Summary of the study : Adapting Open Science

Célya Gruson-Daniel, Groupe Projet Réussir L'Appropriation De La Science Ouverte

Ouvrir la science !

CUBE inno
Innovation is Openness

Summary of the study : “Adapting Open Science”

French Committee for Open Science - Research Data College
“Successfully appropriating Open Science” Working Group

soutenu par
MINISTÈRE DE L’ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Table of contents

1 | Background and objectives of the study....3
The “Adapting Open Science” study....3
2 | Methodology....4
3 | Results....6
3.1 What factors should be considered to better understand the diversity of data practices in research?....6
Typology of practices and personas....6
Navigating practices: approach, tools and status....10
Adapting Open Science according to ….....11
3.2 How can we support the evolution of data-related practices in relation to the incentives and obligations of Open Science public policies?....12
4 | Key takeaways....13
4.1 Understanding research approaches in detail....13
Diversify the terminology used around ‘data’....13
Focus on "quality" in research rather than "reproducibility....13
Pay attention to the different forms of added value derived from research work....14
Research approaches: clinical/ experimental/computational specificity....16
4.2 Understand different practices for making data available....16
Think about the reuse of data and other resources and the audiences involved....16
Distinguish between different limits to availability and levers for improvement....17
Highlighting data conservation and security issues....17
4.3 Learn about learning methods and collaborative practices....18
Discovery and training in tools: an exchange between peers....18
Paying attention to interfaces....18
4.4 Diversify the types of support....19
Distinguish between different support needs....19
Develop a network of data support services as close as possible to the teams....19
Be vigilant about mediation issues within research teams....20
4.5 Considering status and career issues....20
5 | Limitations....21
6 | Conclusion....22
7 | Annexes....23
7.1 Main results of the "Data and Open Science" questionnaire....23
7.2 Access to the study’s various research outputs....27
7.3 Credits....28

1 | Background and objectives of the study

The Adapting Open Science study was carried out as part of the "Successfully appropriating Open Science" project led by the Committee for Open Science. It was carried out by a multi-disciplinary and professional working group of the "Research Data" college.[1] This project ran from May 2020 to December 2021 and was composed of three work streams:

- The design and organization of Open Science Legal Workshops (OSLA)[2]
- Participation in the Electronic Lab Notebook Working Group (ELN WG)[3]
- The Adapting Open Science study, which is the subject of the summary below.[4]

The "Adapting Open Science" study

The "Adapting Open Science" study began with a field survey of research professionals in various disciplines to:

- better understand **the practices associated with data and their evolution with Open Science**,
- understand **the factors that differentiate these practices** (discipline, research approach, etc.),
- provide **support adapted** to the needs of different research communities.

The study aimed to answer two questions:

1. What factors should be taken into consideration to better understand the diversity of practices associated with data in research?
2. How can we support the evolution of data practices in relation to the incentives/obligations brought about by Open Science policies?

1 The Committee's "colleges" are permanent bodies made up of experts on the various aspects of the National Open Science Policy. They examine subjects, provide opinions, propose guidelines and initiate and steer projects. The "Making Open Science Work" project was initiated as a result of the work of one college's working group on data use and governance.

2 The Open Science Legal Workshops (OSLA) were designed and organized online to facilitate dialogue and exchange between lawyers and research professionals. The objective was to collectively raise legal issues related to data and Open Science. Three workshops took place from November 2020 to May 2021 (1/images of research, 2/life cycle of data and 3/personal data) and gathered more than 150 people. A summary presentation and a workshop appropriation kit is available in French on the Open Science Committee website: https://www.ouvrirlascience.fr/atelier-juridique-science-ouverte-synthese-et-recommandations/

3 Participation in the Electronic Lab Notebook Working Group (ELN WG) in order to help to assess the existing electronic lab notebook offer and to define recommendations to facilitate their choice within institutions and their implementation. The report is available in French : Gilles Mathieu, Dominique Pigeon, Tovo Rabemanantsoa, Christophe Chipeaux, Simon Duvillard, et al.. Rapport du groupe de travail sur les cahiers de laboratoires électroniques. [Rapport de recherche] Comité pour la science ouverte. 2021, 68 p. DOI :**10.52949/3**

4 To learn more about the mission "Successfully appropriating Open Science", visit the Committee for Open Science's website and the list of deliverables: https://www.ouvrirlascience.fr/reussir-lappropriation-de-la-science-ouverte/

2 | Methodology

The study was based on **mixed methods** with **two initial qualitative phases** including **interviews** (exploratory, observation of practices), **focus groups, and a study day** dedicated to social science and humanities. **Following this**, a quantitative phase was carried out based on **the design, dissemination and analysis of a questionnaire** to research professors and staff in France (more than 400 responses). The study was finalized by c**ombining the results** of these phases[5] and **a design approach** to facilitate the appropriation of the content (cf. Figure 1).

The research work was part of **a collaborative approach** between the different members of the working group from various disciplines (biology, art history, history, health, Science & Technology Studies - STS) and professional research fields (archives, libraries, research, management and strategy, etc.). **An Open Science approach** (cf. Figure 2) was also tested in order to make the collected information available (in compliance with the GDPR) to facilitate the progress of the research (sharing of intermediate syntheses) and the reproducibility of the quantitative results (scripts, making data available).[6]

