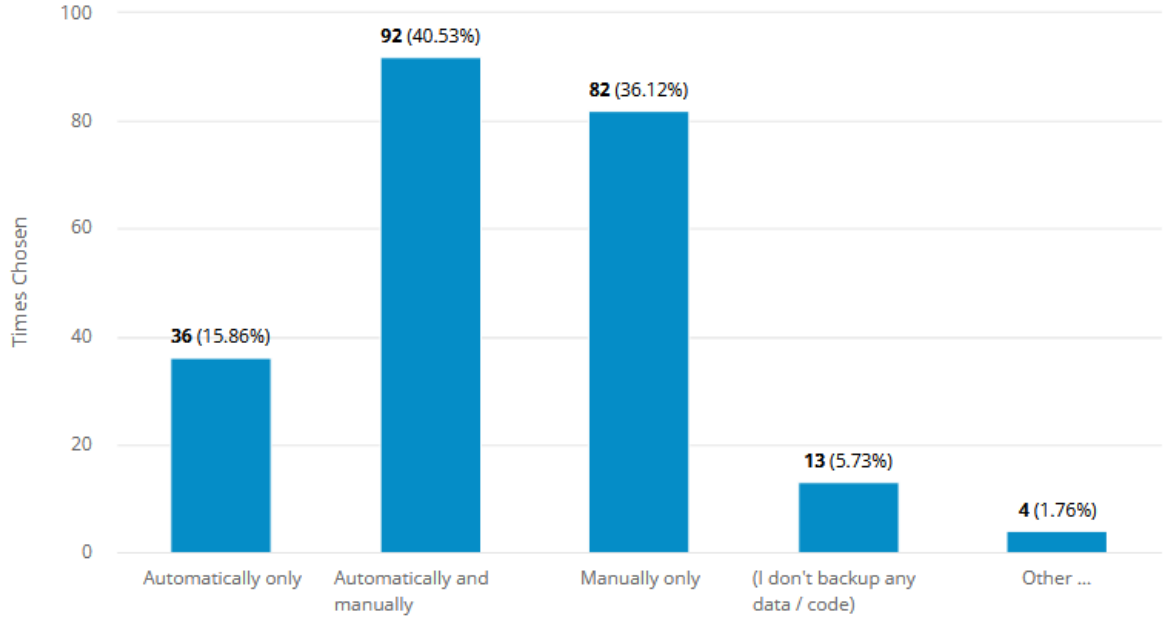
Answer	Times Chosen	Percentage
Institutional / Faculty / Department servers	107	46.72%
Other university / External facility servers	9	3.93%
Laptop / Tablet	47	20.52%
Desktop Computer / Workstation	19	8.3%
External HDD / External SSD	85	37.12%
Cloud (SWITCHdrive, polybox, Google Drive, OneDrive, ...)	77	33.62%
Code sharing platform (GitHub, GitLab, Bitbucket, ...)	104	45.41%
(I don't backup any data / code)	12	5.24%
Other backup solution ...	5	2.18%



A4. How do you backup your research data/code?

Number of responses: 227

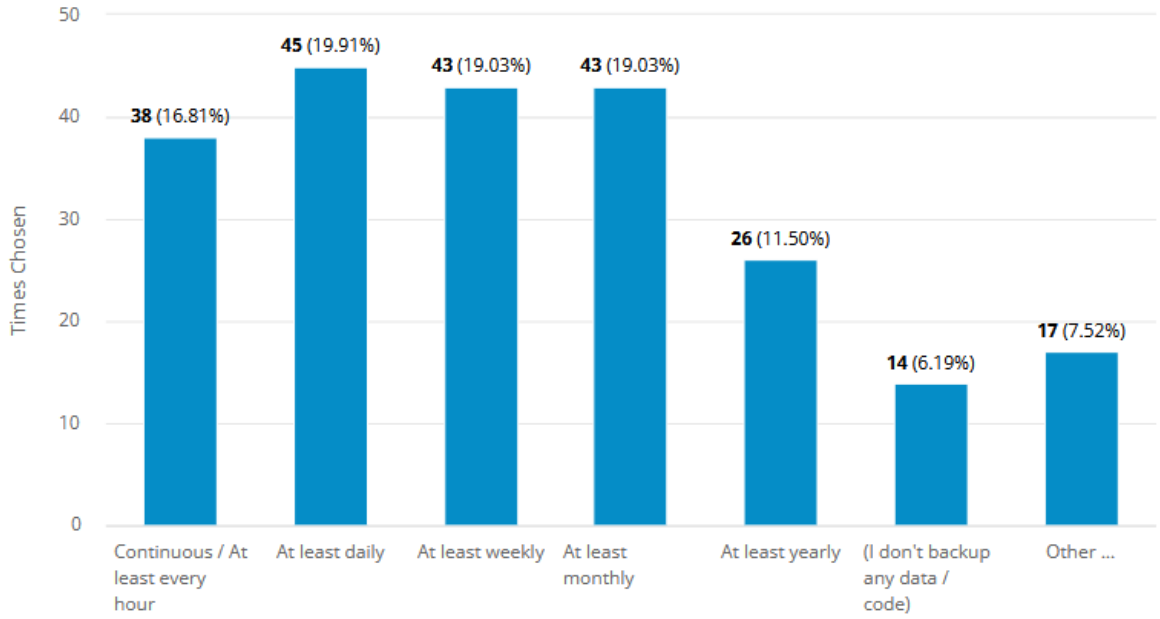
Answer	Times Chosen	Percentage
Automatically only	36	15.86%
Automatically and manually	92	40.53%
Manually only	82	36.12%
(I don't backup any data / code)	13	5.73%
Other ...	4	1.76%



A5. How frequently do you perform backups?

Number of responses: 226

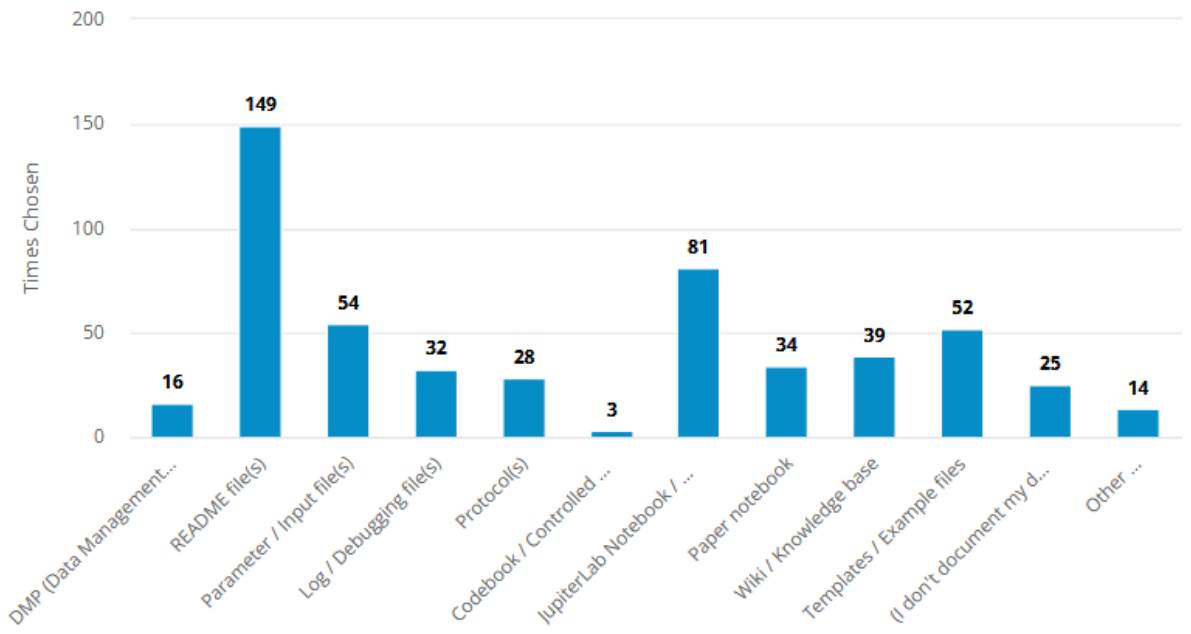
Answer	Times Chosen	Percentage
Continuous / At least every hour	38	16.81%
At least daily	45	19.91%
At least weekly	43	19.03%
At least monthly	43	19.03%
At least yearly	26	11.5%
(I don't backup any data / code)	14	6.19%
Other ...	17	7.52%



B1. Which kind of documentation do you produce to accompany your data / code?

