Rapport d'analyse – Enquête : Les données de la recherche à l'université ParisSaclay, panorama et perspectives

Mireille Brenel, Cédric Mercier, Stela Suhan, Adib Kassas, Claire Ménard, Alicia Ribeiro, Nadège Arnaud, Maximilien Petit, Gaëlle Jaouen, Eva Legras, et al.

Université Paris-Saclay

Rapport d’analyse – Enquête Les données de la recherche à l'université Paris-Saclay, panorama et perspectives

Mireille BRENEL, Cédric MERCIER, Stela SUHAN (Université Paris-Saclay), Adib KASSAS, Claire MENARD, Alicia RIBEIRO (Université d’Evry), Nadège ARNAUD, Claire LEBRETON, Maximilien PETIT (UVSQ), Gaëlle JAOUEN, Eva LEGRAS (AgroParisTech), Annie LE BLANC (CEA Paris-Saclay), Alexia BAUVILLE (Elève conservatrice, Enssib)

université PARIS-SACLAY

Table des matières

Avant-propos 4

1. Méthodologie 5

1.1 L'enquête quantitative 5

1.2. Les entretiens qualitatifs 6

1.3. Biais et limites de la démarche 8

1.3.1. Limites sémantiques 8

1.3.2. Limites méthodologiques 8

1.4. Profils des répondant·es 9

2. Définition et typologie des données de la recherche 14

2.1. Une définition des données de la recherche parmi les répondant·es 14

2.2. Une connaissance du contexte des données de la recherche à géométrie variable 15

2.3. Typologie des données 18

3. Pratiques des données de la recherche 20

3.1 Gestion au sein des laboratoires : les données intégrées au fonctionnement des unités de recherche 20

3.2. Le plan de gestion de données : entre *a priori* négatifs et obligation des tutelles 22

3.3. Le stockage et l'archivage au cœur des pratiques des données de la recherche... 24

3.3.1 Une vision expérimentale de la pérennité des données 28

3.3.2 Archivage et classification des données de la recherche 30

3.4. Partage et publication des données 33

3.4.1. Partage des données : une approche collective de la recherche ancrée dans les pratiques 33

3.4.2. Publication des données : le paysage de l'Université Paris-Saclay 36

3.4.3. Réutilisation des données 42

4. Motivations et freins à la publication des données 43

4.1. Accessibilité des données : des motivations centrées autour des pratiques de la recherche et de la Science Ouverte 44

4.2. Des freins importants : la contradiction inhérente à la publication des données . 46

5. Besoins exprimés par les répondant.es 50

5.1. Autour du Plan de Gestion des Données 52

5.2. Autour des outils de diffusion et du stockage 53
5.3. Un accompagnement aux multiples facettes 55
5.4. Parole libre 56
Conclusion 58

Avant-propos

Cette enquête, proposée par l'ensemble du réseau des bibliothèques et centres de documentation et le Comité de pilotage de la Science Ouverte de l'Université Paris-Saclay en 2021, s'inscrit dans un contexte dynamique de Science Ouverte prônant l'obligation de publication des données de la recherche selon les principes FAIR (Facilement trouvable, Accessible, Interopérable, Réutilisable).

La collaboration entre les établissements de l'Université Paris-Saclay est très forte dans le domaine de la science ouverte, y compris autour du thème des données de la recherche. Cette enquête a d'ailleurs été l'occasion d'impulser une volonté d'actions communes au réseau sur ce sujet. Il est donc apparu pertinent de dresser un paysage le plus complet et élargi possible autour des pratiques de la recherche. Elle s'adressait à tous les chercheur·es, ingénieur·es de recherche et d'études, doctorant·es ou encore personnels administratifs de l'ensemble de l'université, en somme toutes personnes amenées à créer, manipuler, utiliser des données. Tous les établissements de l'Université étaient également concernés. L'ensemble de la communauté scientifique du périmètre recherche de l'Université Paris-Saclay était concernée, et par conséquent les établissements suivants :

- Les établissements composantes : Université Paris-Saclay, AgroParisTech, CentraleSupelec, ENS Paris-Saclay, Institut d'Optique-Graduate School
- Les universités membres associés : Université d'Évry, Université de Versailles Saint-Quentin-en-Yvelines
- Les ONR (Organismes Nationaux de Recherche) : CEA Paris-Saclay (Commissariat à l'Energie Atomique), Inria Paris-Saclay (Institut national de recherche en sciences et technologies du numérique), ONERA Paris-Saclay (Office National d'Etudes et de Recherches Aérospatiales), INRAE (Institut national de recherche pour l'agriculture, l'alimentation et l'environnement), IHES (Institut des Hautes Etudes Scientifiques), INSERM (Institut national de la santé et de la recherche médicale), CNRS (Centre national de la recherche scientifique)

L'enquête répondait à un triple objectif :

- Établir un panorama des données produites au sein de l'Université
- Dresser le paysage des pratiques autour des données de la recherche
- Connaître les besoins des chercheur·es autour des données

Lors de sa mise en ligne entre le 29 avril et le 13 juin 2021, nous avons obtenu 513 réponses complètes.

1. Méthodologie

L'équipe en charge de sa conception a choisi de s'appuyer sur deux méthodes :

- Un questionnaire élaboré autour des données de la recherche, balayant trois grands thèmes : la connaissance du contexte général des données, la typologie des données produites par les répondant·es, leurs pratiques (en terme de stockage/archivage, publication, leurs attentes et besoins dans la gestion de leurs données de recherche)
- Des entretiens qualitatifs menés sur un échantillon de chercheur·es, ingénieur·es et personnels IST (Information Scientifique et Technique) des laboratoires de l'université, sélectionnés sur la base du volontariat parmi les répondant·es à l'enquête.

Le fait de combiner ressources quantitatives et approche qualitative permet d'aboutir à un résultat liant analyse chiffrée et perceptions des acteurs de la recherche. L'analyse des résultats du questionnaire en ligne a porté sur plus de 500 réponses (sur un public cible d'environ 15000 personnes) ; les entretiens ont été menés auprès de 24 volontaires. L'ensemble des constats portés sur les résultats de l'enquête a permis de mieux appréhender les besoins des acteurs de la recherche en termes d'accompagnement sur les données de la recherche.

La question des données de la recherche au sein des universités n'est pas un sujet nouveau, mais il reste difficile à appréhender du fait d'une actualité juridique, scientifique et institutionnelle mouvante. Les pratiques sur la question sont diverses au sein des disciplines, et il est intéressant pour un établissement pluridisciplinaire de cerner l'existant afin de proposer une offre de services adaptée.

1.1 L'enquête quantitative

L'approche quantitative est essentielle et sert à définir les grandes tendances et obtenir des données chiffrées. Le questionnaire en ligne soumis en ligne comporte au total 186 questions, réparties en trois grands thèmes : définition et typologie des données, pratiques des données de la recherche, besoins et attentes autour des données.

Un arbre directif a permis de cibler plus particulièrement des catégories de répondant·es : directeurs et directrices de thèse, directeurs et directrices de laboratoires, porteurs et porteuses de projets, amenés par leurs responsabilités à prendre position sur les enjeux des données de la recherche.

L'objectif fixé de plus de 500 réponses ayant été atteint, la représentativité de l'enquête est considérée comme satisfaisante (à noter toutefois une représentativité variable des établissements composantes, universités associées et organismes de recherche

partenaires). Elle est cependant moins représentative des acteurs de la recherche en sciences humaines et sociales (58 répondant·es seulement), ce qui doit constituer un point d'attention constant lors de l'analyse faisant intervenir une notion de disciplinarité.

Des espaces de réponses libres ont été inclus, ce qui a permis à plusieurs répondant·es de l'enquête de préciser leurs réponses tout au long des questions. En plus de ce dispositif, un espace de parole libre a été ouvert en fin d'enquête.

1.2. Les entretiens qualitatifs

L'organisation d'une campagne d'entretiens qualitatifs en plus de l'enquête quantitative apporte une vraie valeur ajoutée. Les enquêtes sur la gestion des données de la recherche mettent très souvent en avant le fait qu'il y a d'abord un problème de vocabulaire, de terminologie, et que les définitions institutionnelles des données de la recherche ne coïncident pas nécessairement avec les pratiques que l'on observe en détail chez les enseignants-chercheur·es. Rencontrer des chercheur·es au cours d'entretiens, et aborder la question des matériaux de recherche, des données, c'est mettre en lumière des terminologies, des pratiques, des discours et des attentes qui sont riches d'informations. Les entretiens offrent plusieurs avantages :

- Faire remonter efficacement les problématiques des communautés scientifiques qui relèvent des sciences humaines et sociales (SHS), moins représentées dans l'enquête quantitative,
- Créer un premier contact avec des communautés scientifiques qui n'ont pas encore été accompagnées,
- Organiser une collecte de données issues d'entretiens qualitatifs, c'est questionner l'organisation de travail des répondant·es sur le cycle de vie des données de la recherche. C'est engager une réflexion sur la planification, la collecte, l'analyse, la curation, le dépôt, la publication et la réutilisation de données particulières,
- Croiser les données d'une enquête quantitative avec celles qui proviennent d'une campagne d'entretiens qualitatifs est une méthode qui a déjà été utilisée pour d'autres enquêtes sur le sujet au sein des universités (Lille 3 en 2015, Rennes 2 en 2017, Bordeaux Montaigne en 2018, Paul-Valéry Montpellier 3 en 2019, Aix-Marseille en 2019).

Les choix de notre campagne d'entretiens

Différents choix ont été effectués durant la campagne, explicités ici afin de faciliter la réutilisation des données de l'enquête : stratégie d'échantillonnage, méthodologie d'entretien, stockage et traitement des données.

L'échantillonnage s'est fait sur la base d'un appel aux volontaires lors de l'enquête en ligne. Sur les 513 réponses, quarante et une personnes étaient volontaires pour participer à un échange.

Un objectif initial de trente entretiens, ne devant pas dépasser une durée de quarante-cinq minutes, a été posé afin de ménager une charge de travail raisonnable pour les intervieweurs. En tout, finalement, il y a eu vingt-quatre entretiens, correspondant à seize heures et six minutes d'enregistrement audio à traiter. Sur ces vingt-quatre entretiens, cinq personnes travaillent dans le champ des sciences du vivant et de l'environnement, onze personnes en sciences et technologie, huit personnes en sciences humaines et sociales. Ces vingt-quatre personnes proviennent de sept établissements différents de l'Université Paris-Saclay et exercent dans une vingtaine d'unités de recherche. La diversité des profils est donc bien respectée.

Il a été décidé également d'interroger quelques personnes en dehors de l'échantillon de volontaires, afin de recueillir la parole de profils que l'on sollicite rarement sur le sujet. C'est notamment le cas des ingénieur·es d'études et des ingénieur·es de recherche dans les laboratoires. D'un point de vue méthodologique, il semblait plus judicieux d'effectuer des entretiens de type semi-structuré, le but étant de permettre aux personnes interviewées de se sentir libres d'exprimer leur ressenti, leurs attentes et leurs difficultés concernant la gestion des données de la recherche. Un guide d'entretien a été conçu en amont par le groupe de travail. La trame de l'entretien suit celle de l'enquête en ligne (profil du répondant, généralités sur les données de la recherche, stockage des données, archivage des données, publication des données, attentes et besoins).

Il n'a pas été réalisé de transcriptions intégrales d'entretiens cependant les extraits significatifs ont été transcrits et catégorisés.[1]

L'analyse générale de ces extraits fait apparaître un déséquilibre sur certaines catégories :

- La thématique de l'**archivage** est la plus déstabilisante pour une grande partie des personnes interviewées de toutes les disciplines,
- Au contraire, la thématique **« Généralités »** rassemble le plus d'extraits car elle permet aux personnes interviewées d'expliciter aisément leurs matériaux de recherche, et d'aborder des points importants à propos de la gestion des données : définitions, terminologies, données sensibles, plans de gestion des données.
- La catégorie consacrée au **« stockage des données »** laisse apparaître à la fois des stratégies personnelles et institutionnelles diversifiées concernant cet aspect.
- Les échanges de la partie dédiée à la **« publication des données »** génèrent quant à eux des réactions plus marquées.
- Enfin, pour ce qui est de la catégorie des **« attentes »**, elle vient éclairer de façon pertinente les enseignements de l'enquête quantitative.

[1] En utilisant le logiciel MAXQDA Analytics Pro accessible grâce au soutien du dispositif DLab SHS de l'UVSQ

1.3. Biais et limites de la démarche

De façon générale, il paraît vraisemblable que les répondant·es représentent une part plus sensibilisée à la question des données que le public ciblé par l'enquête, et que la communauté de recherche de Paris-Saclay soit dans son ensemble moins sensibilisée aux questions de données qu'elle n'apparaît dans les résultats de l'enquête.

Outre les biais concernant les profils des répondant·es (voir plus bas), le dépouillement des réponses a fait apparaître des limites, concernant notamment l'appréhension de certaines notions abordées dans le questionnaire.

1.3.1. Limites sémantiques

La publication des données de la recherche amène à s'interroger sur deux approches qui sont abordées de manière simultanée par les répondant·es de l'enquête :

- Le partage des données avec des collègues ou des partenaires, soit en cours de projet de recherche, avec les membres du projet ou tout autre acteur pertinent ; soit de manière plus ponctuelle à l'issue d'un projet de recherche, sur sollicitation d'autres équipes par exemple ;
- La publication des données, visant à en communiquer une version documentée et réutilisable, via des outils dédiés garantissant un accès ouvert et pérenne au plus grand nombre.

La distinction entre les deux approches semble, au vu de certaines réponses ponctuelles, avoir pu être source de confusion pour certains des répondant·es.

Bien que les termes « partage » et « diffusion » aient pu être utilisés dans la rédaction des questions de l'enquête, nous préférons désormais privilégier l'utilisation du terme « publication », en tout cas dans toutes nos actions à venir et, autant que possible dans la rédaction de ce rapport d'enquête. La publication des données se fait dans un cadre défini, de la même façon, ou d'une façon très proche de la publication des articles scientifiques (relecture critique, traçabilité, paternité claire dans la citation, identification pérenne). Le choix de ce terme est donc important afin que son utilisation fasse appel aux mêmes référentiels, dans l'esprit des publics ciblés, que pour les articles scientifiques.