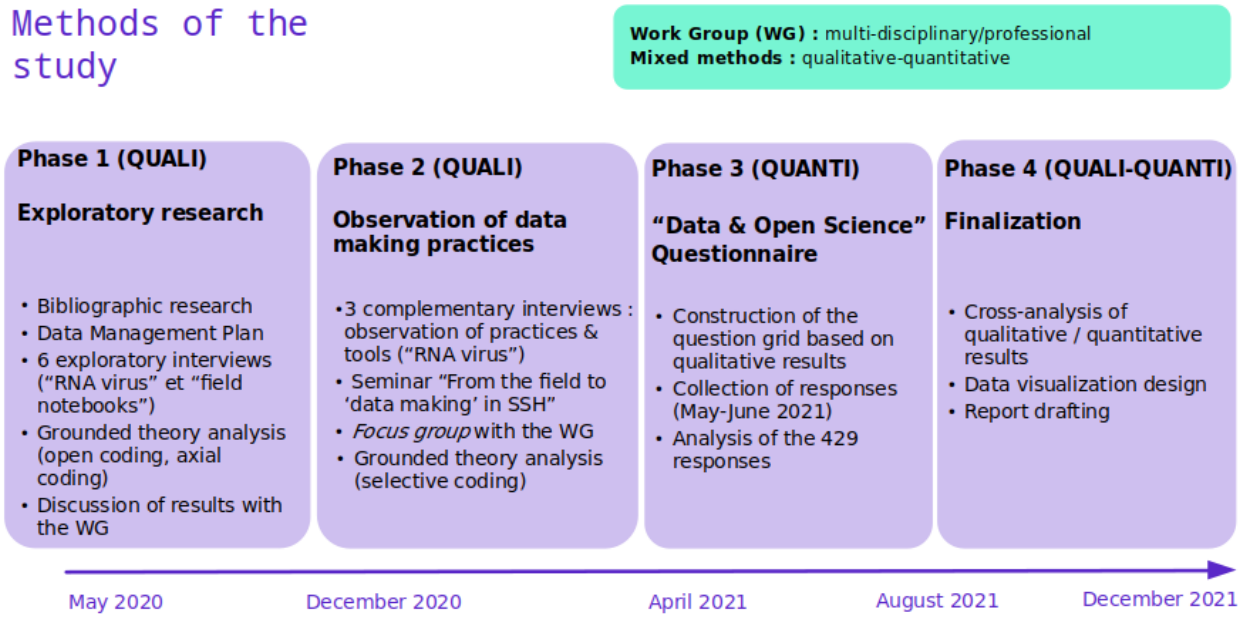


Figure 1: Summary of the different methodological steps of the study "Adapting Open Science"

5 The results of the qualitative phases were obtained through grounded theory analysis. Quantitative results were obtained through univariate (flat and cross tabulation) and multivariate (MCA and HCPC) statistical analysis. For more information, see the methodological guide referenced in section 7.2 (in French).

6

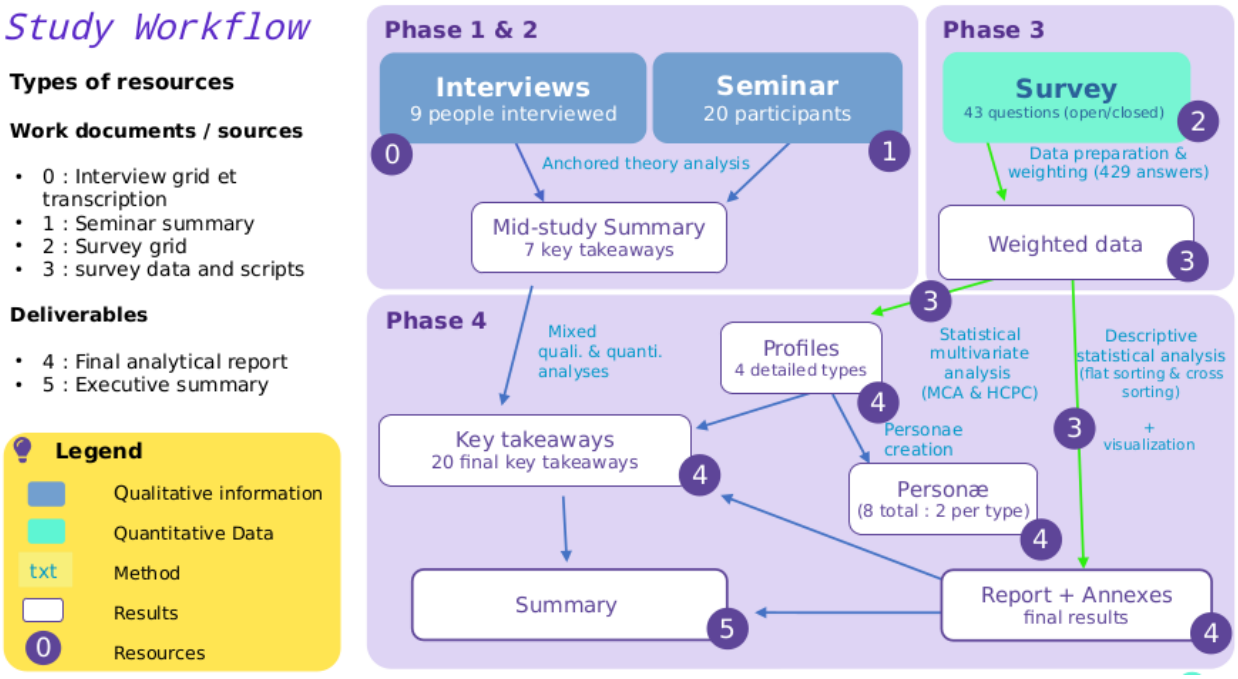


Figure 2: The Open Science workflow according to which information for the study was collected and processed.

3 | Results

3.1 What factors should be considered to better understand the diversity of data practices in research?

Typology of practices and personas

Although disciplines are an important factor in differentiating various data-related practices, the study shows that it is important **to go beyond the single disciplinary reading grid** and distinguish other differentiating factors.

The first exploratory interviews, supported by a review of the literature, led to the definition of **"data-related practices"** as all of the steps necessary to constitute data[7] and **to make it available** (ranging from restricted sharing to open data).

In addition to the disciplinary fields (Sciences Technology and Medicine - STM/ Social Sciences and Humanities - SSH), another factor taken into consideration was **the individual or collective nature of the research work**. Based on these factors, the multivariate analysis and multiple correspondence analysis of the survey responses, these elements made it possible to highlight **4 main types of practices (experimental, collaborative, computational, solitary)**.

- While the "discipline" axis strongly colors these typologies of practices - for example, **the "experimental" profile** is associated with **people from the Earth and Life Sciences** - other profiles such as **"computational"** bring together individuals from different disciplines (from computer science to linguistics) but who share **a common culture of data** and often **a knowledge of free and open source software**.
- Furthermore, within the Social Sciences and Humanities where the representation of a "solitary researcher" may still be dominant, **a "collaborative" profile stands out** with individuals implementing collective practices at different stages of their research, or at least wishing to train in them.
- Finally, **the "solitary" profile** (not restricted to the SSH), includes individuals who **conduct their research alone without necessarily wanting** to do so because of their status or working conditions, for example in the case of PhD students.

7 In STS, several studies looked at the constitution of data/databases and frictionless processes related to it. Data are considered as a construct that is subject to different stages, exchanges, use of tools, processes, until the production of what is called "data" with the aim in particular of being shared, exchanged and having value as evidence. Other concepts at the heart of this study are, for example, those of data journeys, datafication or the "public of data" (Jaton and Vinck 2016; Gruson-Daniel and De Quatrebarbes 2019; Gitelman 2013; Bowker and Star 2000; Heaton and Millerand 2013).