Number of responses: 216

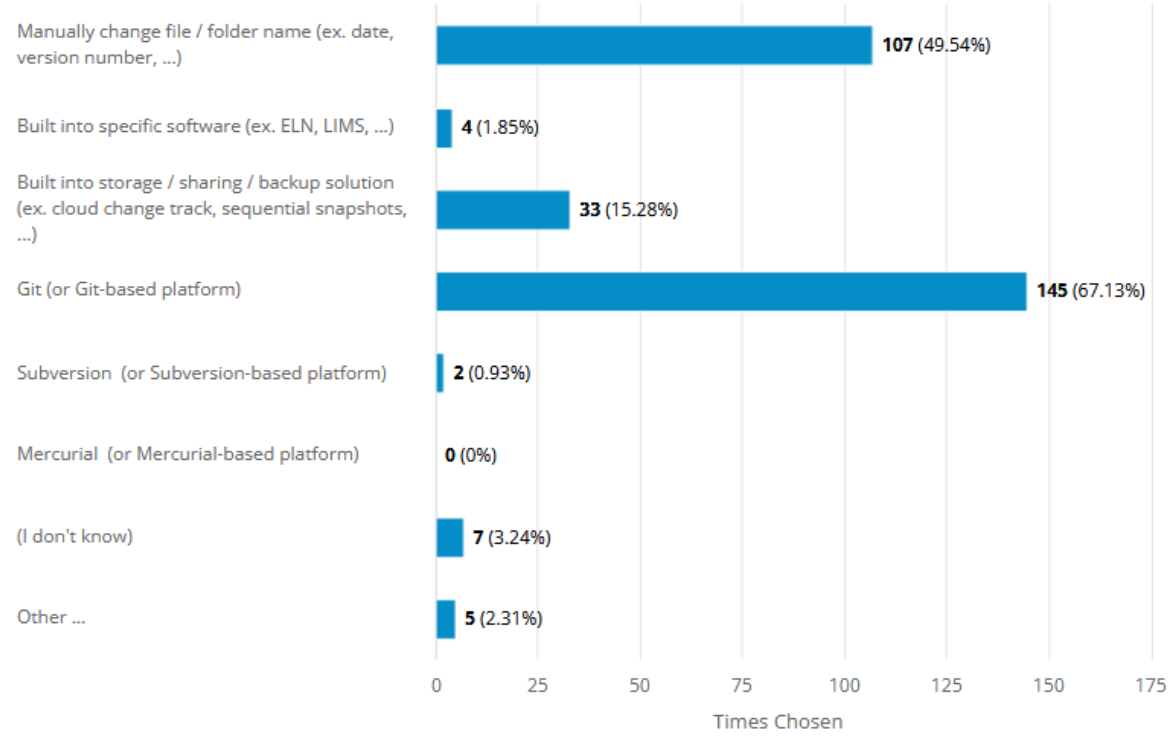
Answer	Times Chosen	Percentage
DMP (Data Management Plan)	16	7.41%
README file(s)	149	68.98%
Parameter / Input file(s)	54	25%
Log / Debugging file(s)	32	14.81%
Protocol(s)	28	12.96%
Codebook / Controlled vocabulary	3	1.39%
JupiterLab Notebook / Other data processing scripts	81	37.5%
Paper notebook	34	15.74%
Wiki / Knowledge base	39	18.06%
Templates / Example files	52	24.07%
(I don't document my data / code)	25	11.57%
Other ...	14	6.48%



B2. How do you manage versions of your data / code?

Number of responses: 216

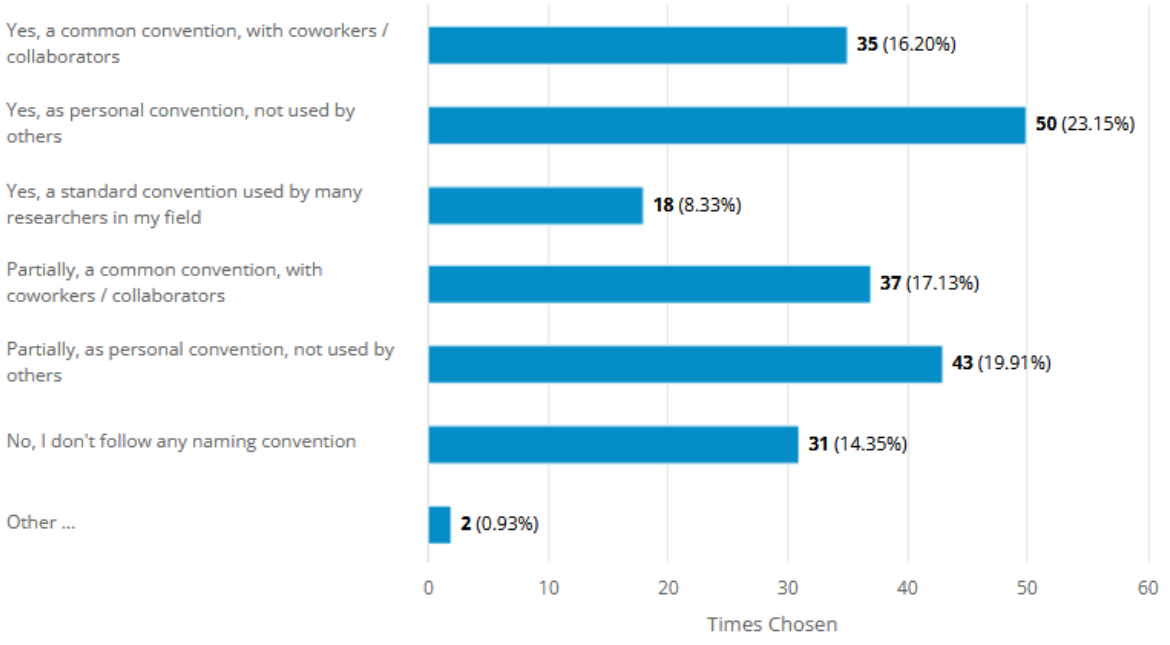
Answer	Times Chosen	Percentage
Manually change file / folder name (ex. date, version number, ...)	107	49.54%
Built into specific software (ex. ELN, LIMS, ...)	4	1.85%
Built into storage / sharing / backup solution (ex. cloud change track, sequential snapshots, ...)	33	15.28%
Git (or Git-based platform)	145	67.13%
Subversion (or Subversion-based platform)	2	0.93%
Mercurial (or Mercurial-based platform)	0	0%
(I don't know)	7	3.24%
Other ...	5	2.31%



B3. Do you follow a naming convention for files / folders?

Number of responses: 216

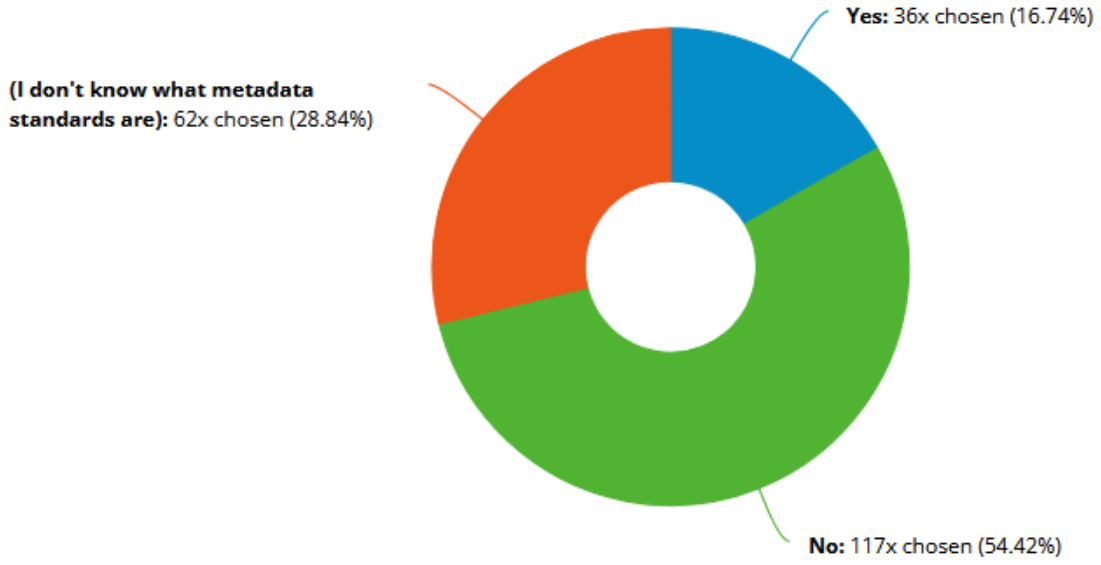
Answer	Times Chosen	Percentage
Yes, a common convention, with coworkers / collaborators	35	16.2%
Yes, as personal convention, not used by others	50	23.15%
Yes, a standard convention used by many researchers in my field	18	8.33%
Partially, a common convention, with coworkers / collaborators	37	17.13%
Partially, as personal convention, not used by others	43	19.91%
No, I don't follow any naming convention	31	14.35%
Other ...	2	0.93%



B4. Do you know any metadata standards?