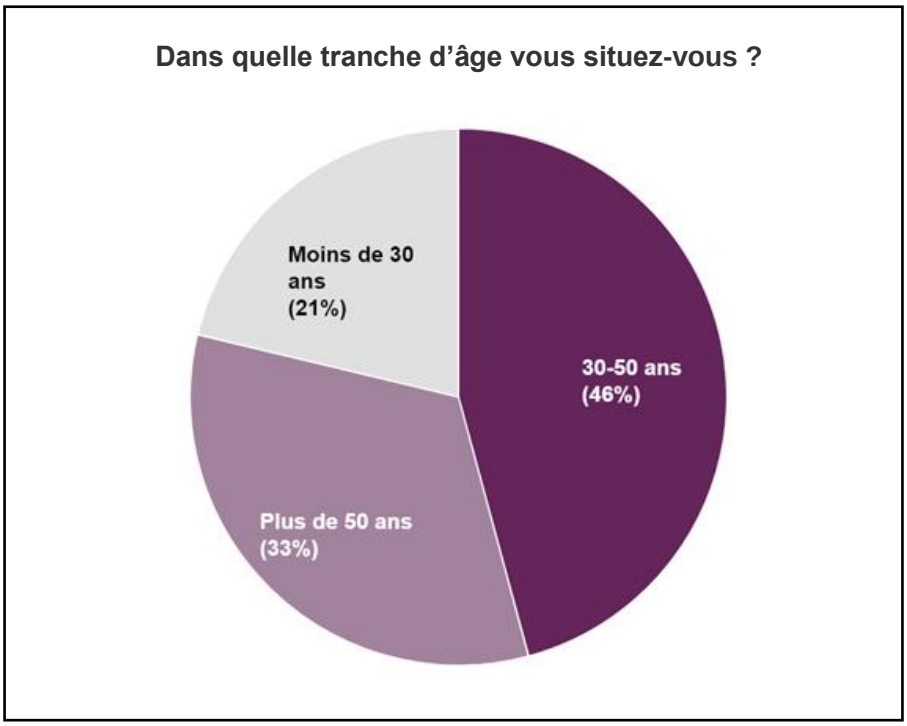
1.3.2. Limites méthodologiques

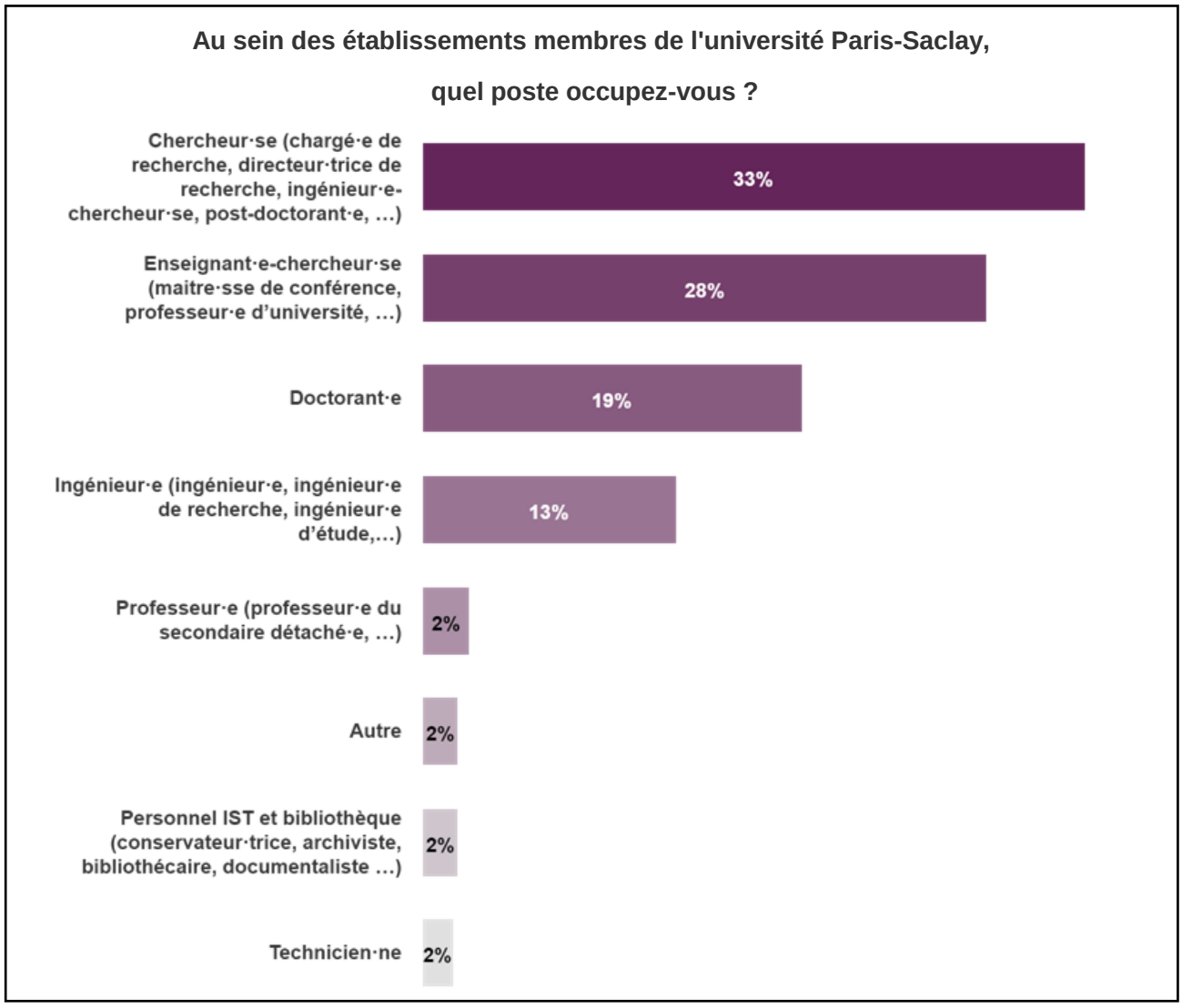
Lors du dépouillement, une difficulté est apparue dans le traitement des réponses : la question de la discipline de recherche du répondant était à choix multiples, et certains ont donc donné plusieurs réponses. Certaines analyses, quand elles impliquent notamment de croiser des données à la question de la discipline, ont donc pu s'en trouver biaisées.

D'autre part, à la question du poste occupé, aucun répondant n'a sélectionné la proposition « praticien hospitalier ». Cette réponse a donc été systématiquement écartée des analyses et des graphiques.

Pour finir, bien que ce filtre n'ait pas été utilisé en tant que tel pour l'analyse des données, il convient de noter une disparité entre établissements du périmètre de l'Université Paris-Saclay parmi les répondant·es, qui ne semble pas refléter entièrement la proportion de leurs effectifs respectifs : si la majorité des réponses provient du périmètre employeur de l'Université Paris-Saclay, du CEA et du CNRS (plus de 57% à eux trois), à l'inverse, les établissements-composantes sont sous-représentés (8 réponses pour AgroParisTech, 18 réponses pour CentraleSupélec, 6 réponses pour l'ENS Paris-Saclay, 3 réponses pour l'IOGS). Certains organismes de recherche semblent également peu présents au regard de leurs effectifs sur le périmètre recherche de l'Université : INRIA, INRAE et INSERM par exemple cumulent ensemble moins de 7% des réponses. Il est ainsi possible que certaines analyses puissent être davantage affinées par les actions de terrain et de cartographie menées sur ces périmètres.

1.4. Profils des répondant·es

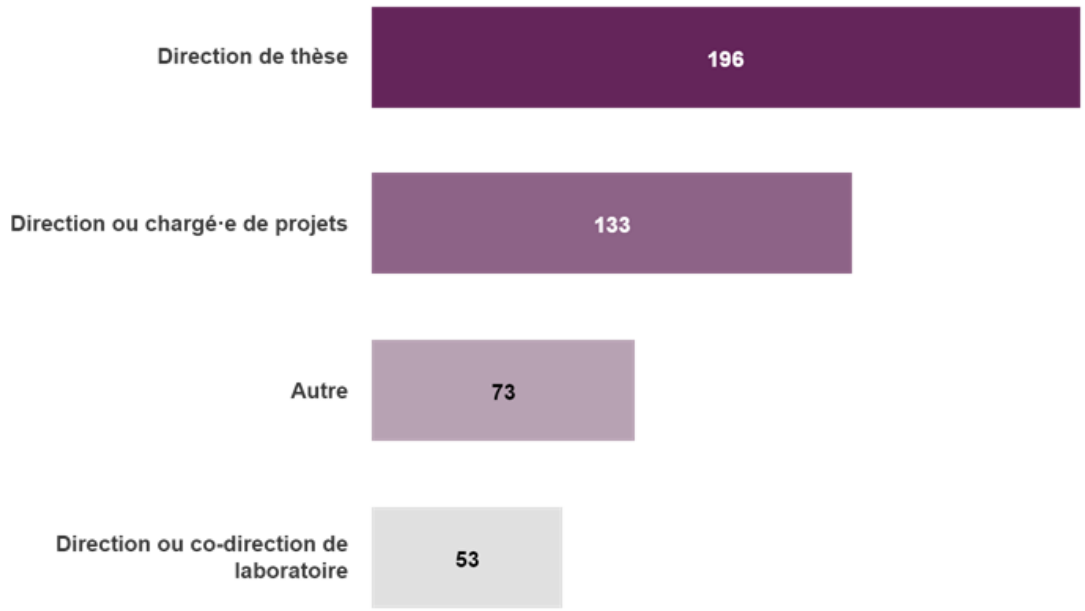




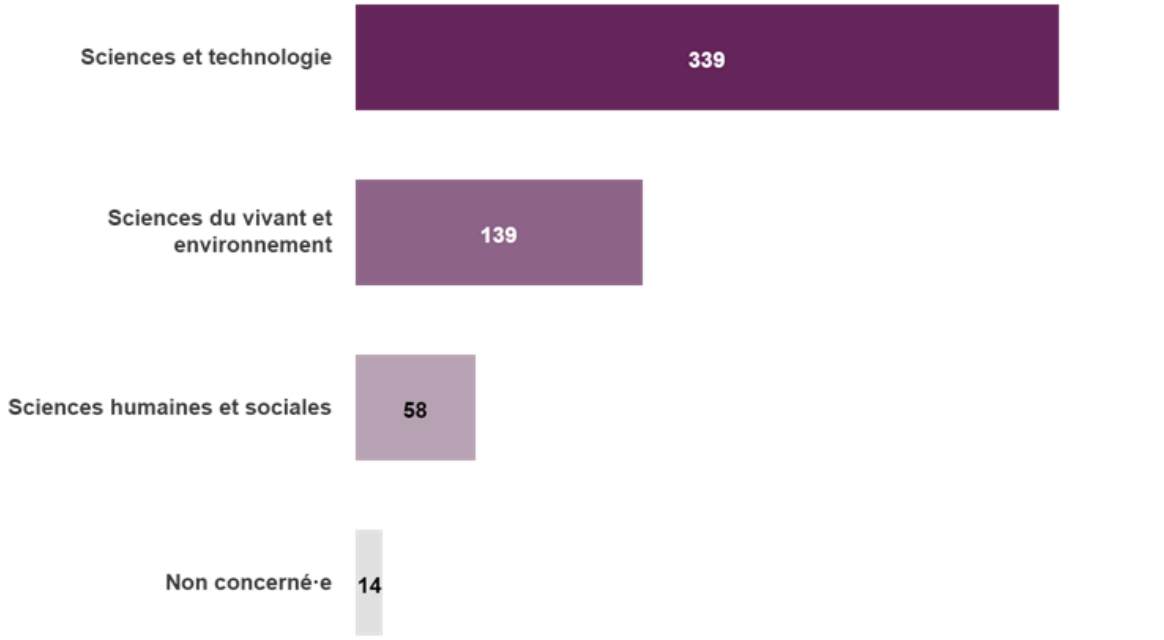
Les 513 répondant·es à l'enquête en ligne suivent un schéma de répartition assez habituel au sein des communautés scientifiques françaises : près de la moitié des répondant·es a entre 30 et 50 ans, et plus de 60 % d'entre eux sont chercheur·es ou enseignants-chercheur·es. Les doctorant·es (19 %) et ingénieur·es d'études ou de recherche (13 %) forment les deux autres catégories principales. Les autres catégories de répondant·es se situent toutes en dessous de 5 % : les professeurs du secondaire, les personnels IST et les technicien·nes. La catégorie "praticien hospitalier" qui été proposée n'a recueillie aucune réponse : elle a donc été supprimée des résultats de l'enquête.

Outre ce poste, occupez-vous d'autres responsabilités au sein de l'établissement ?

(297 répondant·es ; plusieurs réponses possibles)



Quel(s) est/sont votre/vos domaine(s) de recherche ? (plusieurs réponses possibles)



Cette répartition, et notamment la surreprésentation des enseignants-chercheur·es et chercheur·es, est caractéristique de nombreuses enquêtes sur les données de la

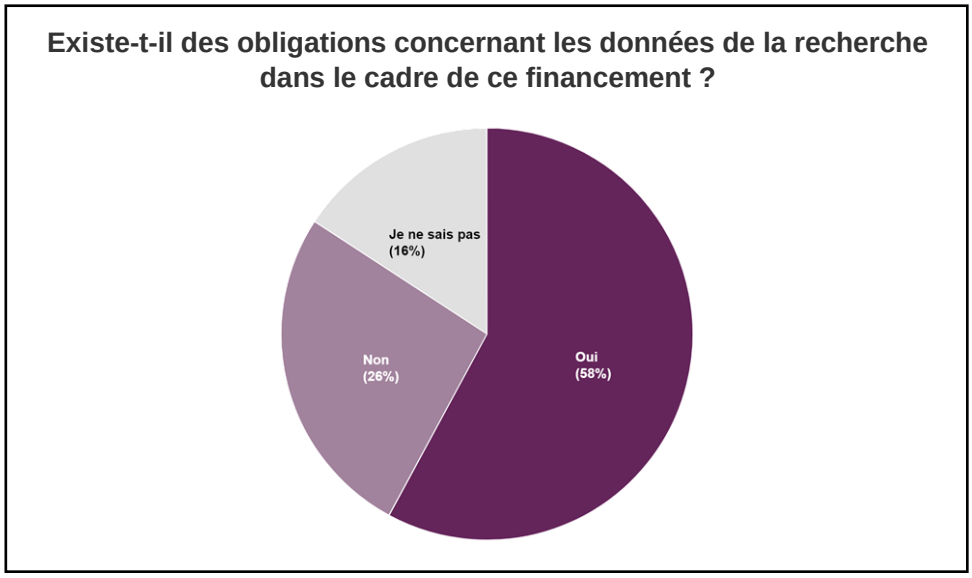
recherche. On peut émettre l'hypothèse que les autres catégories se sentent moins concernées ou légitimes dans ce domaine, et/ou qu'elles ne produisent / ne traitent pas de données de la recherche. Une question de l'enquête permet de cerner plus précisément le degré de représentativité des postes dit « de responsabilité », qui représentent de fait plus de la moitié des répondant·es au questionnaire. Parmi les responsabilités déclarées, il y a 38,2 % de directeurs ou directrices de thèse, 25,9 % de chargés et chargées de projets et 10,3 % de directeurs ou directrices de laboratoire. On peut regrouper les autres responsabilités exprimées en plusieurs catégories : responsables de formations, responsables d'équipe, responsabilités électives (conseil de faculté, Graduate School, conseil d'école doctorale, élu du personnel…). Il y a donc aussi une légère sur-représentativité des fonctions à responsabilité parmi les répondant·es. Plusieurs raisons peuvent expliquer ce phénomène :

- La sur-représentativité des chercheur·es et enseignants-chercheur·es explique mécaniquement cet état de fait, puisque c'est parmi cette catégorie que l'on retrouve la très grande majorité des postes à responsabilité cités dans l'enquête.
- Le même biais mécanique s'applique aussi sur l'âge des répondant·es, puisqu'avec 33 % de répondant·es ayant plus de 50 ans la probabilité d'avoir un poste de responsabilité augmente.
- La question des données de la recherche est souvent abordée lors de la participation à des projets à financements publics ou européens qui incluent la gestion et la publication des données dans leurs obligations de financement. Les chercheur·es en contact direct avec ces projets ont donc une plus grande chance d'être sensibilisés à la gestion des données.
- La question des données est aussi récurrente dans le cadre des thèses, donc de la même façon, les encadrants de thèse sont plus familiers que d'autres chercheur·es avec ces enjeux.