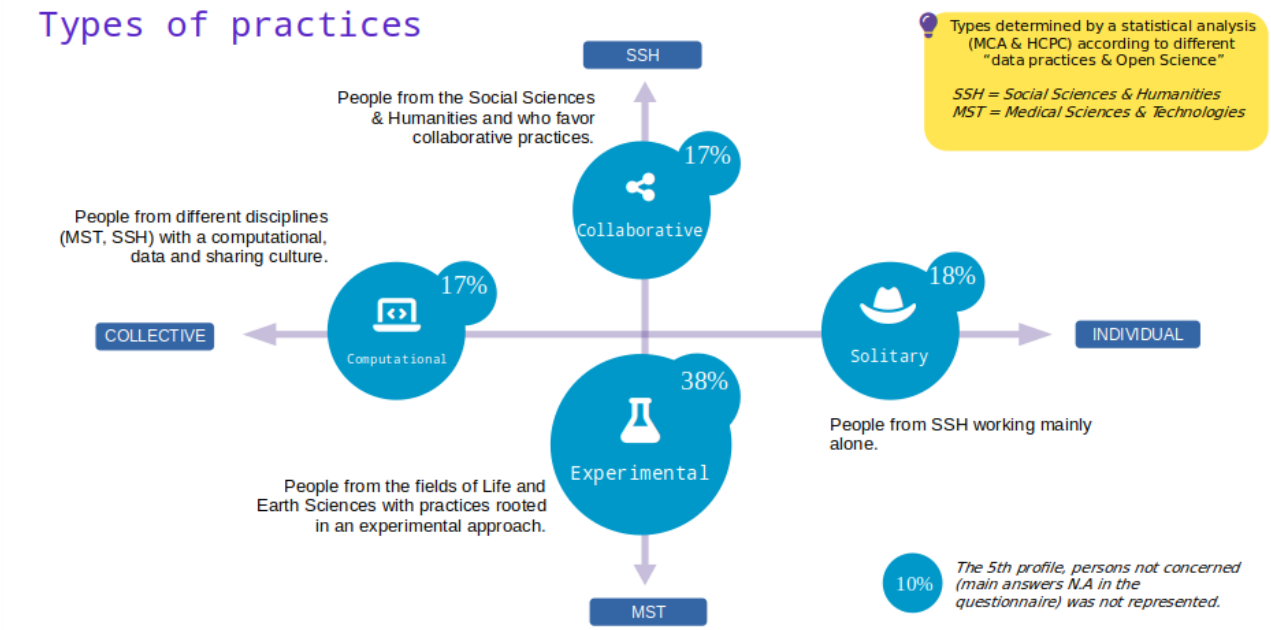


Figure 3: Presentation of the 4 personae according to types of practices associated with the data from the analysis of the questionnaire (MCA then HCPC).

Different personas were produced to give a better understanding of these profiles (types of practices) based on the results of a MCA (Multiple Correspondence Analysis), (see Figures 4 and 5). The personas are fictional characters, made up based on the answers to the questionnaire (the most representative of each class) and the results of the qualitative analysis phases. Each persona is presented in the form of a descriptive sheet and gives an overview of concrete situations encountered by a variety of research professionals (assistant professors, research engineers, researchers, etc.).

Figure 4: Example of two descriptive sheets (personas) illustrating "computational" and "experimental" profile types based on fictitious characters. Picture license: Unsplash.

Figure 5: Example of two descriptive sheets (personas) illustrating "collaborative" and "solitary" profile types based on fictitious characters. Picture license: Unsplash.

Navigating practices: approach, tools and status

In order to obtain **a finer level of granularity in the analysis** of these two general factors (discipline and collaborative/solitary nature), **three additional criteria were explored:**

1. research approach ;
2. learning tools and modalities; and
3. status and function in research.

For the "research approach" axis, we include various elements such as **the research environment** (laboratory, clinic, fieldwork, etc.), data origins (measurement instruments, archives, etc.), the relationship to data, particularly in **the terms used to name**the **criteria associated with research quality**, the **added value** of the research work, and the **steps involved in the making of data**[8]. These criteria, based on the analysis of the questionnaire responses, have made it possible to characterize the various research approaches that influence the relationship with the data. For example, in the context of laboratory work, the added value of research is linked to experimentation and one speaks more readily of measurements and values to qualify the data. In archival work or field studies, the terms corpus and materials are widely used for a main added value associated with the collection of rare data and theoretical work (see table below).

Research Approaches

Understanding research approaches in detail helps to address research communities in differentiated ways.

Work environment	Discipline	Criteria associated with research quality	Words used to describe data	Added-value of research work	Data origins
Laboratory	Medical Sciences and Technologies	Reproducibility, replicability	Measures, values, raw data Sources	Experimentation	Measuring instruments
Field work, Archives,	Social Sciences and Humanities	Explicability, transparency	Corpus, materials, sources Values and raw data	Collection of rare and standard data, theorization	Archives, website, grounds
Clinical	Life and Earth Sciences	Negative data	Pipelines, clean data, negative data Corpus	Modeling, workflow automation	Measuring instruments

Figure 6: Summary table of different research approaches according to work environment (laboratory, field/archive and clinical).

In the **"tools and learning methods"** category, the focus is more specifically on the **material context** (even if digital) of practices through the use of a set of tools. It is a

8 This axis is thus linked to epistemological and methodological principles and criteria of scientificity associated with different research paradigms and epistemic communities. The expression comes from S. Leonelli's book, Data Journey (2021).

matter of considering the ways in which tools are discovered and learned, **the appropriation of the digital work environment**, and the interest **in more or less collaborative work habits** and **the needs identified for support**. This axis offers a distinction between different communities of practice and learning.

Finally, a last category is that of **the status and functions allocated** to individuals in research, for example **the professional category** or **the status and seniority** within research (doctoral student, civil servant, etc.).