Number of responses: 215

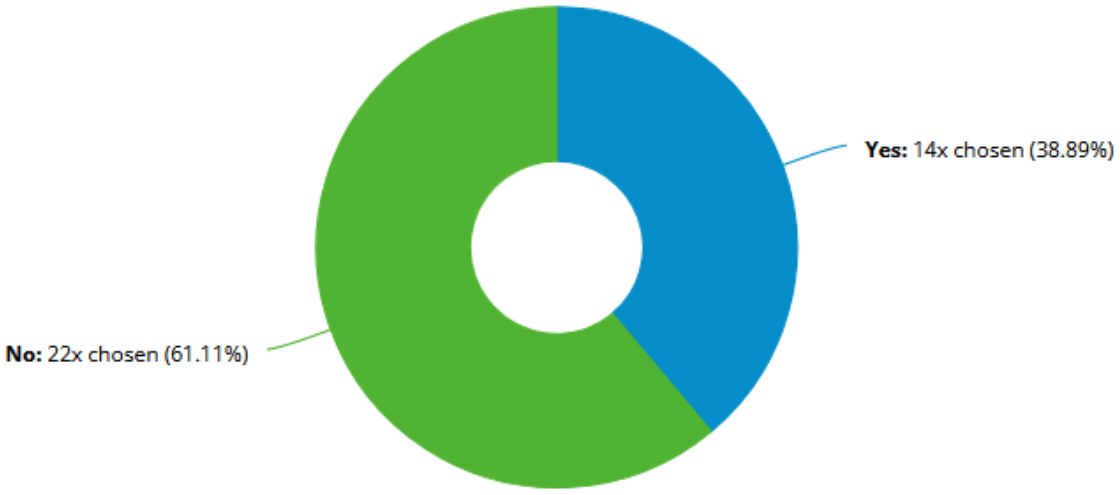
Answer	Times Chosen	Percentage
Yes	36	16.74%
No	117	54.42%
(I don't know what metadata standards are)	62	28.84%



B5. Do you use any metadata standards?

Number of responses: 36

Answer	Times Chosen	Percentage
Yes	14	38.89%
No	22	61.11%



B6. Please, indicate the exact name of your metadata standard(s)?

Number of responses: 14

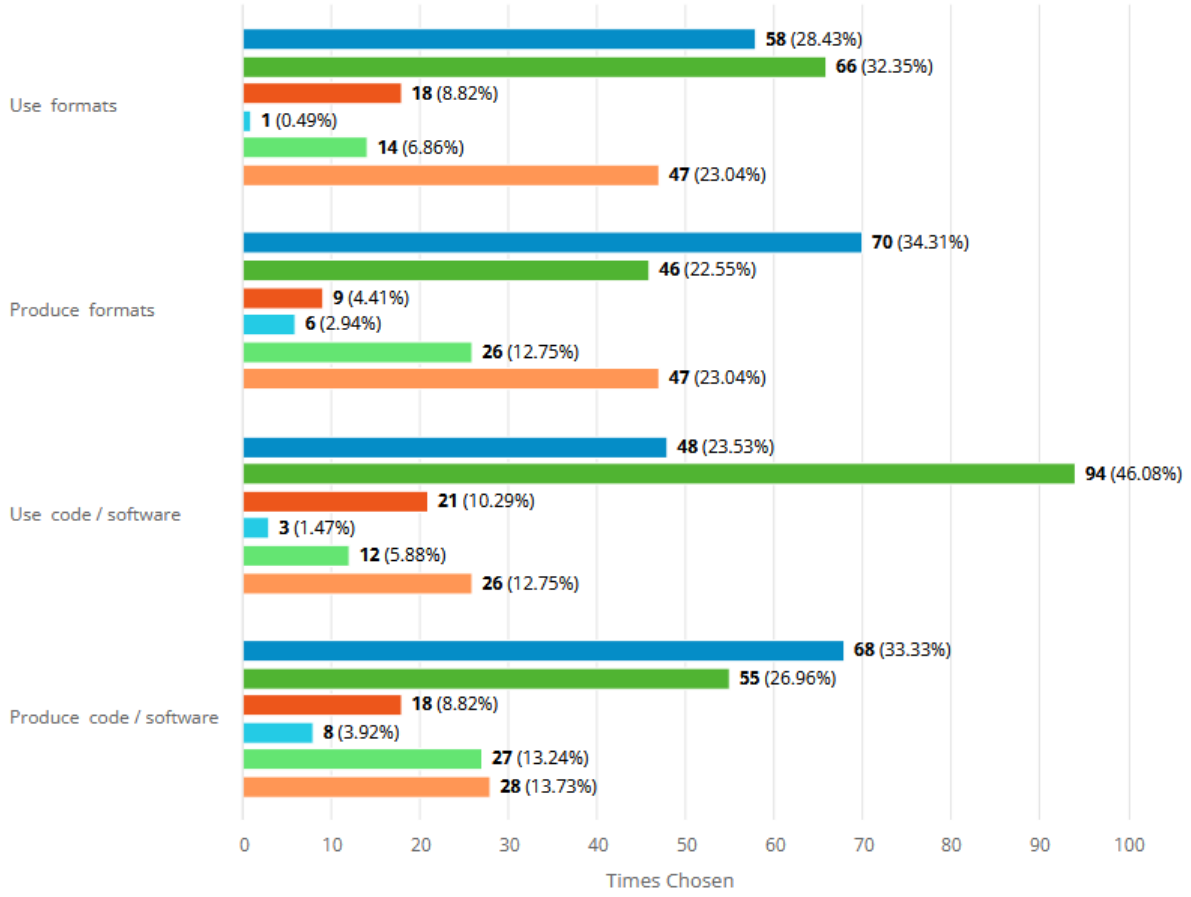
Text answers:

- Standards respectant ISO9001 en France avec un PGD
- Il n'y a pas de standard de métadonnées pour les données de séquençage mais nous utilisons celles qui sont requises pour publier sur GEO (https://www.ncbi.nlm.nih.gov/geo/)
- OME for images
- NWB
- OME, Open Microscopy Environment
- STAC
- BIDS
- https://packaging.python.org/en/latest/specifications/core-metadata/#core-metadata
- OAI-PMH, DataCite, CIF
- BIDS and Frictionless data for data, boutique (and if it counts python standard metadata) for code
- Brain-Score
- Nwb
- Optimade
- PDF properties for documents produced by LaTeX

C1. Do you use / produce open format files or open-source code / software?