La répartition par domaines de recherche est un reflet fidèle du spectre disciplinaire de l'Université Paris-Saclay, avec environ les deux tiers des répondant·es évoluant en sciences et technologies, suivis par les sciences du vivant et de l'environnement, et dans une moindre mesure par les sciences humaines et sociales.

Concernant le domaine des sciences et technologies, c'est la physique qui arrive en premier parmi les répondant·es (30 %), puis suivent l'informatique, les mathématiques, les matériaux et physique des solides, la mécanique des fluides et l'automatique, signal et image (de 14 à 16 % des répondant·es pour chacun). En sciences du vivant et de l'environnement, la biologie moléculaire et la biochimie sont fortement représentées (30 %), ainsi que la génétique (23 %) et les biotechnologies, sciences environnementales, biologie synthétique et agronomie (17 %). En sciences humaines et sociales enfin, l'histoire est majoritaire (26 %), suivie de la sociologie (19 %), les littératures et langues étrangères (17 %) et le droit (15 %).

Les porteurs et porteuses de projet



Le caractère obligatoire de la gestion de données dans le cadre des divers projets publics et/ou européens est une caractéristique connue (comme rappelé plus haut). Par conséquent, l'enquête s'attardait plus spécifiquement sur les porteurs et porteuses de projets (ANR à 75 %, H2020 à 38 %, privé à 44 %, de multiples réponses étant bien sûr possibles. On retrouve également comme financeurs les collectivités territoriales, des associations comme Sidaction ou Fondation de France, et des institutions, CNRS, INSERM, LabEx etc.). En effet, il semblait intéressant de faire un point particulier sur cette part des acteurs de la recherche, considérée normalement comme plus sensibilisée aux attentes relatives à la gestion des données de la recherche.

Les résultats issus de l'analyse de cette catégorie, reportée sur les autres questions du questionnaire, sont particulièrement intéressants. On peut d'ores et déjà noter que plus de la moitié de ces projets entraînent des obligations par rapport aux données de la recherche. Or une partie significative des porteurs et porteuses de projets ont rapporté leur ignorance sur ce point.

En conclusion, les profils des répondant·es correspondent à la communauté scientifique visée par l'enquête, mais présentent des biais de représentativité identifiés. Le choix méthodologique des entretiens individuels adossés au questionnaire en ligne permet d'apporter des nuances de réflexion et de mettre en valeur des profils sous-représentés dans l'enquête.

2. Définition et typologie des données de la recherche

L'enquête en ligne proposait une définition des données de la recherche : « L'expression « DONNÉES DE LA RECHERCHE » désigne généralement les sources, matériaux ou informations collectées et/ou produites au cours de recherches scientifiques ». Cette définition, voulue large, a été bien comprise par les répondant·es.

2.1. Une définition des données de la recherche parmi les répondant·es

En effet, 96 % des répondant·es sont d'accord avec cette définition. Parmi les 4 % à avoir exprimé un désaccord, la principale question semble être la place à accorder au code informatique (est-ce une donnée ou non ?), ainsi que des interrogations autour du contexte de la recherche, à inclure ou non dans les données, et les rôles des mots « sources » et « informations ». La confusion entre ces deux derniers termes dépend peut-être de disciplines où ces notions sont moins prégnantes. La place à accorder au code est une préoccupation qui revient régulièrement dans les tentatives de définition des données de la recherche, et qui avait été évacuée par la définition très générale choisie dans l'enquête. Il est significatif que plusieurs chercheur·es aient choisi de cibler plus précisément cet aspect de la donnée. Il est également à souligner l'importance du lien entre numérique et données de la recherche qui revient à plusieurs reprises en entretien :

> *« Néanmoins, cette idée qu'il y a des données de la recherche, c'est à dire, elle est à mon avis beaucoup due au fait qu'on a des machines qui peuvent manger les données à n'en plus finir. Mais ça dépend des domaines. » (Chercheur en Sciences et technologie)*

> *« Et donc voilà, je pense que les données de la recherche, c'est tout ce qui est numérisable. D'ailleurs on englobe dans les données de la recherche même les logiciels à la fin du compte parce que quand on dit open data et open software c'est un peu tout ça, c'est un peu le même monde quoi. » (Chercheur en Sciences et technologie)*

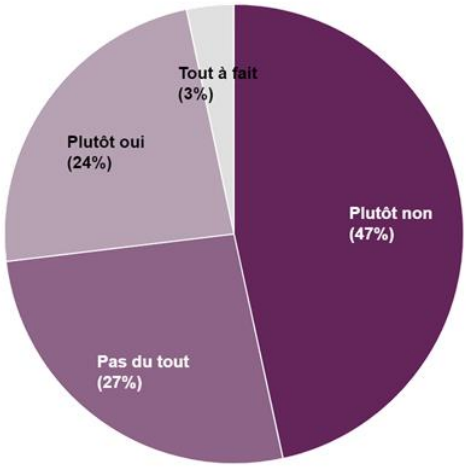
La question du contexte et des outils est aussi soulevée lors des entretiens avec les chercheur·es. Ainsi, un chercheur en sciences et technologie met en avant le lien entre les données, les outils et la vérifiabilité de la recherche :

« Pour moi les données de la recherche seraient les données qui sont collectées lors d'expérimentations et puis qu'on peut du coup mettre à disposition. Et pour moi, ça irait même au-delà des mesures expérimentales, et ça inclut aussi, puisque j'y suis sensibilisé, toutes les chaînes de traitement, tous les outils logiciels, toutes les bibliothèques. Enfin voilà, c'est très bien d'avoir les données, c'est très bien d'avoir les données traitées, mais

si je n'ai pas la capacité de refaire le cheminement pour arriver aux données traitées, ce n'est pas vérifiable et pour moi du coup c'est moins intéressant quoi. »

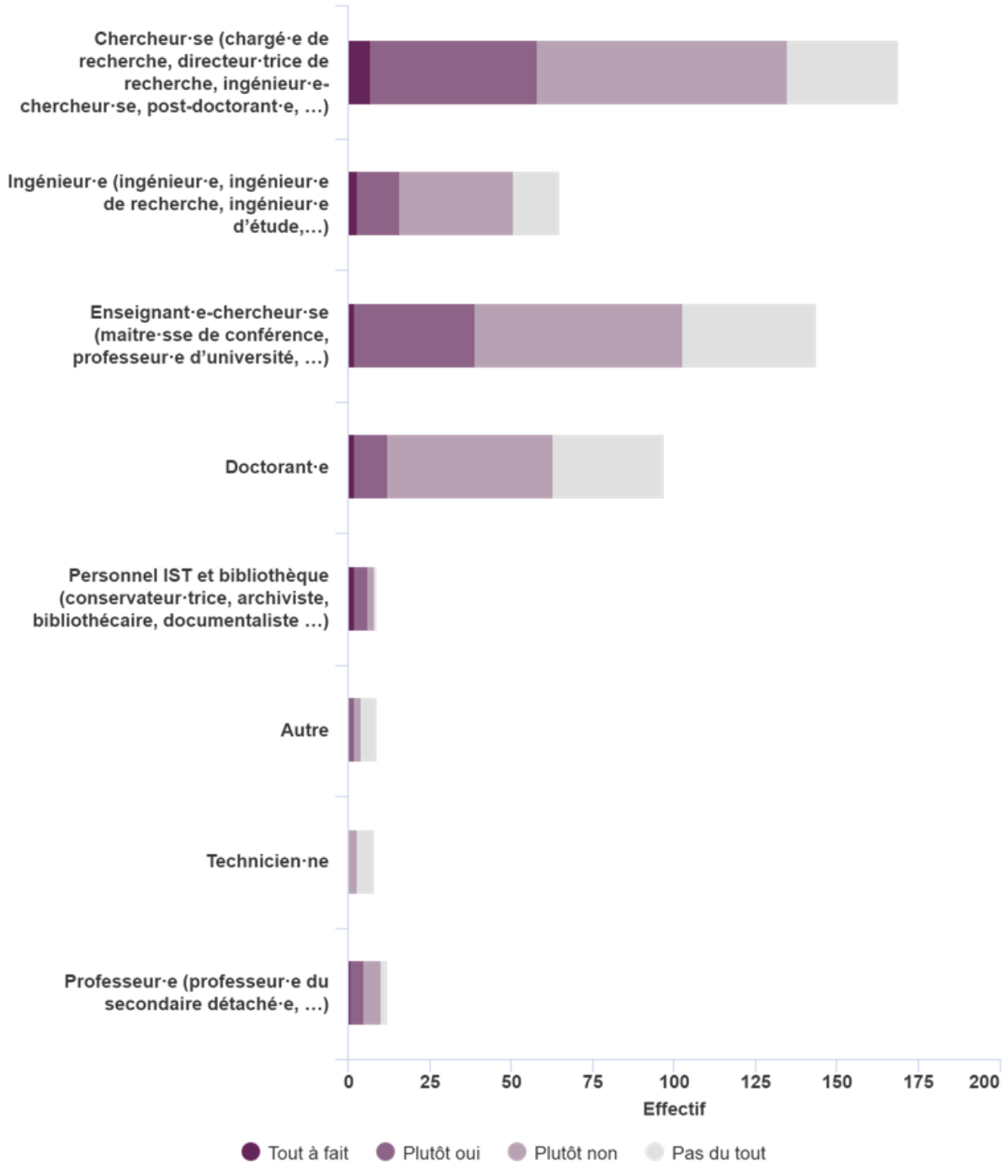
2.2. Une connaissance du contexte des données de la recherche à géométrie variable

Avez-vous connaissance des directives ou recommandations à l'échelle nationale et/ou internationale pour la gestion des données de la recherche ?



Concernant leur connaissance des directives ou recommandations à l'échelle nationale et/ou internationale, les répondant·es affirment en grande majorité n'en avoir plutôt pas (46,6 %) ou pas du tout (26,5 %). Parmi les directives ou recommandations connues, sont citées principalement celles de l'ANR, du Plan National pour la Science Ouverte, et de H2020.

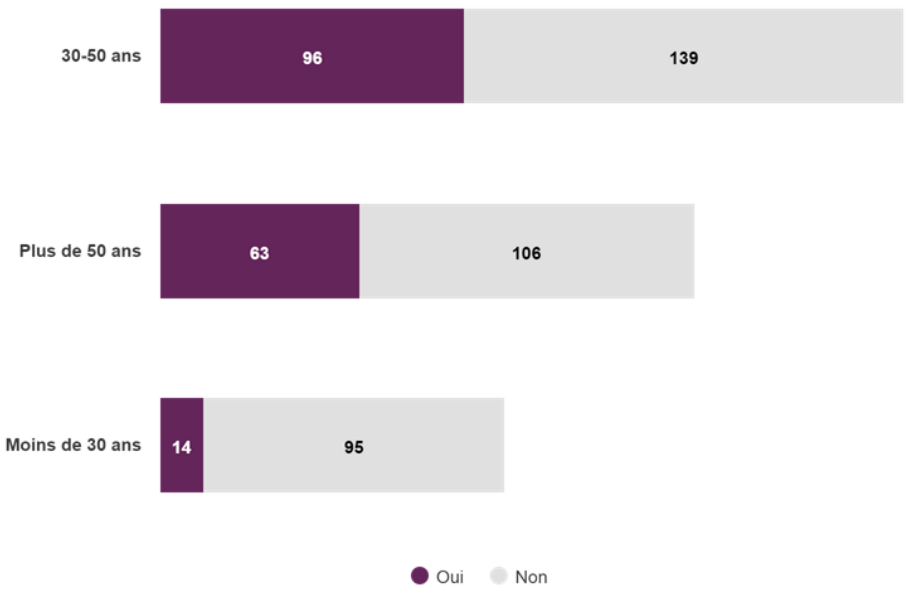
Avez-vous connaissance des directives ou recommandations à l'échelle nationale et/ou internationale pour la gestion des données de la recherche en fonction du poste que vous occupez ?



Sans surprise, les chercheur·es sont les mieux informés à ce sujet, aucun d'entre eux n'a par exemple indiqué n'être pas du tout au courant de ces directives, même si 32 % n'en ont qu'une connaissance approximative. Les doctorant·es sont par contre beaucoup moins au fait des recommandations nationales/internationales sur le sujet, ce qui peut

s'expliquer par le fait qu'ils n'y ont pas ou peu eu affaire. Sans surprise également, il y a une corrélation entre la responsabilité des répondant·es et leurs connaissances des directives nationales et internationales autour des données : 43,4 % des directeurs et directrices ou co-directeurs/co-directrices de laboratoire, 37,6 % des responsables de projets et 32,1% des directeurs et directrices de thèse en ont une plutôt bonne connaissance.

Avez-vous déjà entendu parler de « Plan de gestion des données de recherche » (PGD), ou de son équivalent anglais, « Data Management Plan » (DMP) en fonction des tranches d'âge ?