Adapting Open Science according to …

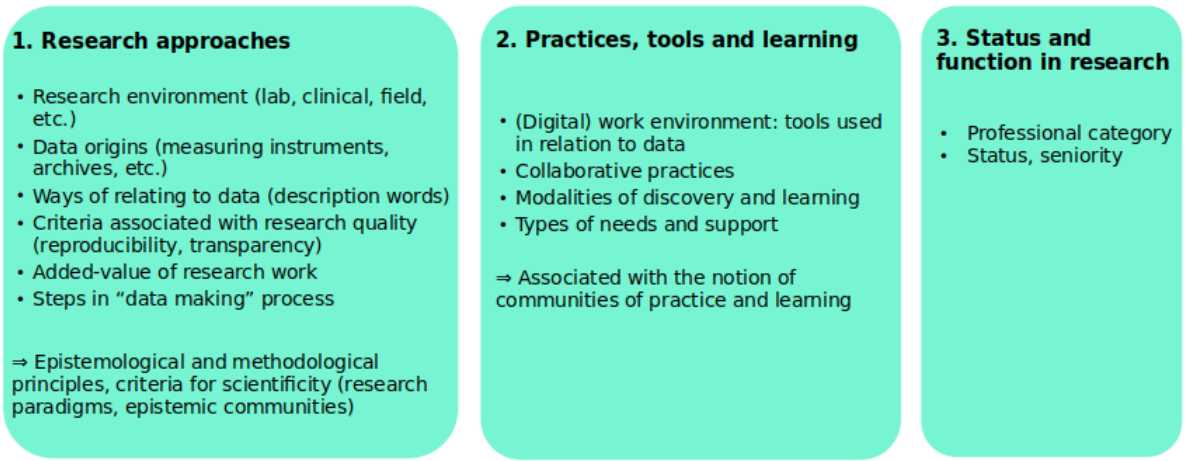


Figure 7: Summary of factors differentiating data-related practices (research approaches, practices/tools/learning, status and function in research) and their characteristics.

For each axis, the criteria aim to provide a detailed understanding of the various types of relationship with data and their representation by research professionals. These criteria can also influence their apprehensions and/or motivations to give open access or share data in an Open Science approach and consequently, must be taken into consideration when deciding which assistance and support solution to adopt.

3.2 How can we support the evolution of data-related practices in relation to the incentives and obligations of Open Science public policies?

In order to respond to this problem, five guidelines for accompanying measures have been determined on the basis of the lessons learned from the qualitative phases (interviews, observations of uses, seminar) and the results of the "Data and Open Science" questionnaires (see appendix).

- Orientation track 1: To understand in detail the research approaches;
- Orientation track 2: To apprehend different practices of provision of data;
- Orientation track 3: To know the modalities of learning and the collaborative practices;
- Orientation track 4: Diversify the types of accompaniment;
- Orientation track 5: Take into consideration the status and the career issues.

Regarding **incentives for Open Science related to research data**, we include, for example, the **application of FAIR principles** for data (Findable, Accessible, Interoperable, Reusable), the implementation of **data management plans (DMP)**, the encouragement of **greater reproducibility** of research work, the implementation of support and the deployment of infrastructures for making data available.

For each track, different themes have been distinguished, each associated with recommendations. The 20 recommendations aim to facilitate the evolution of practices associated with data and Open Science incentives while adapting to the various contexts of academic research.

4 | Key takeaways

The key takeaways are presented according to 5 larger orientations and associated sub-themes. They were formulated inductively from the results of study[9] and reflect the theoretical frameworks employed and the terminologies of the people interviewed.

Each takeaway is followed by a clarification. Access to all the study's outputs and methodological explanations are available in the general presentation (see 7.2 Access to the study's various outputs).

4.1 Understanding research approaches in detail

Diversify the terminology used around 'data'

Key takeaway 1: Do not remain with the notion of "data". Use a more precise and specific set of terms to designate the different objects manipulated and produced during the research.

Explanation : The term "data" is mostly used in the recommendations/incentives to Open Science to designate all information leading to the production of scientific results. Research data has been defined by the OECD as factual records (numbers, texts, images and sounds), which are used as primary sources for scientific research and are generally recognized by the scientific community as necessary to validate research results. Aside from this definition, other terms are more commonly used within the research communities to describe the elements that are used to obtain research results. Thus, it is advised not to remain with the notion of "data", but to use more precise and specific terms to designate the different objects manipulated and produced in the course of research. Depending on the research community one is addressing, several expressions can be used: "databases", "datasets", "corpus", "archives", "sources", "materials", "measurements".

Focus on "quality" in research rather than "reproducibility

Key takeaway 2: Broaden the issues of reproducibility to those of quality in research. Use other terms such as transparency, traceability, and explicability, especially with SSH communities.

Key takeaway 3: Integrate in the reflections on research quality and the availability of research data, the issues of scientific and ethical values (integrity, honesty) and the impact of research in its different dimensions (social, economic, technical, etc.).

9 The key takeaways are derived from the qualitative and quantitative results of the study, i.e. the life science and archaeology interviews, the seminar with research professionals in the humanities and social sciences, as well as the "data and open science" questionnaire. Although a majority of respondents were from the humanities and social sciences, the results and analysis were obtained after weighting different disciplines in order to have a balanced and distributed sample of HE&R disciplines (see the methodological guide on Gitlab: https://code.inno3.eu/ouvert/decliner-so).

Explanation : Today, the issues of reproducibility are an integral part of the discourse and incentives for Open Science[10]. However, it is necessary to detach ourselves from the term reproducibility in order to address more broadly the question of "quality" in research. Indeed, **the notion of "reproducibility" applies more specifically to research involving measurement instruments and the use of computational methods** (verification of calculations based on access to source codes and original data). Other terms are more inclusive to address the issue of research quality more broadly in different research communities. For example, **the principle of "transparency"** is to be used in a privileged way **in multidisciplinary research contexts[11].** The concept of **"explicability"** is used in the context of SSH work that requires the constitution of corpora or the construction of databases. On the other hand, the notion of **"replicability"** can be used preferentially in the framework of experimental research when it is a question of reproducing an experiment. This implies considering access to methodological protocols (not exclusively to data and source codes). Several comments also pointed out the importance of associating **the ethical principles and values (integrity, honesty, etc.) of research and its impacts** (social, economic, technical, etc.) with the reflections on the question of quality in research.

Pay attention to the different forms of added value derived from research work

Key takeaway 4 : Facilitating the availability of data implies taking into account, in a differentiated way, the investment of work required at different stages of the research, the added value created according to the research approach and the repercussions in terms of evaluation and career.

Explanation: When conducting research, different steps are necessary to obtain results that can be shared with the peer community. These steps generate **a more or less important added value according to the time devoted to their realization or to the degree of recognition attributed to this work** by the community. Different types of added value have been distinguished and then correlated to criteria related to the research process. For example :

- **the collection of rare data or data requiring a significant** amount of time is mainly associated with **fieldwork or with archives** and documentary collections in the Social Sciences and Humanities;
- t**he preparation of samples** and the definition of experimental protocols are activities associated with **laboratory research work**;

10 The second national plan for Open Science thus emphasizes the importance of a science that is reproducible, transparent, more efficient and cumulative. The issues of reproducibility of scientific results are addressed in particular in connection with the opening of source codes (third axis). Access to source codes and data are essential elements for reproducible approaches.

11 To reproduce the results of the multivariate statistical analysis of the questionnaire data that shows the correlation between the term "transparency" and the other variables, see the code repository on Gitlab: https://code.inno3.eu/ouvert/decliner-so/

- a **clinical research framework** is more strongly correlated with **added value derived from the automation of workflow processes** and modeling on a large quantity of data.