Number of responses: 204

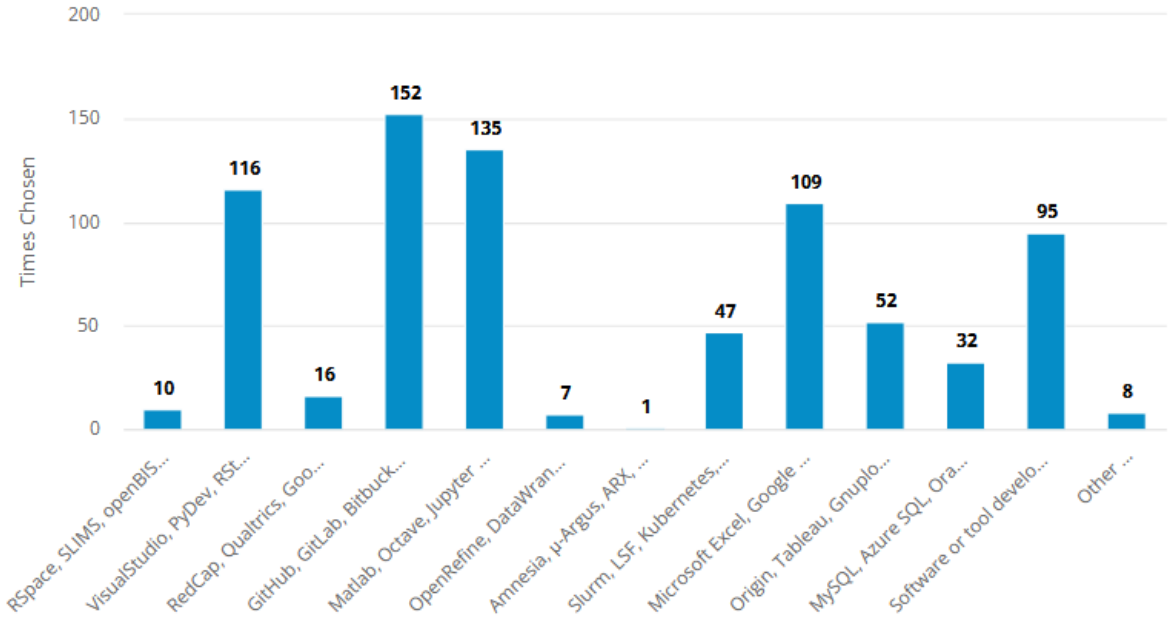
	All open 1		Mostly open 2		Mostly proprietary 3		Only proprietary 4		(Doesn't apply to me) 5		(Don't know) 6				
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%		Ø	±
Use formats	58	28.43 %	66	32.35 %	18	8.82 %	1	0.49 %	14	6.86 %	47	23.04 %		2.94	1.95
Produce formats	70	34.31 %	46	22.55 %	9	4.41 %	6	2.94 %	26	12.75 %	47	23.04 %		3.06	2.05
Use code / software	48	23.53 %	94	46.08 %	21	10.29 %	3	1.47 %	12	5.88 %	26	12.75 %		2.58	1.62
Produce code / software	68	33.33 %	55	26.96 %	18	8.82 %	8	3.92 %	27	13.24 %	28	13.73 %		2.78	1.83



C2. What type of software / platform do you use while working with data / code?

Number of responses: 204

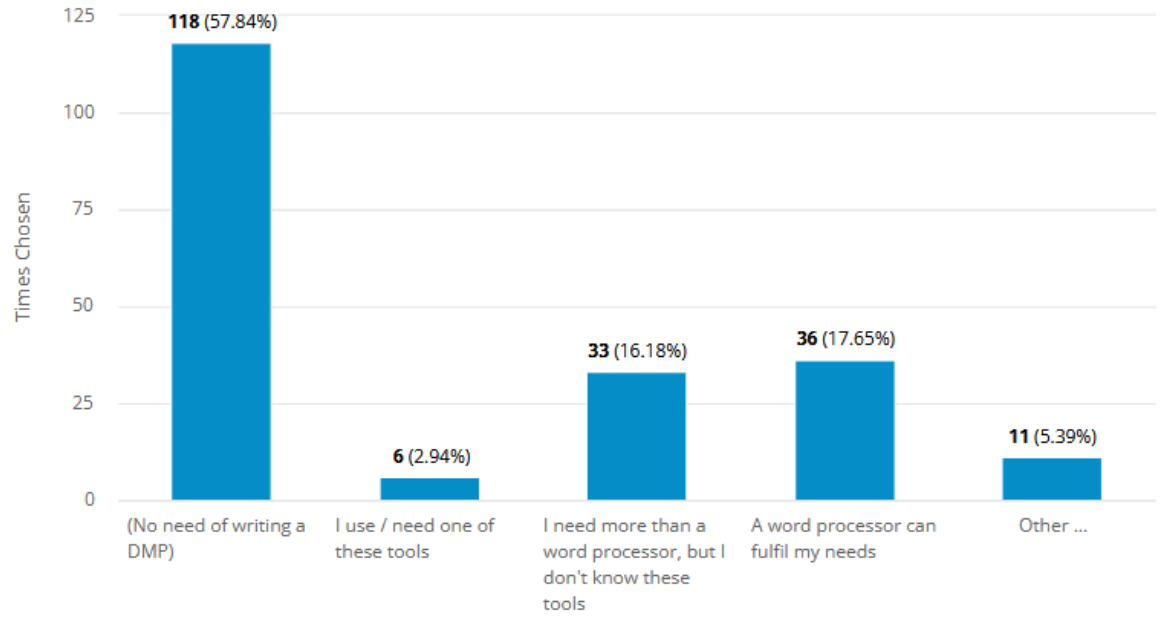
Answer	Times Chosen	Percentage
RSpace, SLIMS, openBIS, eln.epfl.ch, or other ELN or LIMS	10	4.9%
VisualStudio, PyDev, RStudio, or other IDE	116	56.86%
RedCap, Qualtrics, Google Forms, or other surveying platform	16	7.84%
GitHub, GitLab, Bitbucket, or other code sharing platform	152	74.51%
Matlab, Octave, Jupyter Notebooks, or other computing environment	135	66.18%
OpenRefine, DataWrangler, Trifacta, or other tools for data cleaning or preprocessing	7	3.43%
Amnesia, µ-Argus, ARX, or other data anonymization tool	1	0.49%
Slurm, LSF, Kubernetes, or other Computational Resource Management or Scheduling tool	47	23.04%
Microsoft Excel, Google Sheet, LibreOffice Calc, or other spreadsheet software	109	53.43%
Origin, Tableau, Gnuplot , or other graph plotting and data visualization software	52	25.49%
MySQL, Azure SQL, Oracle RDBMS, or other database software or system	32	15.69%
Software or tool developed in-house with python, C, javascript, R, or other programming language	95	46.57%
Other ...	8	3.92%



C3. Do you use / need a specific tool for writing DMPs (Data Management Plans)?

Number of responses: 204

Answer	Times Chosen	Percentage
(No need of writing a DMP)	118	57.84%
I use / need one of these tools	6	2.94%
I need more than a word processor, but I don't know these tools	33	16.18%
A word processor can fulfil my needs	36	17.65%
Other ...	11	5.39%



"Other ..." text answers:

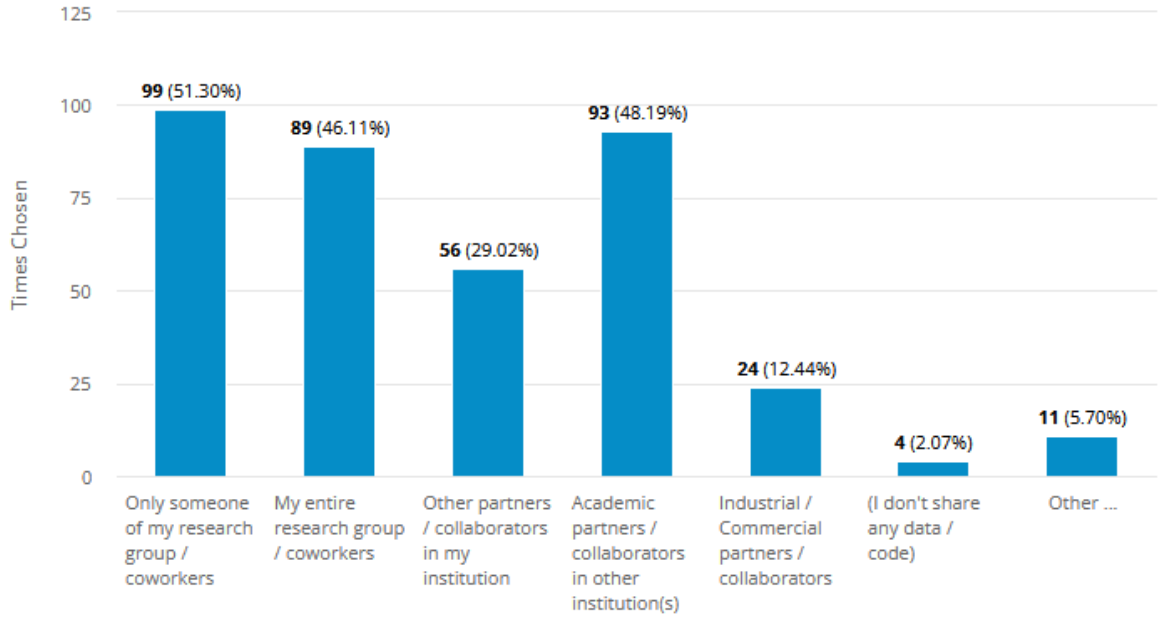
- I donk know what a DMP is

- I was told to not write a DMP
- Do not know what DMPs are
- Don't know what DMP are
- I have never used any but would be interested to learn about them
- I don't know what it is
- I don't know what DMP is
- LaTeX
- Je ne sais pas ce qu'est un DMP
- https://www.materialscloud.org/dmp
- Je ne sais pas

D1. With whom do you share your research data / code?