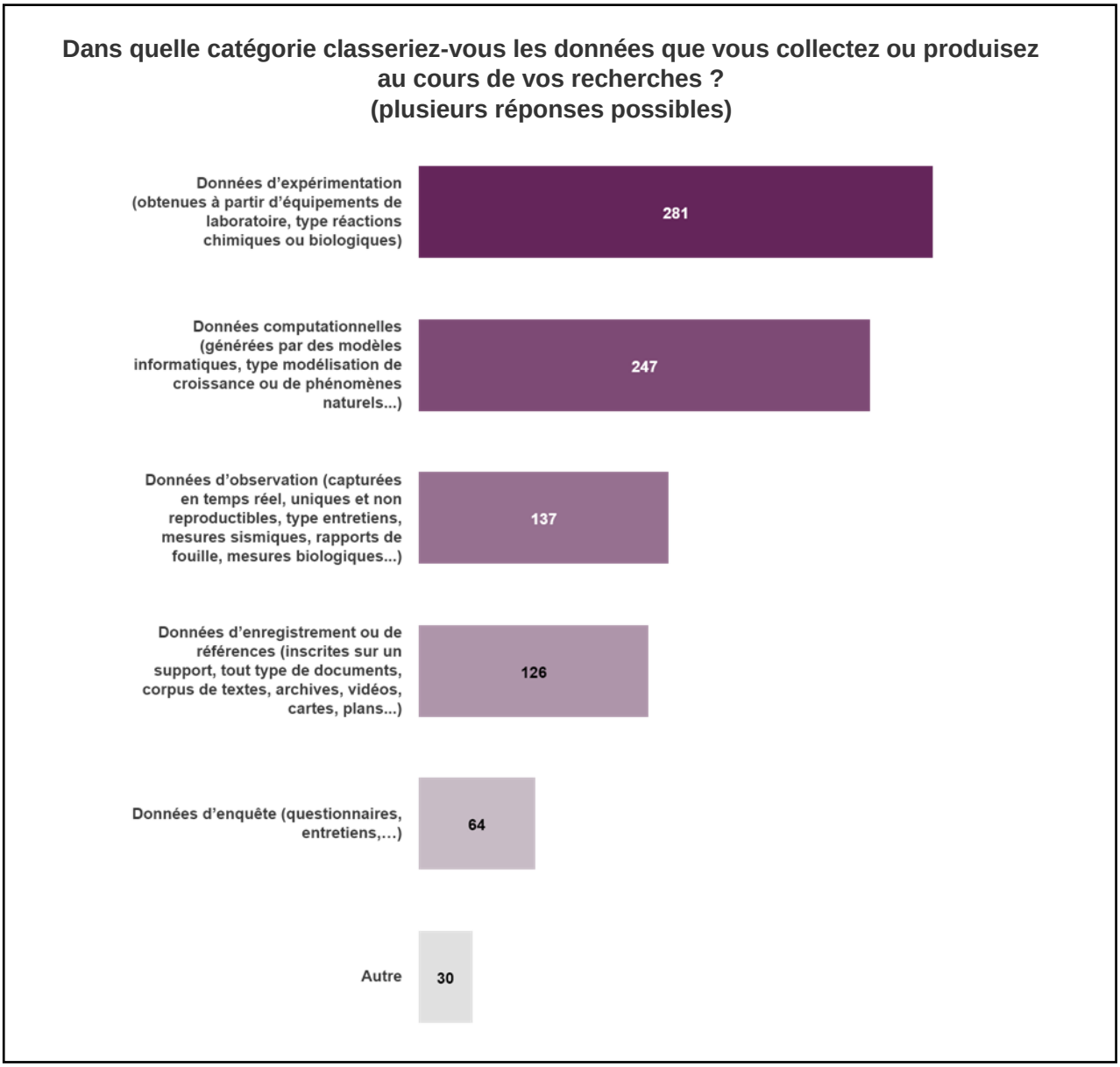
Néanmoins, une grande majorité des répondant·es (66,3 %) n'a pas entendu parler des plans de gestion de données (PGD, ou DMP en anglais, pour *Data Management Plan*). Cette méconnaissance est d'autant plus vraie pour les moins de 30 ans, ce qui recoupe les observations précédentes.

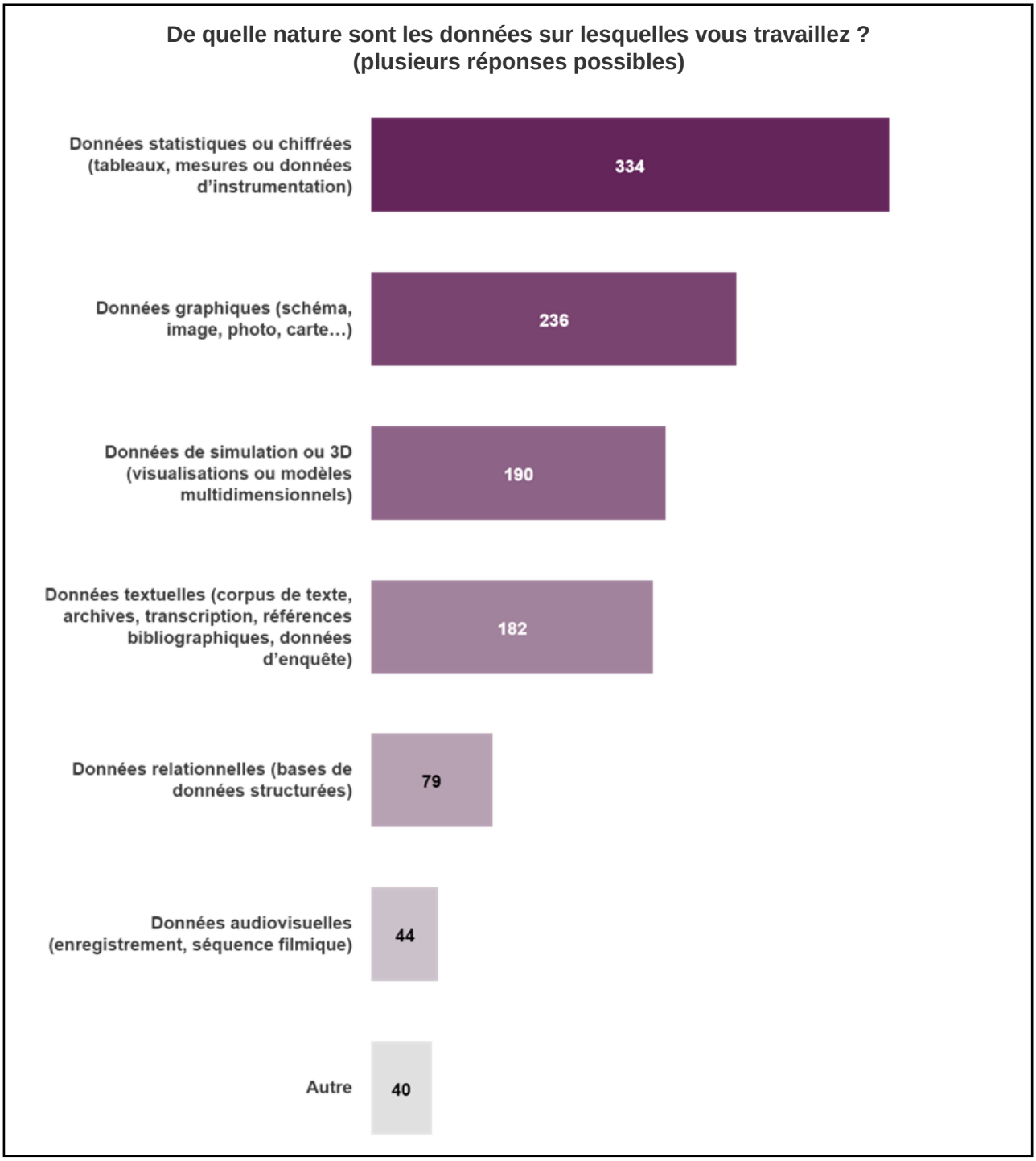
Si l'on met en concordance la connaissance de ce qu'est un PGD et le poste occupé/la responsabilité, notons que la moitié est des chercheur.es (50,3%), 60,4 % des directeurs/directrices ou co-directeurs/co-directrices de laboratoire et 53,4 % des responsables de projets.

On observe donc une forte disparité au sein de la communauté scientifique visée par l'enquête : suivant l'expérience, les responsabilités ou les fonctions des répondant·es, le niveau d'information basique (directives nationales, PGD) sur les données de la recherche subissent une forte corrélation. Les disciplines en revanche sont moins significatives. On pourrait en conclure que le savoir relatif aux données de la recherche est de type empirique, c'est-à-dire qu'il repose sur une confrontation directe entre l'acteur de la recherche et une situation où il est en mesure de se poser des questions relatives aux

données (projets, plan de gestion de données etc.). Les données sembleraient peu présentes à l'inverse dans le parcours de formation du doctorant.

2.3. Typologie des données





Les données sont majoritairement sous forme numérique pour tous les domaines. On constate une part plus importante des données papier en sciences humaines et sociales mais cette tendance est à nuancer compte tenu du faible panel de répondant·es. 53 % des répondant·es indiquent produire du code informatique, ce qui est un pourcentage significatif, et contribue à expliquer pourquoi l'inclusion ou non du code sous le terme « donnée de la recherche » est porteuse d'enjeux pour de nombreux répondant·es. La problématique était ainsi fréquemment abordée pendant les entretiens :

« Maintenant on commence à rentrer dans le dur, c'est à dire associer vraiment la gestion de données ouvertes avec le logiciel libre. Moi je pense qu'on ne peut pas raisonnablement faire des données ouvertes avec des logiciels fermés. » (Chercheur en Sciences et technologie)

La grande majorité des répondant·es ne collectent pas de données à caractère personnel (75 %) ni de données confidentielles liées au secret industriel (69 %). Il est à noter qu'une partie des répondant·es ne peut pas déterminer le caractère « personnel » des données (10 % de « je ne sais pas »).

C'est en sciences du vivant que l'on rencontre le plus de données sensibles (plus de 80 %), alors que les données liées au secret industriel (environ 28 %) se trouvent le plus souvent en sciences du vivant et en sciences et technologies.

3. Pratiques des données de la recherche

La partie la plus importante du questionnaire et des entretiens se concentre sur les pratiques effectives des acteurs de la recherche concernant les données, et suit un découpage centré sur le cycle de vie de la donnée : l'enquête et les entretiens s'intéressent successivement à la manière dont sont produites les données, à la gestion des données, au stockage pendant la production, à l'archivage après publication, à la publication et à leur réutilisation.

3.1 Gestion au sein des laboratoires : les données intégrées au fonctionnement des unités de recherche

La gestion des données de la recherche s'inscrit dans les problématiques actuelles de la vie des laboratoires. Si l'on analyse dans cette partie les questions qui n'ont été posées qu'aux directeurs et directrices de laboratoires ou d'unités de recherche : 70 % des répondant·es ont déclaré que cette question a déjà été abordée ou discutée avec leurs collègues de laboratoire. L'enquête montre ainsi un intérêt pour les enjeux actuels relatifs aux données de la recherche. Il semble avéré qu'une réflexion est menée en interne au niveau des laboratoires sur le rôle et la place des données dans le travail de recherche.

Les réponses au questionnaire ne révèlent pas une différence significative en fonction du domaine de recherche, même si on peut voir une légère prépondérance dans les sciences et technologie (76,6 %) par rapport aux sciences du vivant et environnement (70,6 %) et aux sciences humaines et sociales (50 %). Sur l'ensemble des répondant·es au questionnaire, 42 % ont indiqué qu'au sein de leur laboratoire existe déjà une personne référente sur la gestion des données de la recherche.

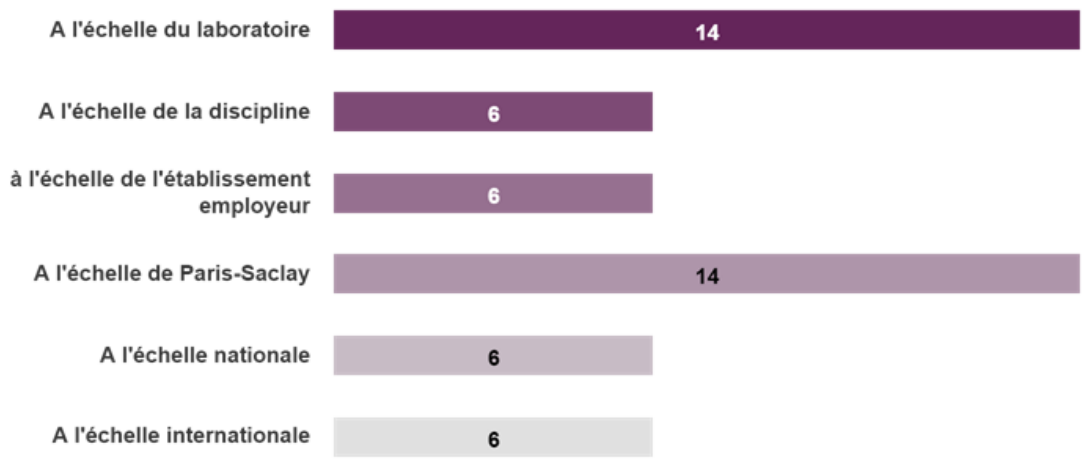
Dans la plupart des cas (81,8 %), cette personne a un profil d'ingénieur·e (de recherche ou d'études). Cela permet d'affirmer que, au niveau des laboratoires, on attribue à un

ingénieur·e la fonction et la mission d'accompagnement des chercheur·es dans la gestion des données de la recherche.

On constate également que la majorité (78 %) des répondant·es se montre favorable à une politique de gestion des données scientifiques : 38 % envisagent de mettre en place une telle politique au niveau du laboratoire ou de l'unité de recherche. Il n'est pas observé de différence substantielle en fonction du domaine de recherche, ni du poste occupé au sein de l'Université Paris-Saclay. Toutefois, on peut constater que les chercheur·es et chercheuses manifestent davantage un intérêt positif envers cette politique, probablement car ils sont directement concernés. La relation avec l'établissement employeur n'induit pas non plus de différences.

En revanche, on note une corrélation significative entre la considération d'une politique de gestion des données au niveau du laboratoire ou de l'unité de recherche et l'avis sur la nécessité d'une politique commune.

A quelle échelle pensez-vous qu'une politique commune pour la gestion des données de recherche soit nécessaire ? (25 répondant·es ; plusieurs réponses possibles)



De l'ensemble des répondant·es, 47 % déclarent qu'une politique commune de gestion des données de la recherche est nécessaire.