Paying attention to these different research approaches, as well as the forms of added value generated according to the contexts, is important in order to identify blockages in the provision of data. Some research approaches (technique improvement, automation, modeling) may encourage the provision of data, while other approaches may discourage it (rare data collection, time-consuming sample preparation).

Research approaches: clinical/ experimental/computational specificity

Key takeaway 5 : Consider the frictional elements of the research environment. This environment is made up of cultures that coexist while sometimes being in opposition. On the one hand, a technical-industrial culture aims at improving processes, risk management, and efficiency. On the other hand, there is a culture of independence and freedom of researchers associated with the claim of a posture of "craftsmanship" and creativity in the scientific approach.

Explanation: Even when research studies address the same objects of study, this does not mean that the research approaches are identical. For example, studies on RNA viruses (HIV, SARS, etc.) can be carried out in the context of clinical studies on the one hand, and in the context of basic bench or computational research on the other. In the case of clinical research, for example, a structuring of the data (with strict regulations on their use) is planned from the start (*Case Report Form*) and the added value of the work comes from an automation and an improvement of the protocols. Whereas in the framework of fundamental biology, the collection of rare data obtained following time-consuming experiments as well as their interpretation are at the heart of the scientific approach with the defense of a posture of craft and creation. These different approaches and added value sometimes meet within the same research projects with the criticism of an "engineering" of research on the one hand and a lack of quality control on the other hand.

4.2 Understand different practices for making data available

Think about the reuse of data and other resources and the audiences involved

Key takeaway 6 : Making data available requires thinking about the "data audience", the temporality, the sharing modalities (legal, financial, technical, etc.) and the necessary support.

Key takeaway 7: In addition to data, other objects (protocols, source codes, etc.) can also be made available under specific conditions to be defined.

Explanation: Concerning the reuse of data, at the heart of Open Science policies, it is necessary to take into consideration the "data audience" which influences the way in which data are made available but also the modalities of sharing and contextualizing this information. Sharing data among colleagues (peer community) to ensure reproducibility will not imply the same work in terms of explicitness and contextualization of the data as making it available to a wider audience with the objective of a broader dissemination of research results. This also implies thinking about the support needed for sharing (help with data structuring, outreach videos that broadcast research work to the greater public) as well as the attached ethical and legal issues. Moreover, data sharing is not the only element to include in an Open Science approach. For example, in the case of

experiment replicability, sharing the protocol is essential, as is making the source codes available for reproducing the analysis of specific data.

The term "making available" is used in the study to distinguish different practices including:

- sharing restricted to a targeted and known public (via email for example);
- putting the data online on a site/warehouse with access control or not;
- the opening of data on a repository with an open license (*open data*).

Distinguish between different limits to availability and levers for improvement

Key takeaway 8: Differentiate the reasons limiting the availability of data (too much time required, lack of habit, competitive advantage not to share) to provide appropriate responses.

Key takeaway 9: Encourage journal editorial boards to build on existing national policies regarding regarding data and source codes associated with scientific publications.

Explanation: The main reasons limiting the availability of data are mainly lack of familiarity with these practices, too much time needed to make them available, and a desire to add value to the data storage (and retain information) to maintain a competitive advantage. Secondary reasons include questions about the risks of additional bureaucracy generated by making data available, as well as legal and ethical issues surrounding access to personal data. There is little awareness of the obligations to make data available, and these obligations are mostly from journal editorial boards or ethics committees. Making committees aware of the issues involved in making data available is a key element for taking these practices into account in the evaluation and recognition of research work, as their role in this process is important[12].

Highlighting data conservation and security issues

Key takeaway 10: Raise awareness of the distinction between data storage and archiving, which involve different services and different infrastructures as the need for a possible selection of data in order to differentiate data to be kept from data to be destroyed

Key takeaway 11: Prioritize and/or highlight the security features and reliability elements offered by the research infrastructures made available for data storage.

Explanation: As far as data storage is concerned[13], it is mostly done on external media and professional computers. Nevertheless, in the Social Sciences and Humanities, the

12 Moreover, the Law for a Digital Republic (2016) can be recalled in these committees with article 30 which prevents publishers from limiting the dissemination and reuse of data associated with publications when they are the result of research financed for at least half by public funds.

13 By data storage, we include data processed by researchers during the collection and analysis phase as well as data produced for release (sharing/opening). The notion of “data life cycle” was not addressed by the respondents and there was little or no knowledge of data archiving processes (choice of data to be destroyed or archived).

use of personal computers is frequent, especially for doctoral students, which does not facilitate the follow-up of data, their security or their reuse at the end of a project. The communities are particularly vigilant about data security (encrypted data, risk of hacking, etc.) and question the reliability of institutional infrastructures. Cloud solutions such as Google Drive or Dropbox are mostly used for file sharing. Moreover, at present, the difference between storage and archiving remains blurred for the research communities. Archiving services are rarely used, because storing data seems to be a sufficient action for research professionals to preserve their data.

4.3 Learn about learning methods and collaborative practices

Discovery and training in tools: an exchange between peers

Key takeaway 12: To facilitate the appropriation of new practices, take into consideration the specificities of community meetings and learning (laboratory life, study days and conferences, social networks, etc.).

Explanation: In addition to discovering tools on one's own, the role of other people within research teams (team members or other teams) is essential to build up one's digital work environment. Habits are often formed as soon as the first research internships in a master's degree with training within the teams (internship supervisor, "laboratory" life for work on the "bench", etc.). In the Social Sciences and Humanities, seminars and informal times play an important role in discovering new tools and sharing practices. Social networks also represent spaces for exchanging and discovering practices, which are considered useful especially when different communities meet.

Seminar: « from the field to the 'making of data' in SSH »

As part of the survey (phase 2), a study day was dedicated to the study of "data making" practices in SSH and allowed three key issues to emerge:

- Common issues in "data making" practices[14],
- Reconfiguration of research groups,
- Environment and recognition of "data making" work.

Paying attention to interfaces

Key takeaway 13: Pay particular attention to data processing and analysis interfaces so that they do not become "black boxes" and "dead ends" (lack of interoperability, proprietary formats, etc.).

Key takeaway 14: Be vigilant about the new turnkey solutions that are being developed for data analysis and manipulation.

14 A French-language summary of the workshop's findings can be found here : https://pad.inno3.eu/4-3ODX5JSBCRcXEWFwK2FA#

Key takeaway 15: Provide training in computational practices, but without wanting to turn all research actors into data scientists. Provide sufficient background to be able to exchange and understand the issues.