Number of responses: 193

Answer	Times Chosen	Percentage
Only someone of my research group / coworkers	99	51.3%
My entire research group / coworkers	89	46.11%
Other partners / collaborators in my institution	56	29.02%
Academic partners / collaborators in other institution(s)	93	48.19%
Industrial / Commercial partners / collaborators	24	12.44%
(I don't share any data / code)	4	2.07%
Other ...	11	5.7%

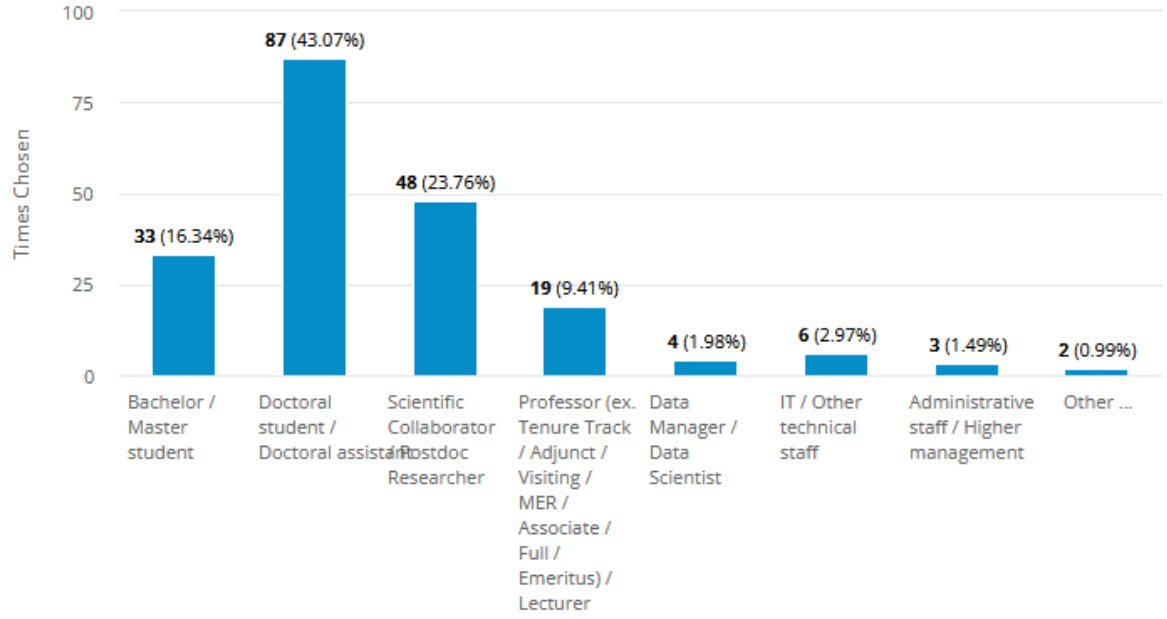


D2. What is your role?

Number of responses: 202

Answer	Times Chosen	Percentage
Bachelor / Master student	33	16.34%
Doctoral student / Doctoral assistant	87	43.07%
Scientific Collaborator / Postdoc Researcher	48	23.76%
Professor (ex. Tenure Track / Adjunct / Visiting / MER / Associate / Full / Emeritus) / Lecturer	19	9.41%

Data Manager / Data Scientist	4	1.98%
IT / Other technical staff	6	2.97%
Administrative staff / Higher management	3	1.49%
Other ...	2	0.99%

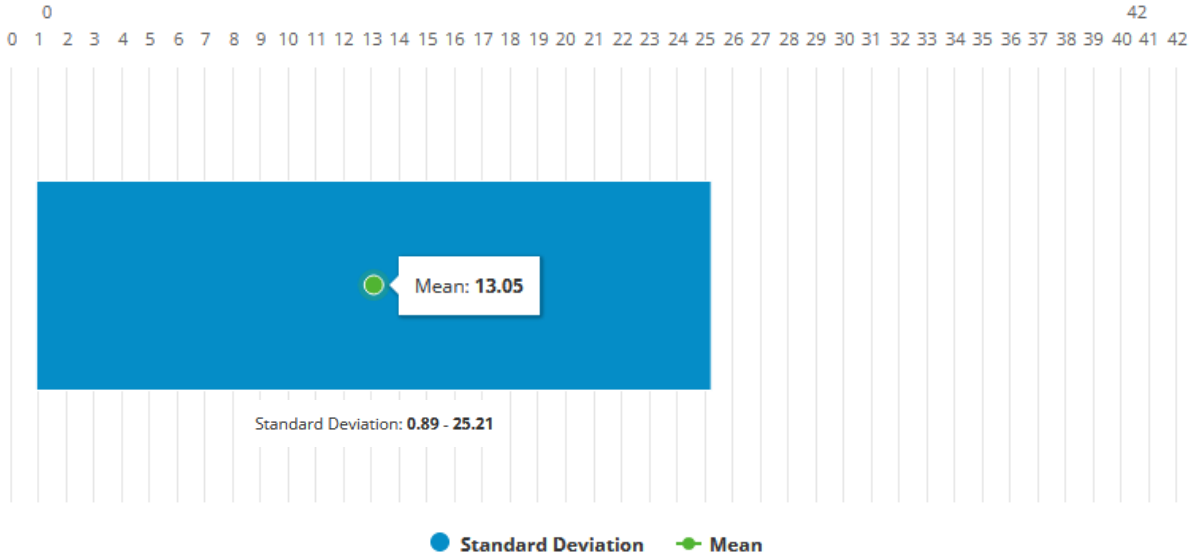


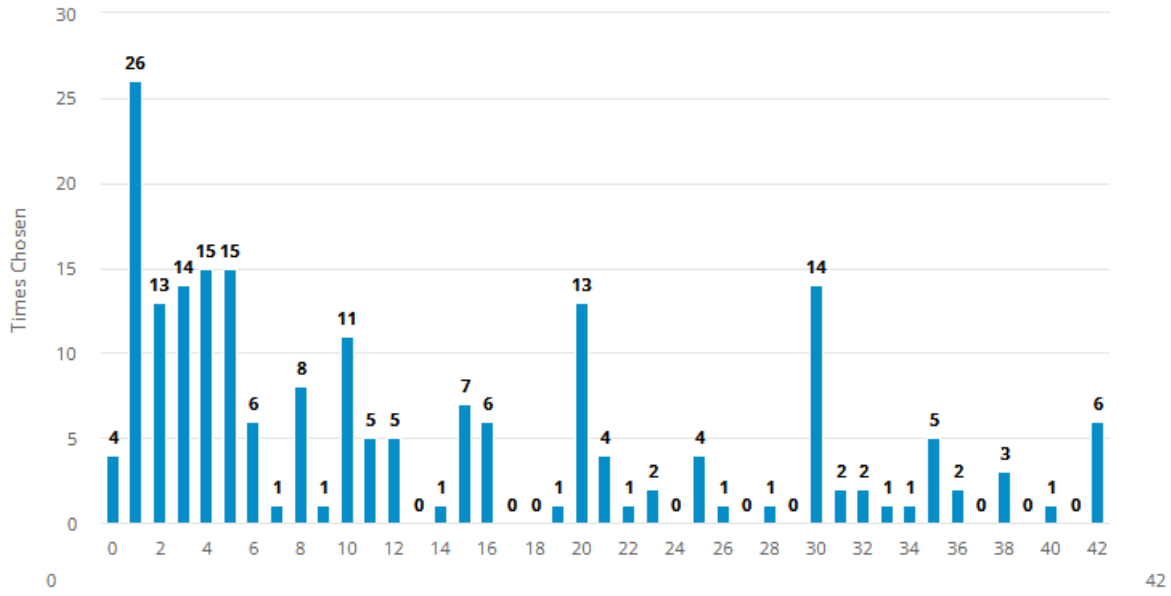
D3. In average, how many hours of work do you put every week into managing data / code? (Estimate)