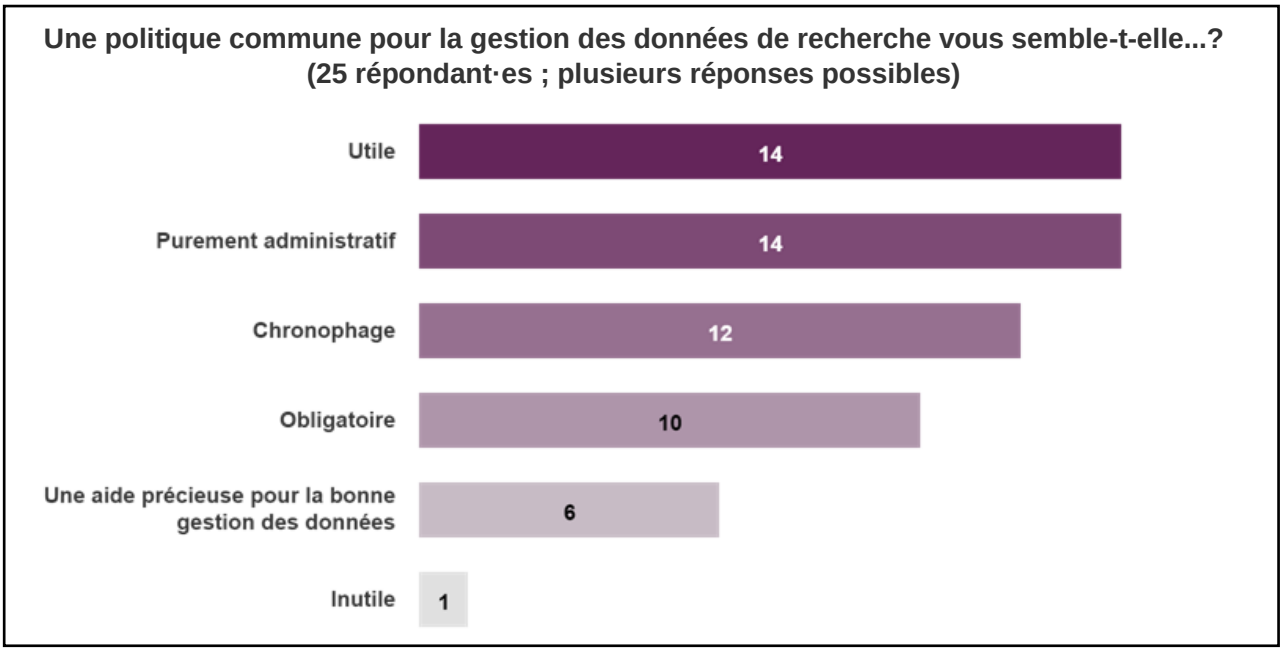
56 % des répondant·es voient la nécessité d'une politique commune à l'échelle du laboratoire, exactement le même pourcentage que ceux qui placent cette politique à l'échelle de l'université Paris-Saclay.

En ce qui concerne le domaine de recherche, 44 % des répondant·es en Sciences du vivant et l'environnement, 56 % en Sciences et technologie et 50 % en Sciences Humaines et Sociales (bien que moins significatif car sous-représentés dans l'enquête) se prononcent pour la nécessité d'une politique commune de gestion des données de recherche.

3.2. Le plan de gestion de données : entre *a priori* négatifs et obligation des tutelles

Le plan de gestion de données est loin d'être connu par tous les acteurs du monde de la recherche, notamment dans ses aspects techniques. Un professeur d'immunologie rapporte en entretien sa conception empirique d'un PGD. La confusion entre pratiques de stockage et d'archivage et planification de gestion des données dans le cadre d'un projet, commune à plusieurs des interlocuteurs, révèle une certaine familiarité avec le terme sans toutefois pouvoir lui associer un contenu précis.

> *« Plan de gestion de données. Alors ça me parle, mais pas sous ce terme-là, mais c'est plus comment est-ce qu'on pratique ses sauvegardes ?, comment est-ce qu'on manage ses données, c'est à dire comment est-ce qu'on les intitule, où est-ce qu'on les range, c'est ça que vous voulez dire ? »* (chercheur en sciences du vivant et environnement)

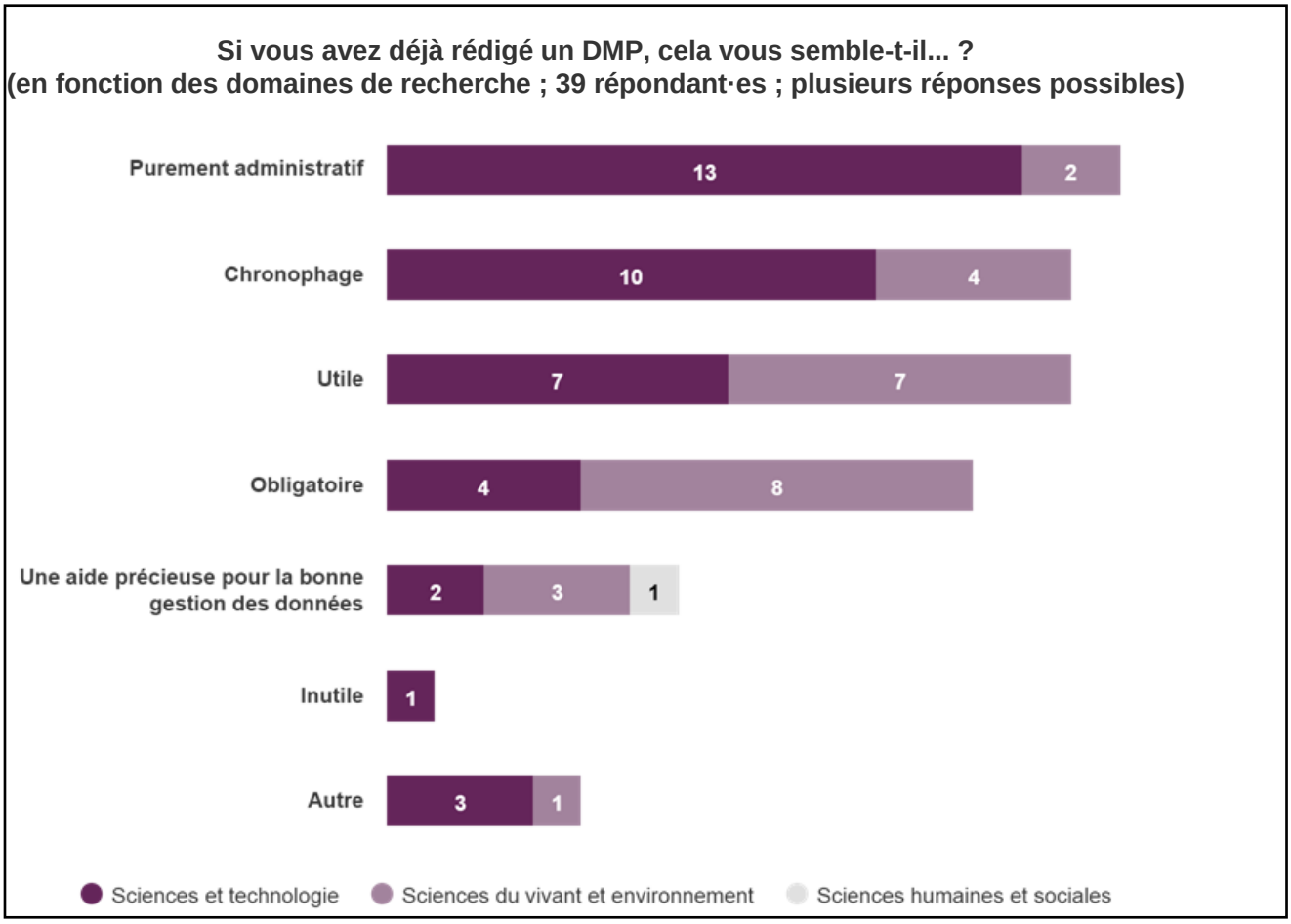


C'est parmi les personnes ayant déjà eu affaire à une obligation institutionnelle de PGD que l'on retrouve les avis les plus significatifs. Parmi les répondant·es de l'enquête, 23 % ont déjà rédigé un plan de gestion de données. Les *a priori* sur le PGD sont globalement négatifs : il est souvent vu comme purement administratif, chronophage et obligatoire, même si 36 % lui reconnaissent une utilité.

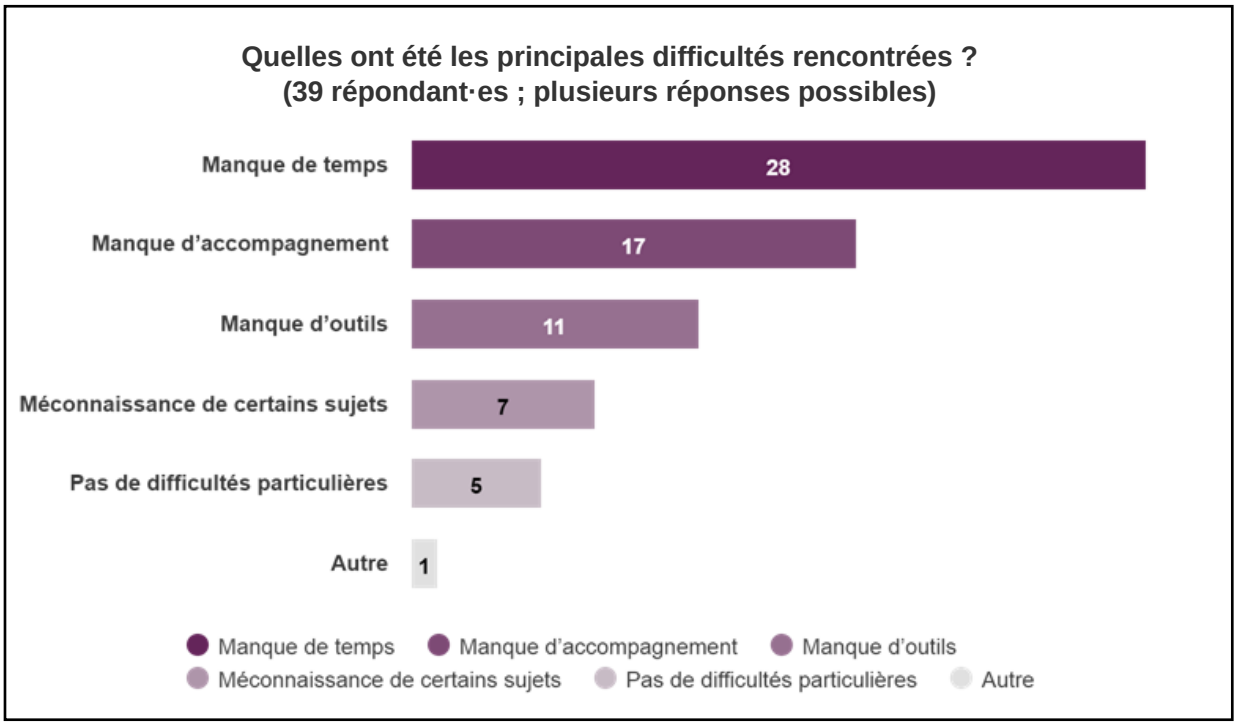
Cette vision contrainte du PGD se retrouve aussi dans les entretiens avec les chercheur·es, comme cet enseignant-chercheur en sciences et technologie qui relate son expérience avec les plans de gestion de données :

« C'est juste une contrainte. C'est-à-dire que ça ne nous règle pas les vrais problèmes qui sont : comment est-ce qu'on peut garder des données longtemps ? Qu'est-ce qu'on garde et qu'est-ce qui est partageable largement et avec quelles informations ? Et ça, moi je veux bien faire des plans, mais ça ne va pas répondre à mes questions. » (chercheuse en sciences et technologie)

Aucun répondant de moins de 30 ans ne s'est prononcé sur l'utilité du PGD, ce qui renforce l'hypothèse d'une approche par l'expérience.



L'analyse par discipline révèle une plus forte vision de l'utilité d'un PGD en sciences du vivant et de l'environnement.



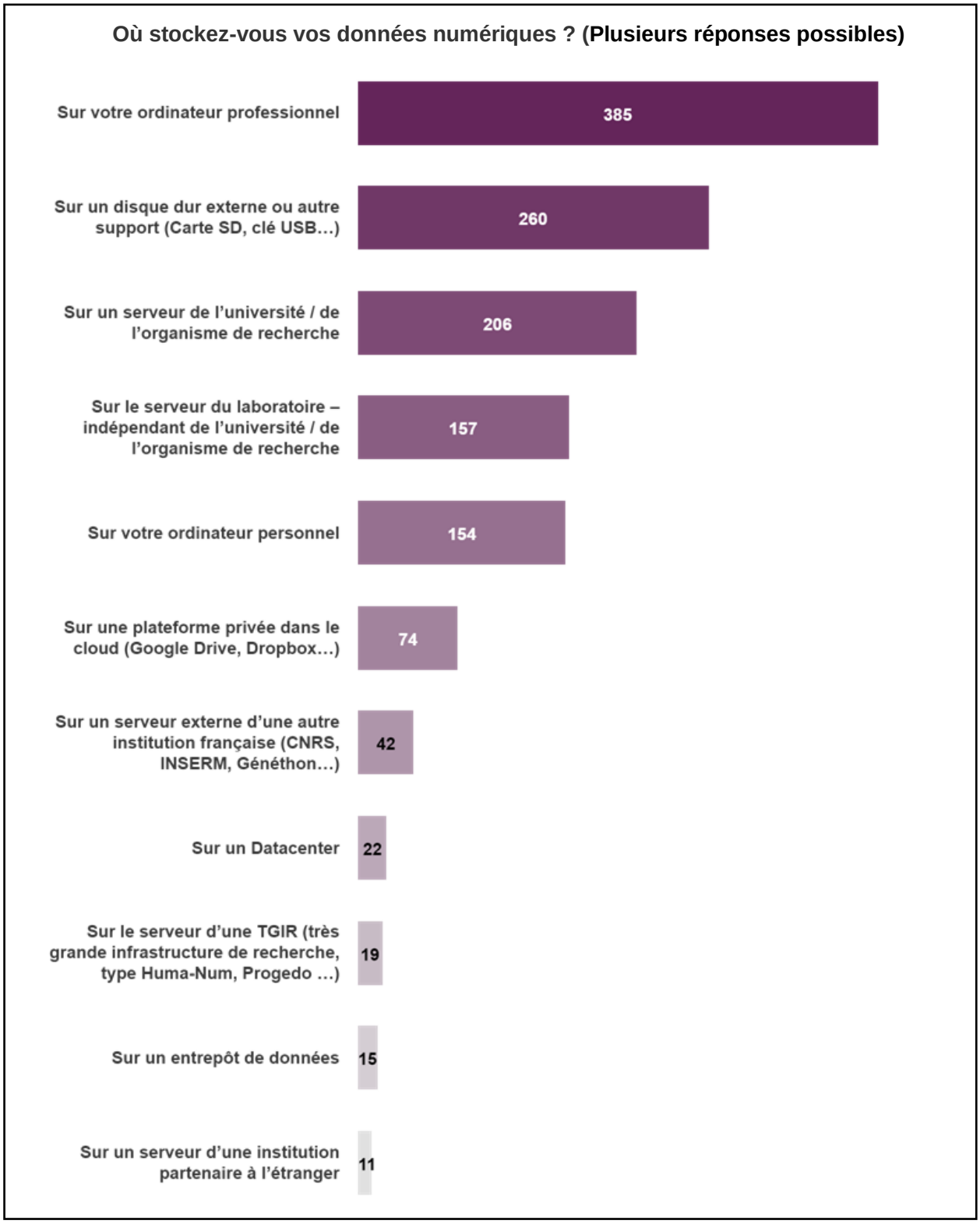
À la question « Quelles ont été les principales difficultés rencontrées pour remplir un PGD ? », les réponses appellent clairement une aide pour cette tâche et le caractère chronophage ressort de nouveau.