Explanation: Graphical interfaces are essential in the data processing and analysis stages. They help to easily manipulate data, filter information, classify it, visualize it quickly and dynamically, and facilitate continuity between different actors with differentiated access layers. For teamwork, extractions of graphs or elements from a database facilitate exchanges and collective analysis. However, turnkey tools, also called "*click and play*", make the underlying algorithmic processing invisible. With the rise of data science platforms (AI-oriented statistical data science tools), a literacy in computational principles is necessary from the first cycles of higher education in order to cultivate a critical eye regarding these turnkey platforms.

4.4 Diversify the types of support

Distinguish between different support needs

Key takeaway 16: In addition to support for storing, archiving and making data available, offer support for describing and mediating data for various audiences in different formats as well as for addressing legal and ethical issues.

Explanation: The requests for assistance formulated by the research communities consist first of all in requests for human and financial resources: creating or renewing permanent positions , financial assistance for access to databases, or for digitization. As far as data is concerned, the assistance requested concerns storage during data processing, archiving and availability. For the Social Sciences and Humanities, support for dissemination to the general public in the form of videos or blog posts is an important issue that is often not covered in research project budgets.

Develop a network of data support services as close as possible to the teams

Key takeaway 17 : Facilitate a "a network of data support services" at different scales by diversifying the support and accompaniment methods through 1) the development and maintenance of infrastructures, and 2) acculturation within research teams through support persons already present in the daily life of the teams to play a mediation role, understand the needs and the culture of the laboratory or team.

Key takeaway 18 : Be careful about adding additional "data referent" functions to the workloads of people already in place, to the detriment of creating stable and permanent positions dedicated to data availability missions.

Explanation: In addition to the implementation of one-stop portal and national infrastructures to support data-related practices, the people interviewed for the study were in favor of a network as close to the teams as possible. Stable and permanent relays within the teams are requested, although there is some mistrust as to the

additional workload that would be generated by adding a new "data referent" function to the people already on the job, particularly research or study engineers.

Be vigilant about mediation issues within research teams

Key takeaway 19 : Pay attention to the necessary translation and mediation issues that arise when managing and making data available within research communities. This involves finding "common denominators" among tools and documentation that are being used, as well as data and protocol standardization processes.

Explanation: For many, adapting to new data processing, analysis, and sharing practices is accompanied by new and/or complementary work processes and environments to be appropriated. This also reconfigures the working methods between different team members (IT departments, engineers, researchers, etc.) with a set of possible frictions. The constitution of databases between different disciplinary or professional profiles as well as their availability in data warehouses (sharing or opening) crystallize tensions (constitution of vocabularies, reduction of the complexity of a study, recognition of the people who participated in the creation of the database, etc). Nevertheless, these new objects are also a way to build new practices adapted to the skills of each person. Building the necessary dialogue and understanding between different people and skill-sets (translation of specific vocabulary, encouraging exchanges through mediation processes, etc.) requires time and sometimes financial, material or organizational support.

4.5 Considering status and career issues

Key takeaway 20 : Give greater consideration in the career development and evaluation of research professionals to the work of "data development" and making data available.

Explanation: The work of "constituting data" and making data available often requires time, for example, collecting sparse data, formatting/cleaning data, adding documentation, adding metadata, posting to repositories. It is important to recognize the time spent on these activities in the evolution of careers, especially in the case of people with a status and function that can lead to solitary work, a context in which these tasks are even more invisible. Indeed, if some researchers prefer to work alone and not to change their practices by choice or by political positioning , others have a solitary and "non-sharing" approach imposed. This is the case, for example, for doctoral students who are interested in Open Science topics, but for whom data sharing activities are not a priority, nor for their supervisors. For post-doctoral researchers, in the same way, the search for a position often takes precedence over developing these practices, even if this may lead some to develop a visibility and networking strategy around these practices.

5 | Limitations

A first analysis of the results of the questionnaire showed an over-representation of research communities in the Social Sciences and Humanities (SSH). Following this, the results were weighted according to the current distribution of researchers in different disciplinary categories[15]. The way the questionnaire was distributed certainly explains this over-representation. The questionnaire was shared on discussion lists and social networks followed by members of the Adapting Open Science working group. Several lists were associated with the Social Sciences and Humanities (history, sociology, economics, etc.) and the announcement circulated more widely in these communities. In view of the results, the questionnaire would benefit from being shared more widely within institutions in order to refine the results concerning disciplines that are currently under-represented and to confirm the relevance of the factors differentiating the practices highlighted.

15 We have taken as a reference the 2019-2020 data from the State of Higher Education, Research and Innovation in France - n°14 - April 2021 https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/T579/les_personnels_enseignants_de_l_enseignement_superieur_public_sous_tutelle_du_mesri/

6 | Conclusion

This study aimed to study the current practices associated with data in various research communities and to best accompany their evolution in a digital context and public policy that are favorable to Open Science. The objective was to present with a fine granularity elements explaining the diversity of research practices within what is called "Science" in order to better decline and adapt Open Science measures according to epistemic communities or practices. More than a simple disciplinary view, the typology of practices highlighted - and their illustration by personae (typical profile) - shows the importance of considering the solitary or collaborative nature of the work that is part of diverse social, methodological and technical fabrics.

A better appropriation of new practices associated with data by the communities requires an in-depth understanding of different research approaches, as well as a look at the tools and devices used and their learning and discovery modalities. Through the differentiating factors defined, the orientations and recommendations proposed, this study wishes to help those involved in Open Science policies and projects to better dialogue with the research professionals they are called upon to accompany, as well as to diversify the types of assistance offered.

For the people concerned by these practices and subject to their evolution, the study wishes to participate in a step back and reflexivity. It is a question of better understanding "our practices" and/or having a framework of explanation on the practices of other colleagues. Far from wanting to decide or judge the quality of the norms to be applied within research teams or collectives, this study is rather about giving leads to adapt the modalities of interaction between research professionals, to understand the reasons for frictions or blockages to Open Science measures and their incentives, as well as to make available elements of argumentation and debate so that these changes in practices are an enlightened and desired act.

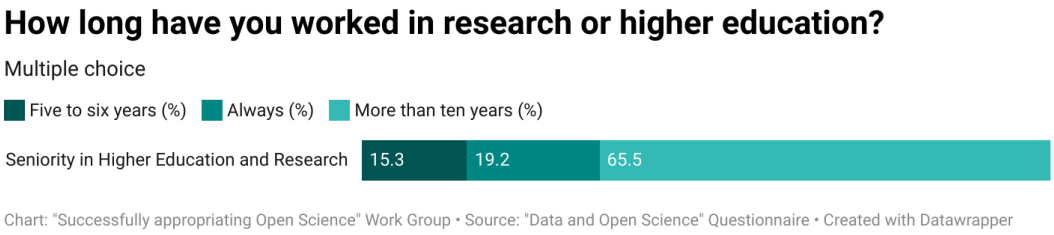
7 | Annexes

7.1 Main results of the "Data and Open Science" questionnaire

429 responses were obtained to the "Data and Open Science" questionnaire, which provided an overview of the current practices of research professionals[16].