Number of responses: 202

Slider Position	Times Chosen	Percentage
0 (0)	4	1.98%
1	26	12.87%
2	13	6.44%
3	14	6.93%
4	15	7.43%
5	15	7.43%
6	6	2.97%
7	1	0.5%
8	8	3.96%
9	1	0.5%
10	11	5.45%
11	5	2.48%
12	5	2.48%
13	0	0%
14	1	0.5%
15	7	3.47%
16	6	2.97%
17	0	0%
18	0	0%
19	1	0.5%
20	13	6.44%
21	4	1.98%
22	1	0.5%
23	2	0.99%
24	0	0%
25	4	1.98%
26	1	0.5%
27	0	0%
28	1	0.5%

29	0	0%
30	14	6.93%
31	2	0.99%
32	2	0.99%
33	1	0.5%
34	1	0.5%
35	5	2.48%
36	2	0.99%
37	0	0%
38	3	1.49%
39	0	0%
40	1	0.5%
41	0	0%
42 (42)	6	2.97%
Ø	13.05	
±	12.16	

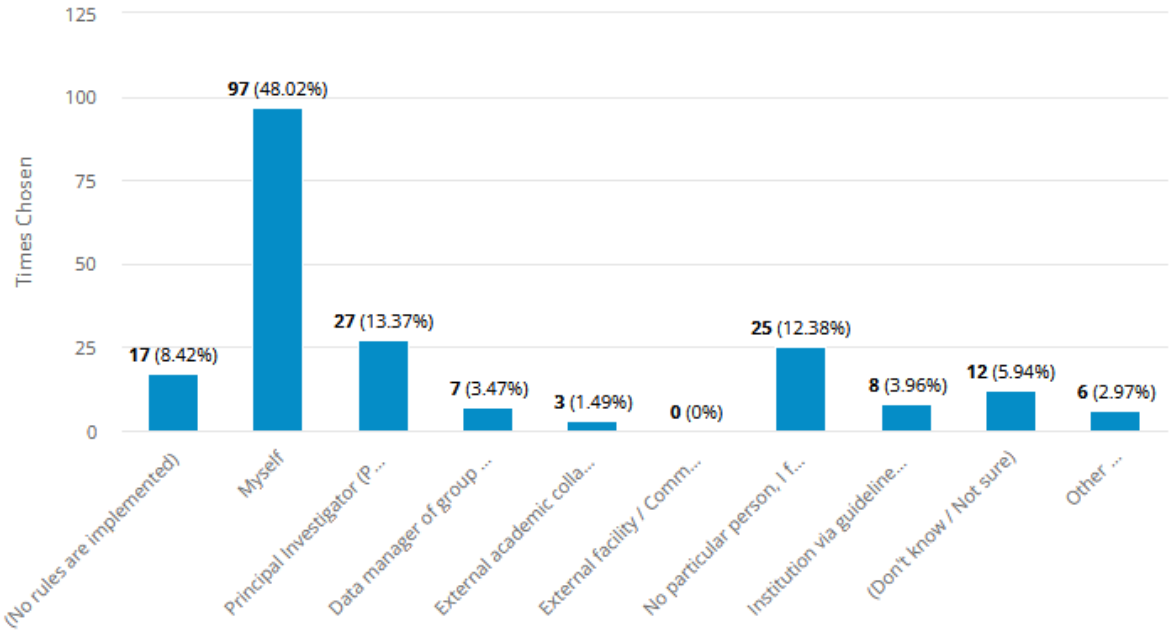




D4. Who sets for you the rules for managing your data / code?

Number of responses: 202

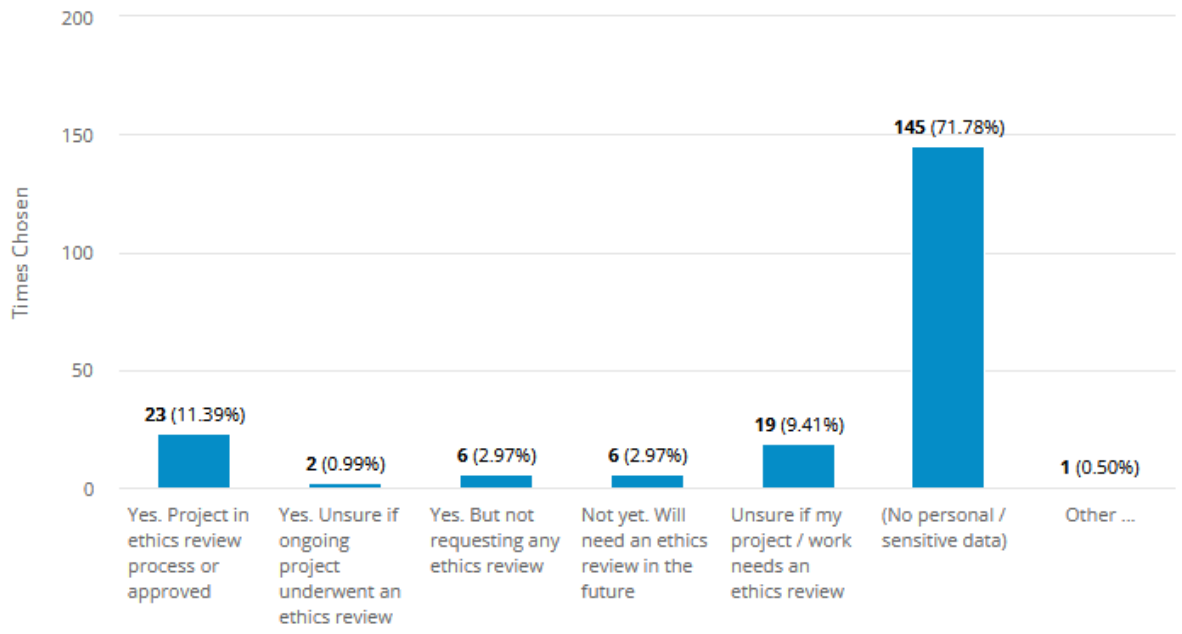
Answer	Times Chosen	Percentage
(No rules are implemented)	17	8.42%
Myself	97	48.02%
Principal Investigator (PI) / Supervisor	27	13.37%
Data manager of group / unit / consortium	7	3.47%
External academic collaborator (in other research group or other university, hospital, ...)	3	1.49%
External facility / Commercial partner	0	0%
No particular person, I follow the historical conventions of my group / unit / consortium	25	12.38%
Institution via guidelines / policy	8	3.96%
(Don't know / Not sure)	12	5.94%
Other ...	6	2.97%



D5. Do you work with personal / sensitive data? And has your project undergone an official ethics review?