Cependant, pour cet accompagnement, les répondant·es se réfèrent très majoritairement à leurs collègues, à plus de 59 % ; viennent ensuite le laboratoire en général (10 %), la Direction de la recherche, la bibliothèque, le service juridique (tous à 5 %). 23 % d'entre eux n'ont pas trouvé d'interlocuteur pour les aider face à leurs difficultés ; on relève ici un manque de visibilité des interlocuteurs experts sur cette question.

Pour conclure, si la méconnaissance de certains sujets est évoquée parmi les difficultés rencontrées pour remplir un PGD, il ressort un manque d'information juridique, une demande d'accompagnement, et l'interrogation sur « l'Intérêt réel d'une telle démarche stérile et destructrice de temps de recherche ? » (phrase prononcée lors d'un entretien individuel) malgré une utilité reconnue par une minorité.

3.3. Le stockage et l'archivage au cœur des pratiques des données de la recherche

Le principal enseignement de l'enquête en ce qui concerne le stockage et l'archivage des données est qu'une importante variété de situations existe selon les acteurs de la recherche, les établissements, les disciplines, les laboratoires : aucune politique de conservation ne se dégage réellement à une échelle supérieure à celle de l'individu.



Les pratiques de stockage sont personnelles, propres à chaque acteur de la recherche et multiplient souvent les supports, comme l'explique ce doctorant en sciences humaines et sociales pendant les entretiens :

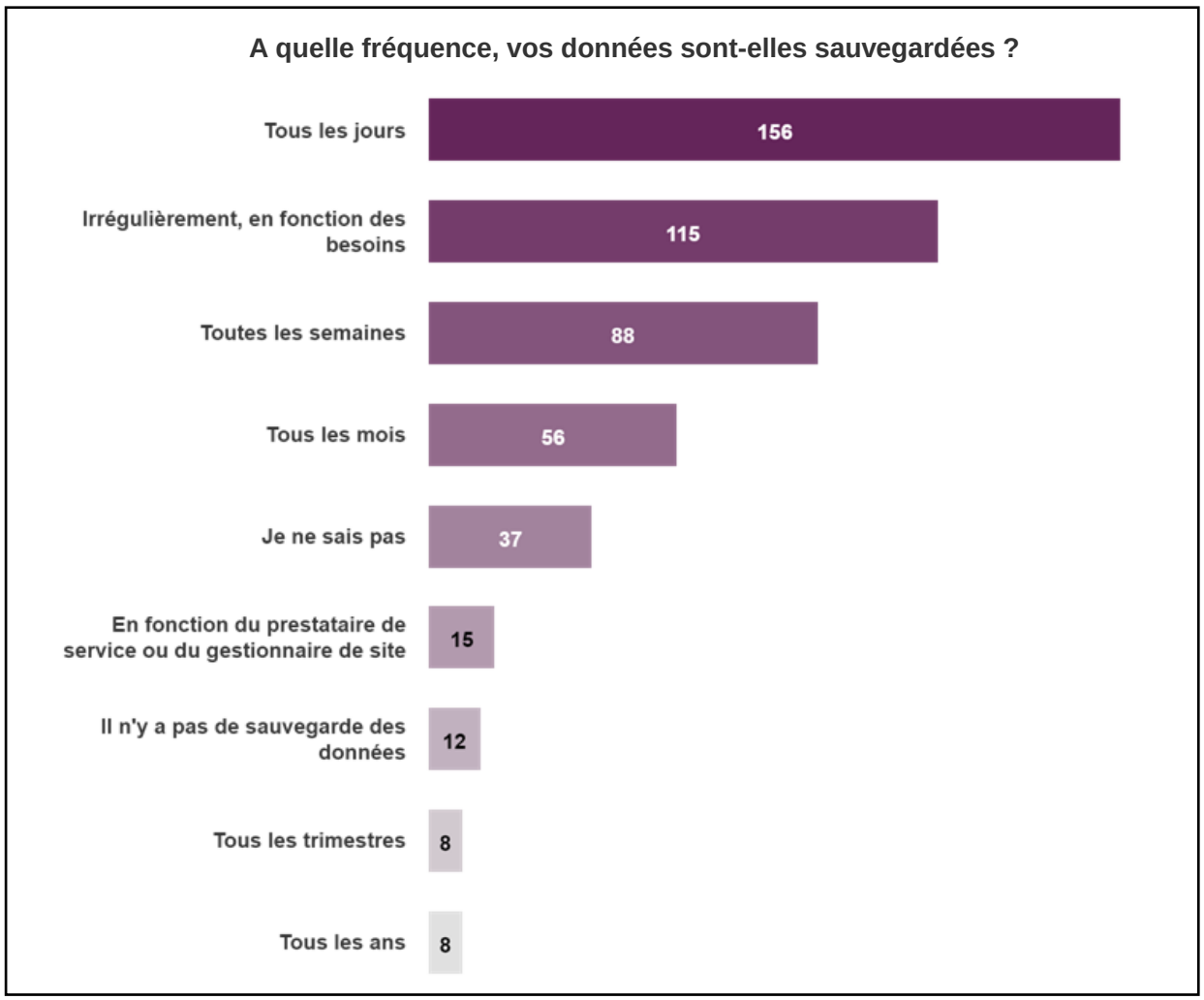
> *« Comme tout doctorant, je pense, ou chercheur, la peur de perdre ses données motive mes pratiques. Donc j'ai démultiplié mes supports de stockage, j'ai un ordinateur, j'ai un deuxième ordinateur, que j'utilise peu et qui est stocké ailleurs, et puis j'ai deux disques durs que je mets à jour tous les mois au minimum. »*

La multiplication des supports de stockage est souvent vue comme une sécurité supérieure dans la sauvegarde des données, mais elle relève aussi très majoritairement de démarches et d'outils personnels : clé USB, disques durs, ordinateurs professionnels ou personnels, cloud personnel etc. Les solutions institutionnelles sont moins utilisées, à l'exception des serveurs de laboratoire, ou d'université. Lorsque le serveur institutionnel est utilisé, il est presque toujours doublé par une autre sauvegarde matérielle pour les répondant·es de l'enquête. L'une des raisons de cette double sauvegarde systématique est explicitée par une ingénieure en sciences du vivant et environnement, qui met en avant la gestion parfois difficile de ces serveurs :

> *« Très souvent on nous dit qu'il n'y a plus de place, il n'y a plus de place. Ah non, mais vous ne pouvez pas imaginer, hein. Parce que l'EEG, par exemple, on a une étudiante, c'est plusieurs téra de données puisque les bébés dorment pendant 1 h et il y a 5 ans de données donc c'est énorme. Et c'est régulièrement qu'on nous dit qu'il faut faire du ménage, voilà, ça, ça fait partie du travail en équipe avec les ingénieurs des autres labos. »*

La peur de perdre ses données est le moteur primordial des pratiques de sauvegarde multiples. Plusieurs répondant·es et interviewés en entretien expriment aussi une préoccupation face à l'utilisation de serveurs privés, tout en leur reconnaissant une facilité d'utilisation. Par exemple, cette pratique d'hébergement de données d'un doctorant en sciences humaines et sociales :

> *« J'essaye au maximum d'auto-héberger toutes les données qui sont à ma disposition. C'est un serveur, bon ce n'est pas l'idéal, que je loue chez OVH et sur lequel je fais tourner une instance Nextcloud. »*



La fréquence de sauvegarde des données est majoritairement connue des répondant·es de l'enquête. L'intervention manuelle que demande la manipulation personnelle des moyens de sauvegarde matériels (clés, disques durs…) explique une fréquence irrégulière importante (23 %), tandis que le recours à des serveurs ou des solutions extérieures suppose une sauvegarde quotidienne (32 %) : ces deux solutions sont perçues comme plus sécurisantes, bien que découlant de logiques très différentes.

Très liée à la question du stockage, la perte de données est une expérience courante : plus du tiers des répondant·es en ont subi.

Sur les 169 répondant·es concernés, une forte proportion a subi des pertes importantes - près de 29% ont perdu plus d'1 To de données, 23% entre 100 Go et 1 To, près de 18% entre 20 et 100 Go.

141 répondant·es ont donné des détails sur les circonstances de la perte de données, certaines réponses comportant la relation de plusieurs incidents, soit 166 au total.

Plus de la moitié (96 réponses, soit 58% du total) relève de problèmes matériels : formatage, pannes, corruption de données, coupure électrique, virus, incendie ont

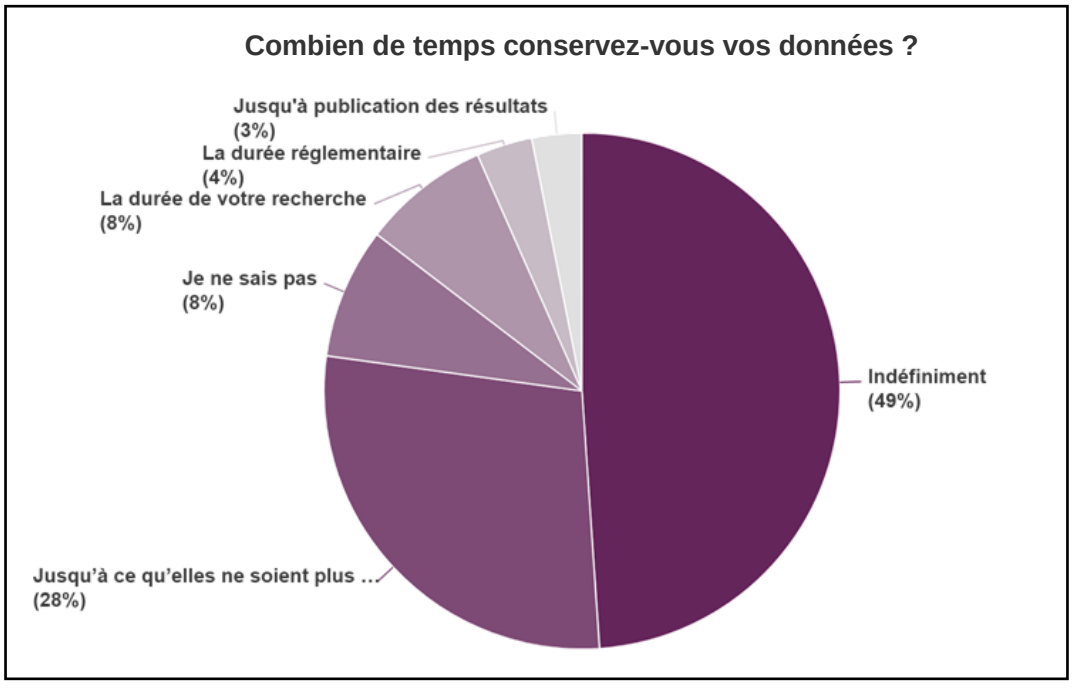
occasionné des pertes totales ou partielles. Parfois, il s'agit d'une panne du système institutionnel, ce qui rend bien pertinent le fait de ne pas s'appuyer uniquement sur un service, même rendu par l'institution.

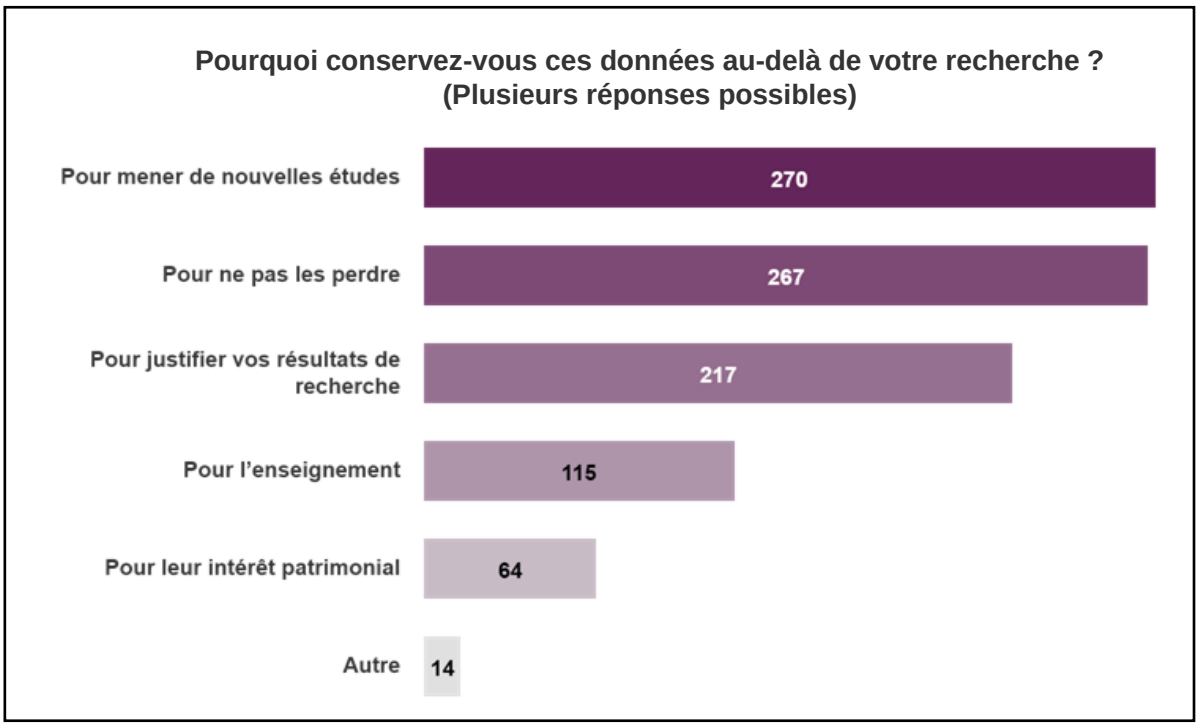
La 2e raison est l'obsolescence des matériels, des logiciels (24 réponses, soit 14%) : « impossibilité d'ouvrir d'anciens formats numériques », « stockage ancien sur support type disquette - ou via des logiciels anciens n'ayant pas de versions utilisables sous un système d'exploitation actuel ».