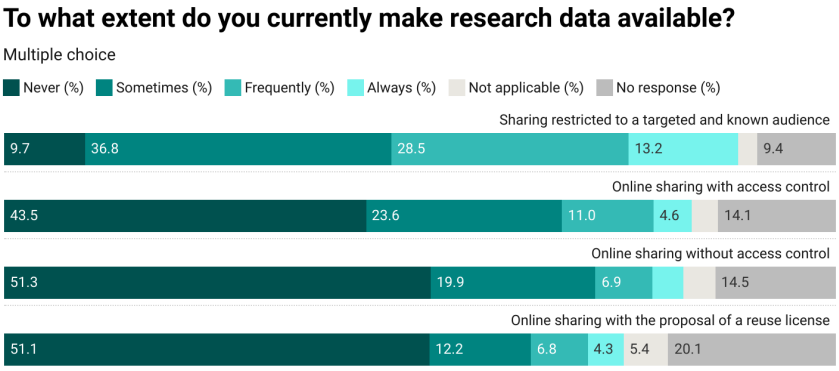
Population

- Gender: 47.8% and 44.8% (other: 7.4%).
- Main function: mainly "tenure-track" (55%) and "non-tenured" (20%) teacher-researchers.
- Seniority: Mostly more than 10 years (65.5%) (see graph below) with civil servant status (68.7%).
- Results weighted according to disciplinary categories to be representative of the 2019-2020 data (State of Higher Education, Research and Innovation in France n°14)[17].



Data "sharing" practices and obligations

- Restricted sharing for a targeted and known public remains the majority practice (78.5%) (see graph below).
- Online sharing with the proposal of an open license (open data) represents a little more than 20% of practices (see graph below).



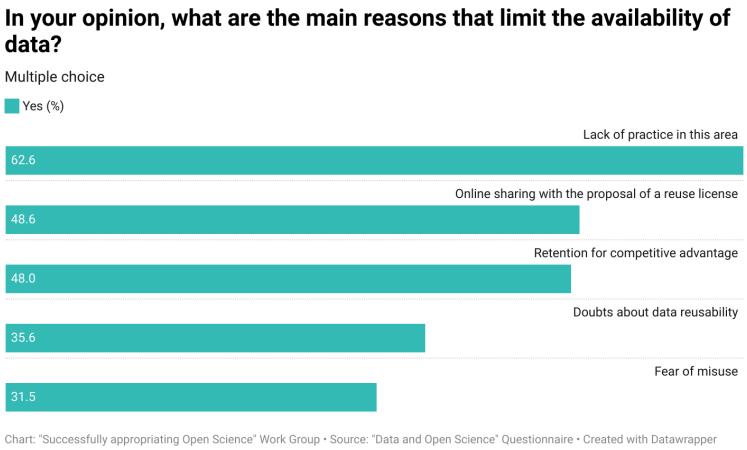
16 Results presented are those after weighting of disciplines in order to reach a representative sample of the disciplinary distribution in Higher Education and research based on the 2019-2020 Data - The State of Higher Education, Research and Innovation in France (n°14 - April 2021).

17 SIES, Sous-direction des systèmes d'information et des études statistiques, eds. 08. The State of Higher Education, Research and Innovation in France 2021. Paris: SIES.

- The obligations to make information available are not well known and concern mainly obligations from an editorial or ethical committee (for example, in biomedical research).

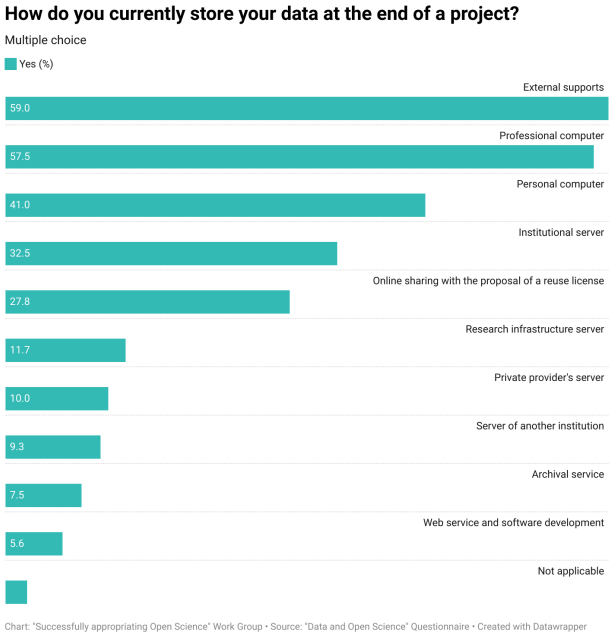
Data Reuse and Limitations on Availability

- Nearly 50% of respondents report that they often and/or sometimes reuse data that has already been produced or published.
- More than 45% of respondents consider that their data would be potentially reusable.
- The main reasons limiting data availability are primarily (see chart below):
  - lack of familiarity with these practices (63%);
  - too much time required (49%);
  - a desire to retain data to maintain a competitive advantage (48%).



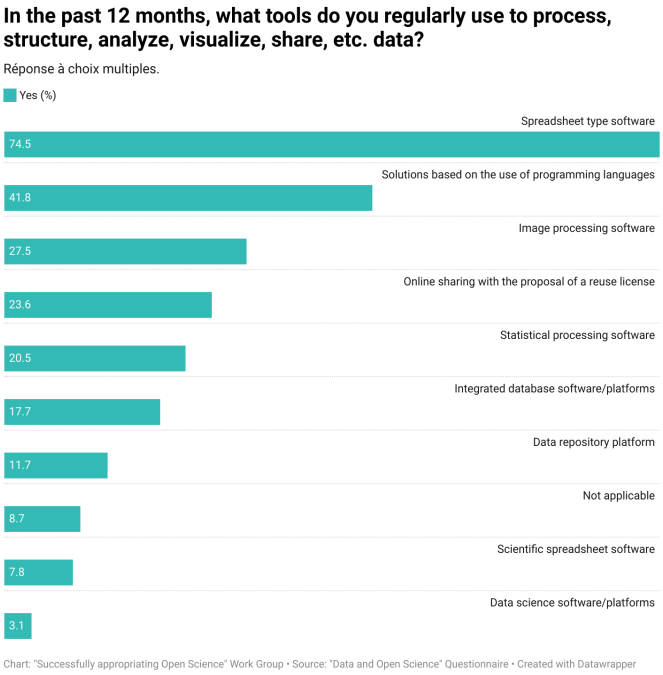
Data storage

- The majority of data storage is done on external media (59%) and professional computers (57.5%).
- There is little use of archive services (7.5%) (see graph below).