Number of responses: 202

Answer	Times Chosen	Percentage
Yes. Project in ethics review process or approved	23	11.39%
Yes. Unsure if ongoing project underwent an ethics review	2	0.99%
Yes. But not requesting any ethics review	6	2.97%
Not yet. Will need an ethics review in the future	6	2.97%
Unsure if my project / work needs an ethics review	19	9.41%
(No personal / sensitive data)	145	71.78%
Other ...	1	0.5%



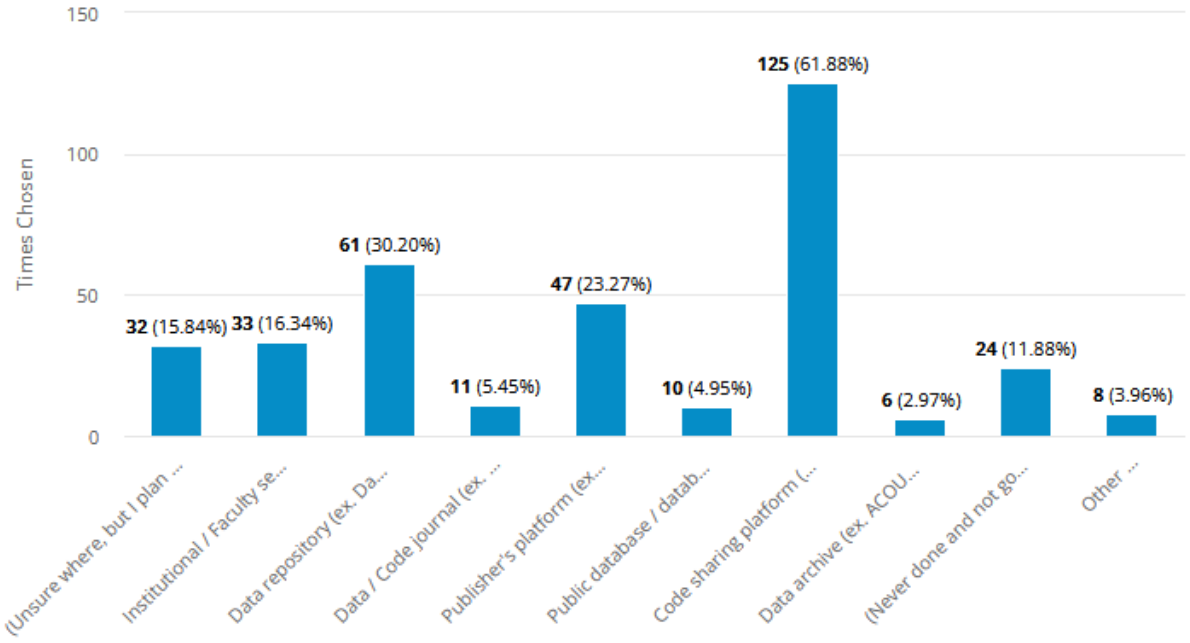
"Other ..." text answers:

- One project that touches PII reviewed and confirmed; all others don't touch PII

E1. Where do you publish / disseminate your data / code (or plan to)?

Number of responses: 202

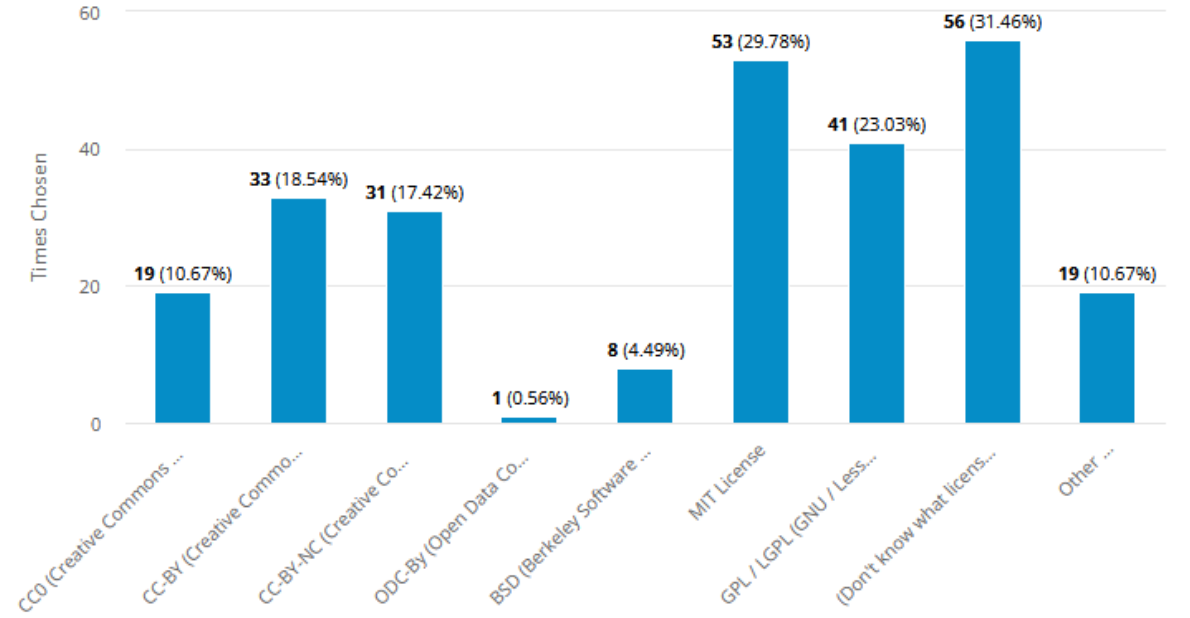
Answer	Times Chosen	Percentage
(Unsure where, but I plan to)	32	15.84%
Institutional / Faculty server or webpage	33	16.34%
Data repository (ex. DaSCH, EnviDat, ERIC, ETH Research Collection, MaterialsCloud, SWISSUBase, Yareta, Zenodo, ...)	61	30.2%
Data / Code journal (ex. Data in Brief, JOSS, ...)	11	5.45%
Publisher's platform (ex. supplementary material)	47	23.27%
Public database / databank (ex. Genbank, RECIFS, ...)	10	4.95%
Code sharing platform (ex. Bitbucket, GitHub, GitLab, ...)	125	61.88%
Data archive (ex. ACOUA, DaSCH, ETH Data Archive, OLOS, SWISSUbase, ...)	6	2.97%
(Never done and not going to)	24	11.88%
Other ...	8	3.96%



E2. Under what license(s) do you publish / disseminate your data / code (or plan to)?

Number of responses: 178

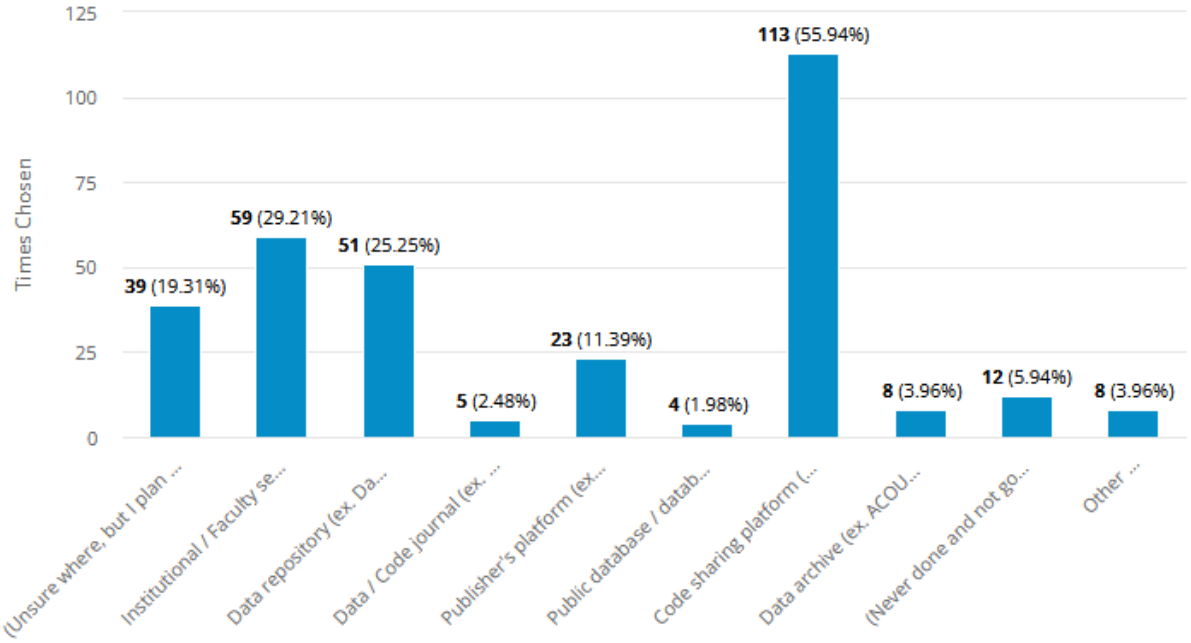
Answer	Times Chosen	Percentage
CC0 (Creative Commons Universal)	19	10.67%
CC-BY (Creative Commons Attribution)	33	18.54%
CC-BY-NC (Creative Commons Attribution-NonCommercial)	31	17.42%
ODC-By (Open Data Commons Attribution)	1	0.56%
BSD (Berkeley Software Distribution)	8	4.49%
MIT License	53	29.78%
GPL / LGPL (GNU / Lesser General Public License)	41	23.03%
(Don't know what licenses are)	56	31.46%
Other ...	19	10.67%



E3. Where do you archive / preserve your data / code for the long-term (or plan to)?