Viennent ensuite à part égale (14 réponses, 8%), des erreurs de manipulation, et l'absence ou incomplétude du suivi, qu'il s'agisse de doctorant·es/étudiant·es parti·es avec les données, ou de cahiers de laboratoire incomplets. Un répondant détaille : « Nous avons des données aveugles, provenant d'étudiant·es qui partent après leur projet d'étude. Il est parfois difficile de relier un projet, aux données qui ont été générées. » Un ancien doctorant explique : « Poste de travail de thèse mal sauvegardé par le service informatique à l'issu de mon départ. Données partiellement récupérées à mon retour en post-doc dans la même équipe. Environ 80% de pertes. » Un 3e indique : « Prise de fonction sans sauvegarde de mon prédécesseur car il n'y a aucun serveur commun. »

Enfin, le vol concerne 10 réponses (6%), et la perte d'un matériel (ordinateurs, papiers perdus dans un déménagement…), 8 réponses (5%).

3.3.1 Une vision expérimentale de la pérennité des données



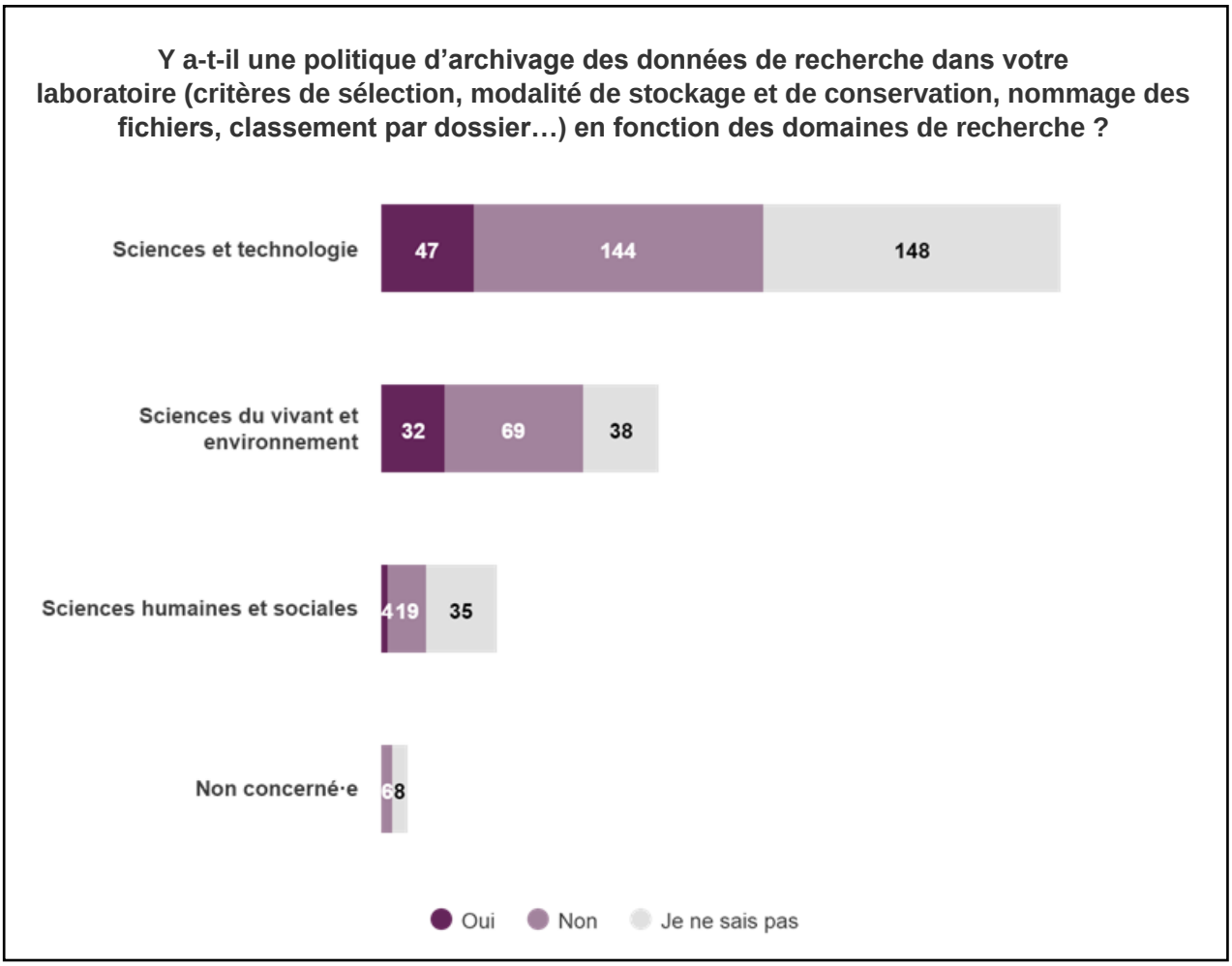


Aucune politique sur la durée de conservation des données ne semble se dégager de l'enquête. Les répondant·es gardent majoritairement indéfiniment leurs données, au-delà de la durée réglementaire. Ils n'évoquent cependant aucune raison principale : « pour ne pas les perdre » arrive en tête des réponses, couplée soit avec « pour mener de nouvelles études » (68 %), soit avec « pour justifier vos résultats de recherche » (55 %). L'intérêt patrimonial est moins critique (16 %).

On aurait donc tendance à penser que cette attitude est due à un principe de précaution sans analyse réelle de l'impact d'une politique de conservation. La tendance majoritaire est une conservation de longue durée.

Il n'y a pas de variations significatives selon les postes, sauf en ce qui concerne les personnels IST qui sont plus attachés à l'aspect patrimonial.

3.3.2 Archivage et classification des données de la recherche



La différence entre le stockage et l'archivage des données n'est pas clairement établie pour la majorité des répondant·es ; cette constatation est répétée à plusieurs reprises dans les entretiens :

> *« Mais voilà, c'est vrai que je n'ai pas encore réfléchi à ces problématiques d'archivage par rapport en tout cas à ma politique de stockage adoptée. C'est quelque chose que je n'ai pas encore envisagé. »* (chercheur en sciences humaines et sociales)

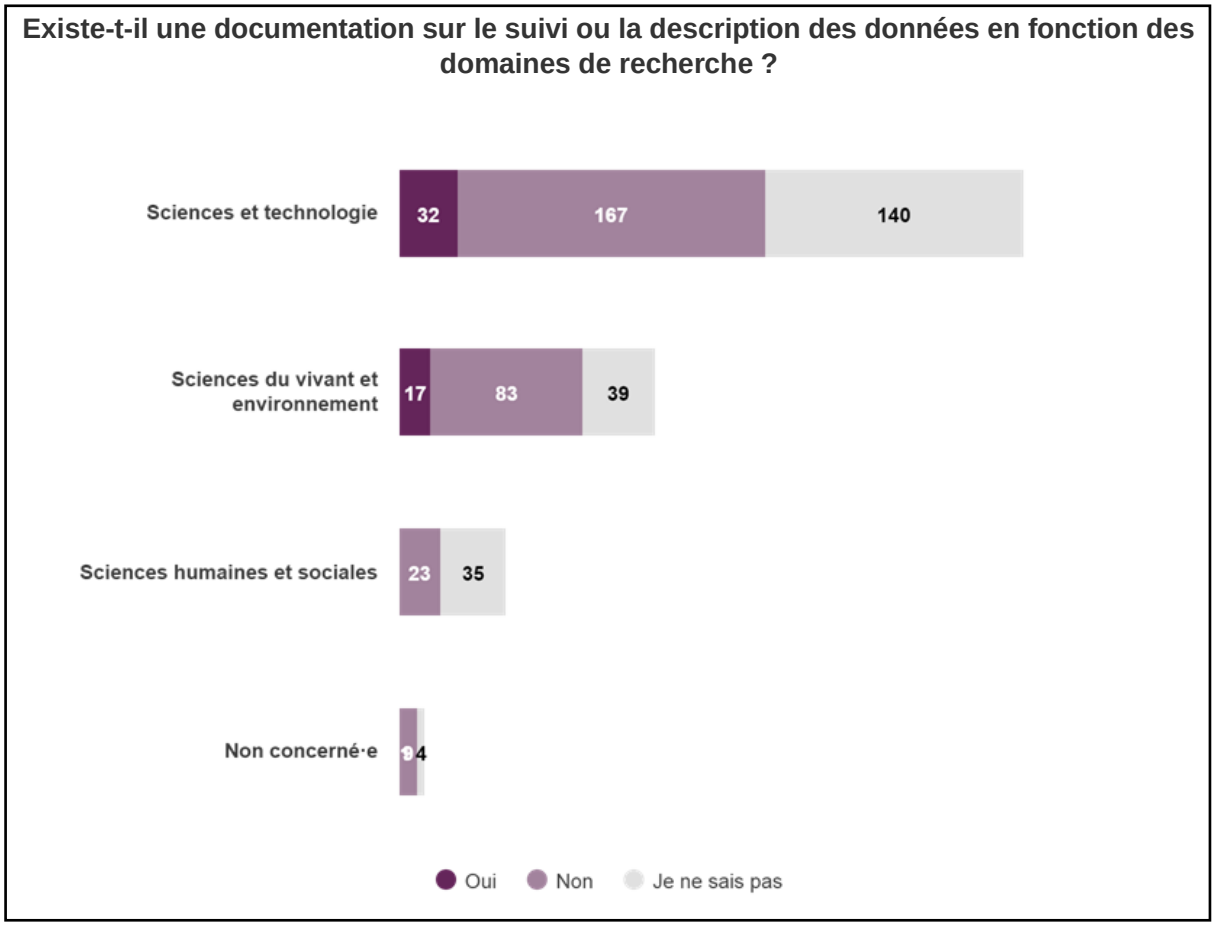
La demande récurrente est surtout qu'il existe une expertise au niveau du laboratoire, échelle privilégiée par les interlocuteurs des entretiens comme par les répondant·es de l'enquête.

> *« Moi je pense que ce qui serait le mieux c'est qu'il y ait un référent au niveau du laboratoire. Parce que voilà chaque laboratoire, chaque discipline a ses particularités aussi et il va y avoir des petites différences j'imagine entre certains labos. Donc, du coup qu'il y ait vraiment quelqu'un au laboratoire, c'est beaucoup*

plus en proximité, ça pourrait aussi aider plutôt qu'à l'échelle d'une composante, ou de toute l'université. » (chercheur en sciences et technologie)

« Non, il faudrait qu'une personne suffisamment compétente vienne dans le laboratoire, nous demande nos besoins et nous propose finalement une solution clé en main. Parce que c'est vrai qu'au quotidien, on trouve toujours autre chose à faire que traiter ce problème d'archivage des données, quoi. » (chercheur en sciences et technologie)

De manière significative, la majorité des répondant·es (71 % de oui) organisent leurs données de la recherche, c'est-à-dire qu'ils adoptent des pratiques d'archivage sans pour autant les apparenter au terme. La très grande majorité d'entre eux utilise une classification et une organisation personnelle (90 %), qui suit parfois une organisation collective mise en place au niveau du laboratoire (23 %). Dans ce cas, seul un haut niveau de connaissance des principes FAIR[2] par le répondant pourrait garantir la possibilité de réutiliser ces données ou de publier ces données.



[2] Définition des principes FAIR selon le site DoRANum : "La notion de FAIR data recouvre les manières de construire, stocker, présenter ou publier des données de manière à permettre que la donnée soit facile à trouver, accessible, interopérable et réutilisable."

Les répondant·es n'utilisent pas d'outil spécifique de suivi des données. Ils ont recours au tableur (47 %) ou au cahier de laboratoire (33 %) comme outil de suivi des données avec une forte proportion en sciences et technologie pour l'utilisation d'un logiciel spécialisé (78 %). L'utilisation de ces logiciels répond à des spécificités de données dans ces disciplines, mais introduit aussi une vulnérabilité technologique qui peut nuire à leur pérennité.

Bien que faisant partie des éléments à fournir dans la plupart des modèles de PGD, et incluse dans les principes FAIR, la question des standards de métadonnées apparaît être un angle mort de la gestion des données de recherche. Moins de 13 % des répondant·es ont renseigné la question « Suivez-vous un standard de description des données ? ». Sur ce faible pourcentage, seuls 4 répondant·es ont indiqué suivre un standard. La problématique de la description des données se retrouve dans les entretiens :

> *« Je sais qu'à chaque fois que j'ai eu affaire à des gens qui cherchaient à faire des bases de données ou autre, la butée qu'on a, c'est la sémantique. On ne parle pas forcément avec les mêmes mots de la même chose. Et derrière voilà, c'est soit on a affaire à un puriste qui dit "mais si on ne peut pas remplir toutes les cases, on ne pourra pas utiliser les données donc ce n'est pas la peine de le sauvegarder ».* (technicien de recherche en sciences et technologie)

> *« C'est un problème d'indexation car après c'est de la métadonnée. On va très vite vers des fichiers qui sont conséquents et un nombre de fichiers assez astronomique. Alors tout vouloir conserver, c'est bien. Après, est-ce que ce sera exploitable un jour et de quelle manière, j'en sais rien. »* (technicien de recherche en sciences et technologie)

De la même façon, seuls 9 % des répondant·es ont indiqué savoir qu'il existait une documentation sur le suivi ou la description des données dans leur laboratoire ou établissement, contre 51 % de non, et très significativement, 40 % de « je ne sais pas ».

Il ressort donc de l'enquête que l'archivage est un domaine encore largement sous exploré par les acteurs de la recherche ; néanmoins, les besoins sont clairement identifiés :

- Politique d'archivage commune au niveau du laboratoire
- Personne référente disposant de compétences d'archivage adaptées aux pratiques des disciplines de recherche
- Langage commun de description des données.

On constate que ces interrogations plus techniques sont souvent soulevées par des profils d'ingénieur·es ou de professionnels de l'IST, et que la dimension locale et spécifique des solutions d'expertise et de recours est privilégiée.

3.4. Partage et publication des données

3.4.1. Partage des données : une approche collective de la recherche ancrée dans les pratiques

Profil des répondant·es partageant des données

Le partage des données avec d'autres personnes est une pratique courante parmi les répondant·es : 72 % déclarent ainsi donner accès à leurs données, ce qui est pleinement cohérent avec la nature collaborative de la recherche scientifique.

Des disparités se manifestent néanmoins en fonction des champs disciplinaires : le partage est très répandu dans les disciplines expérimentales (83 % en sciences du vivant et de l'environnement, 74 % en sciences et technologies), mais semble moins pratiqué dans les sciences humaines et sociales d'après les répondant·es (43 %). Cela pourrait s'expliquer par les différences de méthodes scientifiques propres à chaque domaine, les sciences humaines et sociales abritant des disciplines où la recherche se pratique de manière plus individuelle que dans des disciplines appelant la mise en commun de plusieurs compétences (ingénieur·es et technicien·nes, etc.).