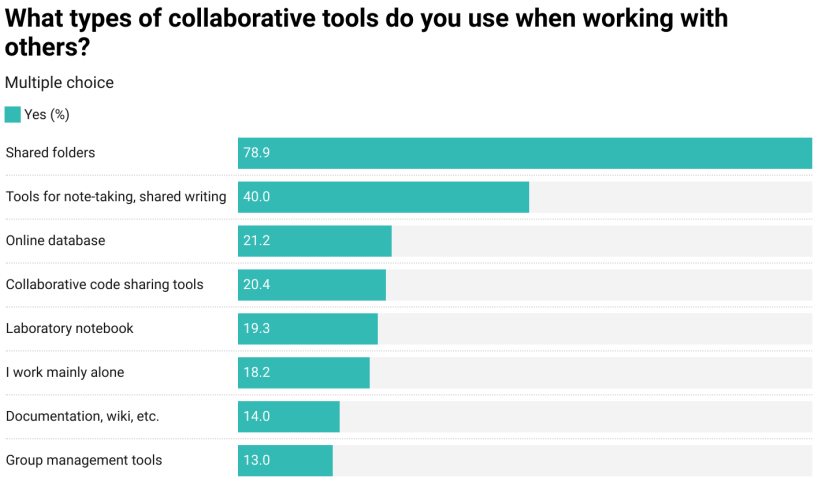
Tools used associated with the data

- Majority use of spreadsheet software (Excel, Calc) (74.5%).
- More than 40% use solutions based on the use of programming languages (R, Python).
- QGIS is one of the most frequently cited data analysis and visualization software (24%).
- Integrated database software/platforms (18%) frequently cited are FileMaker, PostgreSQL, MySQL.
- Data warehouse platforms were used by only 12% of respondents.
- The most widely used operating system is Windows (62%) versus 26% for MacOS and 12% for Linux and other Unix.



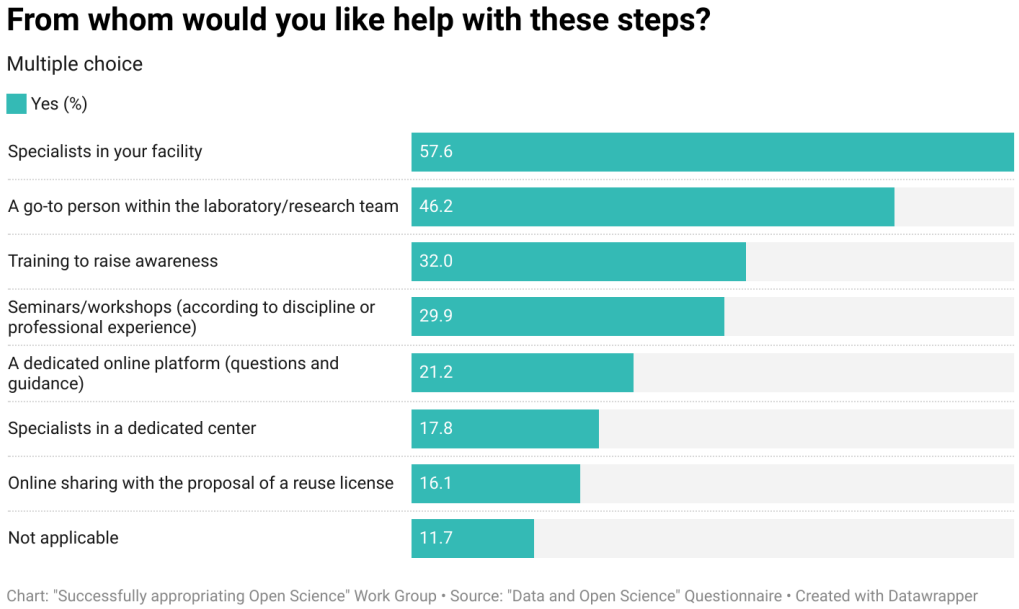
Collaborative practices

- Shared note-taking tools are used by 40% of respondents.
- The use of non-institutional tools is common (DropBox, GDrive, etc.).



Needs and support

- The majority of respondents are aware of available support options (over 55%).
- Those who have received support are generally satisfied (over 60%).
- The necessary support associated with the data concerns first and foremost the storage and conservation of the data (48%), followed by making it available (40%).
- The presence of specialists within the institution (58%) and referents within laboratories/research teams (46%) are the most popular support methods (see graph below).



7.2 Access to the study's various research outputs

- Final report and annexes (interview question grid, survey question grid, data management plan – DMP) in French: on HAL (DOI: 10.52949/27)
- Summary in French: on HAL (DOI: 10.52949/28)
- Summary in English: on HAL (DOI : 10.52949/29)
- "Raw" data collected from the survey: on Recherche.data.gouv.fr (DOI 10.57745/V64RYT)
- Reproducible source code for analyzing the survey data and methodological guide: on Gitlab
- Editorialized website presenting all project-related content: on PubPub

7.3 Credits

The study was carried out by an interdisciplinary working group in which different higher education and research professions were represented. This allowed a variety of expertise and skills (quantitative and qualitative analysis, various types of feedback, etc.) to take part in the project:

- Anne Vanet: Vice-President Digital and Open Science (University of Paris), Director of the Institut Jacques Monod genoinformatics cluster
- Hélène Chambefort: Responsible for the Archives (INSERM)
- Marie Herbert: Head of the Collex/Persée project (Lyon 1 University)
- Juliette Hueber: Editorial and document engineering manager (InVisu CNRS/INHA)
- Claire Lemercier: Director of research CNRS at the CSO SciencePo

The team was accompanied by the consulting firm inno³:

- Célya Gruson-Daniel: Associate researcher at COSTECH (UTC), consultant in charge of project management, design of the study, collection and analysis of qualitative information, redaction of the report and summary
- Benjamin Jean: Lawyer, president of inno³
- Romain Rouyer : Designer in charge of data visualizations and graphic illustrations
- Tamara Glushetckaia : content editing, layout and charts
- Maya Anderson-González : content editing and proofreading in English
- Emilien Schultz (SciencePo) for the reproducibility work done on the survey analysis