Number of responses: 202

Answer	Times Chosen	Percentage
(Unsure where, but I plan to)	39	19.31%
Institutional / Faculty server or webpage	59	29.21%
Data repository (ex. DaSCH, EnviDat, ERIC, ETH Research Collection, MaterialsCloud, SWISSUbase, Yareta, Zenodo, ...)	51	25.25%
Data / Code journal (ex. Data in Brief, JOSS, ...)	5	2.48%
Publisher's platform (ex. supplementary material)	23	11.39%
Public database / databank (ex. Genbank, RECIFS, ...)	4	1.98%
Code sharing platform (ex. Bitbucket, GitHub, GitLab, ...)	113	55.94%
Data archive (ex. ACOUA, DaSCH, ETH Data Archive, OLOS, SWISSUbase, ...)	8	3.96%
(Never done and not going to)	12	5.94%
Other ...	8	3.96%

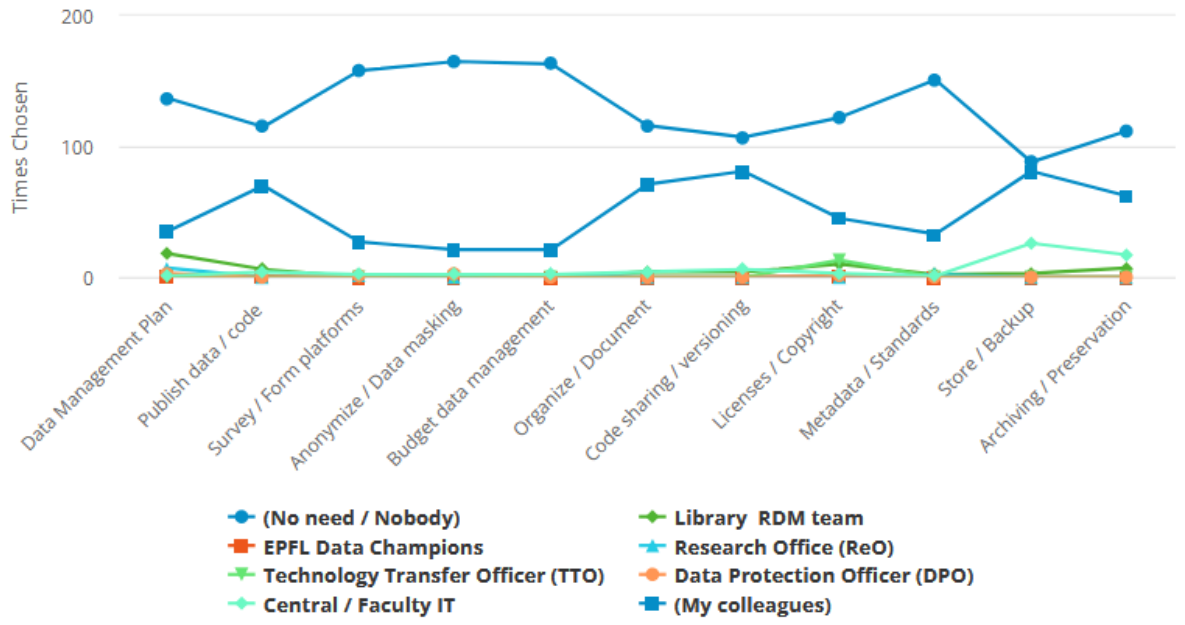


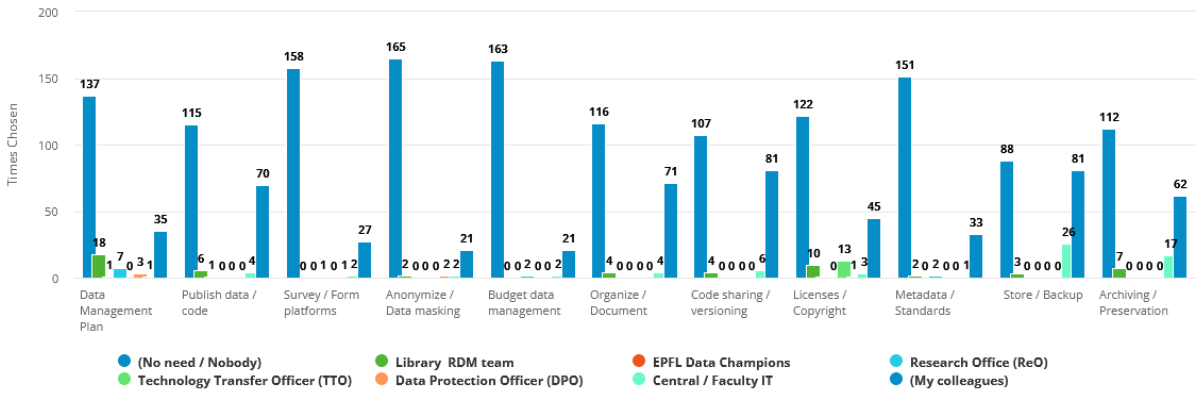
F1. Which EPFL service / support have you consulted for these various data / code management topics?

Number of responses: 187 [NB: this is by excluding the 5 more answers provided by whoever selected the answer "Another institution" for question "What is your main institutional affiliation?"]

	(No need / Nobody) 1		Library RDM team 2		EPFL Data Champions 3		Research Office (ReO) 4		Technology Transfer Officer (TTO) 5		Data Protection Officer (DPO) 6		Central / Faculty IT 7		(My colleagues) 8	
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%
Data Management Plan	137	67.82 %	18	8.91 %	1	0.5 %	7	3.47 %	0	0%	3	1.49 %	1	0.5 %	35	17.33 %

Publish data / code	115	58.67 %	6	3.06 %	1	0.51 %	0	0%	0	0%	0	0%	4	2.04 %	70	35.71 %
Survey / Form platforms	158	83.6 %	0	0%	0	0%	1	0.53 %	0	0%	1	0.53 %	2	1.06 %	27	14.29 %
Anonymize / Data masking	165	85.94%	2	1.04 %	0	0%	0	0%	0	0%	2	1.04 %	2	1.04 %	21	10.94 %
Budget data management	163	86.7 %	0	0%	0	0%	2	1.06 %	0	0%	0	0%	2	1.06 %	21	11.17 %
Organize / Document	116	59.49%	4	2.05 %	0	0%	0	0%	0	0%	0	0%	4	2.05 %	71	36.41 %
Code sharing / versioning	107	54.04%	4	2.02 %	0	0%	0	0%	0	0%	0	0%	6	3.03 %	81	40.91 %
Licenses / Copyright	122	62.56%	10	5.13 %	1	0.51 %	0	0%	13	6.67 %	1	0.51 %	3	1.54 %	45	23.08%
Metadata / Standards	151	79.89 %	2	1.06 %	0	0%	2	1.06 %	0	0%	0	0%	1	0.53 %	33	17.46 %
Store / Backup	88	44.44%	3	1.52 %	0	0%	0	0%	0	0%	0	0%	26	13.13 %	81	40.91 %
Archiving / Preservation	112	56.57 %	7	3.54 %	0	0%	0	0%	0	0%	0	0%	17	8.59 %	62	31.31 %





F2. What type of support are you most receptive to?

Number of responses: 192

Answer	Times Chosen	Percentage
On-site training	53	27.6%
Remote training	29	15.1%
e-Learning training platform	46	23.96%
Webpage / Online documentation	115	59.9%
Remote consulting by email	48	25%
Remote consulting by video conference	26	13.54%
In-person 1-to-1 consulting	58	30.21%
In-person group / lab consulting	65	33.85%
Infopoint / Physical help desk	16	8.33%
Other ...	0	0%