De même, on constate des disparités en fonction du poste occupé : les doctorant·es (dont 82 % sont en sciences expérimentales) semblent moins enclins à partager leurs données (54 %) que les personnels occupant des positions plus confirmées ; cela pourrait s'expliquer par une prudence accrue, ou un manque d'habitude, au cours de cette période de leur carrière, cohérente avec les résultats obtenus pour la publication des données (cf. point suivant). Par ailleurs, les personnels disposant d'un ETP dédié à la recherche plus élevé sont les plus nombreux à déclarer partager leurs données : 85 % pour les chercheur·es et ingénieur·es contre 64% pour les enseignants- chercheur·es.

On peut également relever que la part de répondant·es en situation de responsabilité est un peu plus importante à déclarer partager ses données (environ 80 %) que celle déclarant ne pas être concernée (68 %). En revanche, pour les personnes concernées, la nature de cette responsabilité (direction d'unité, de projet, etc.) ne semble pas beaucoup influer sur cette pratique.

Nature des données faisant l'objet d'un partage

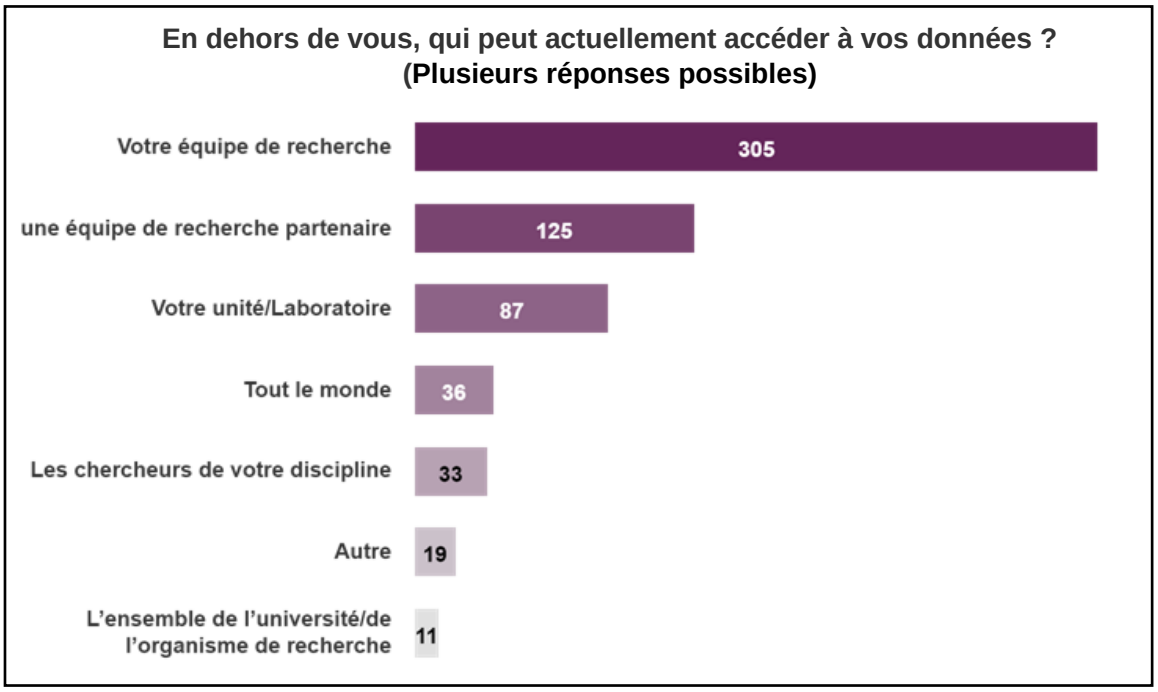
Le type de données ne semble pas avoir une influence significative sur leur propension à être plus ou moins partagées, mais la part des données textuelles et audiovisuelles faisant l'objet d'un partage semble légèrement moins importante (66 % et 68 %) que les autres types de données (plus proches des 80 %). Ce constat parait cohérent avec celui réalisé sur l'influence du domaine disciplinaire, les premières correspondant plutôt aux sciences humaines et sociales tandis que les autres types de données sont plutôt produites par les sciences expérimentales.

Il semble exister en revanche une corrélation entre données de gros volumes (> 1 To) et une part de partage plus élevée (85 %) que pour les données de volume moindre. On peut supposer que la nature de ces jeux de données va de pair avec l'implication de nombreux acteurs pour la gestion et l'analyse. Par ailleurs, ce sont les disciplines qui déclarent partager le plus leurs données qui sont aussi celles qui produisent les plus gros volumes d'après les répondant·es.

La présence de données personnelles dans les données traitées semble avoir une influence : les répondant·es déclarant ne pas partager leurs données sont nombreux (76 %) à déclarer avoir des données à caractère personnel, ce qui semble cohérent avec les précautions usuelles relatives à ces données.

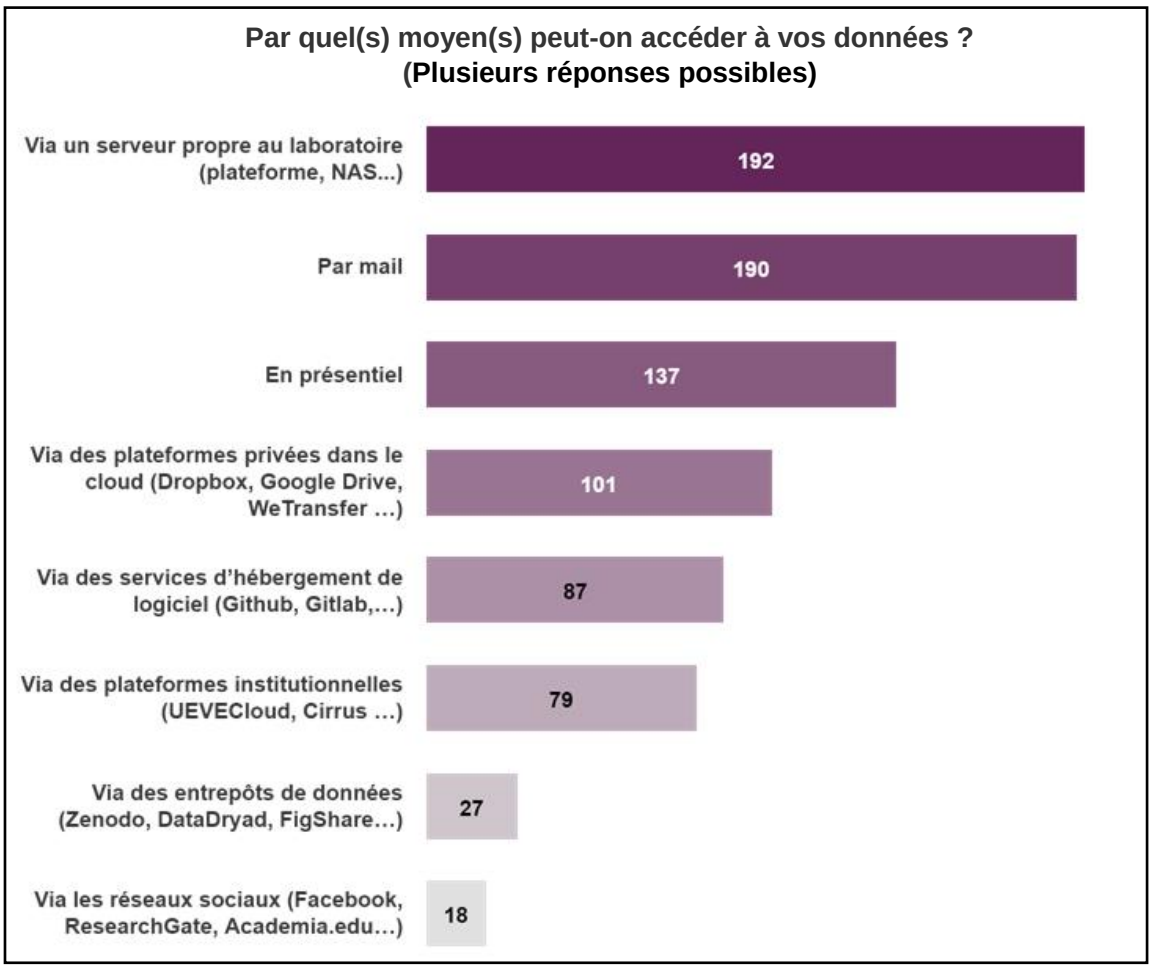
Les répondant·es indiquant avoir des données confidentielles liées à un secret industriel sont proportionnellement un peu plus nombreux à déclarer partager leurs données (78 %) que les autres répondant·es (70 %) : l'hypothèse peut être émise que ces données sont principalement produites dans le cadre de partenariats avec des acteurs privés, nécessitant le partage de données entre partenaires.

Profil des personnes qui accèdent aux données



Le partage des données se réalise surtout au niveau des équipes de recherche (82 % des répondant·es) ; l'ouverture extérieure voire à tous reste minoritaire (34 % pour une équipe de recherche partenaire, moins de 10 % pour un cercle plus large) – à noter toutefois que cette dernière catégorie relève davantage de la publication, approche explorée plus en détail dans la partie suivante. La proportion relativement faible de répondant·es déclarant partager leurs données avec leur unité de recherche (23 %) par

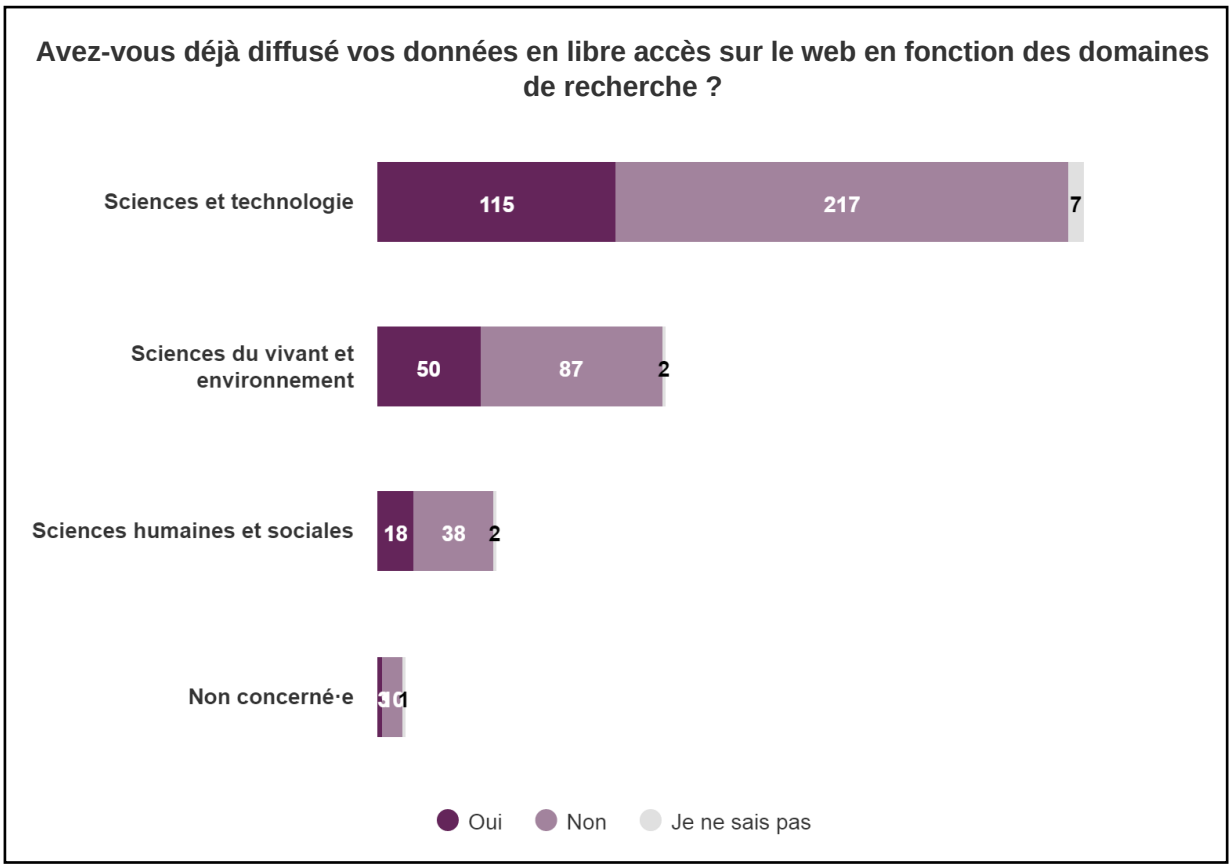
rapport au grain « équipe de recherche » interroge sur les dynamiques de consolidation des données à l'échelle unité – à moins que le terme « équipe » n'ait été compris de différentes façons.



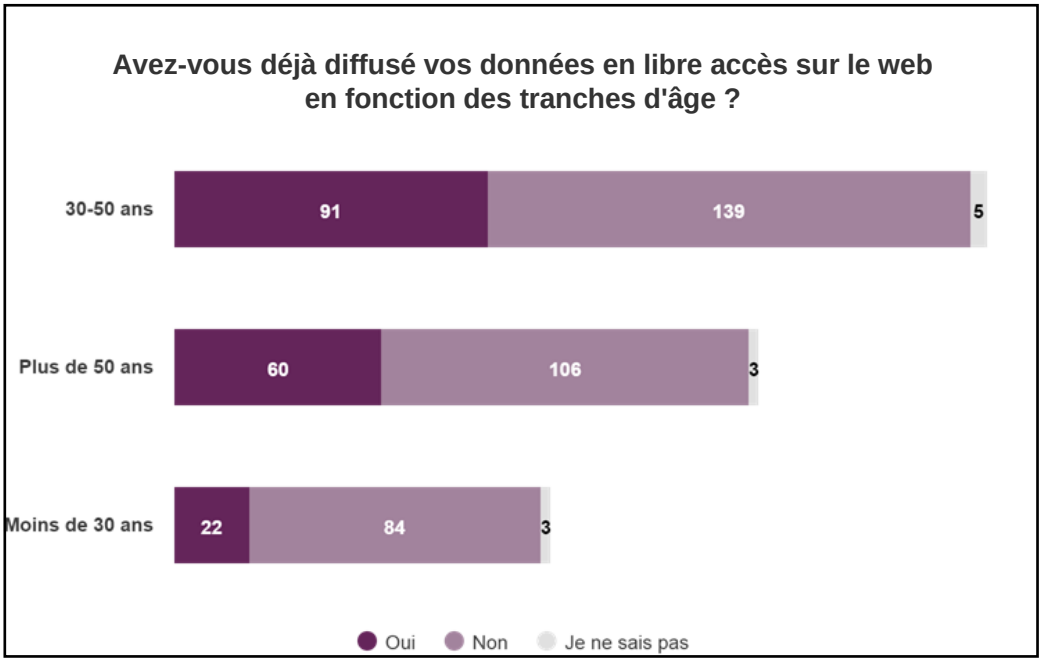
L'e-mail (51 %) et les serveurs propres à l'unité (52 %) sont les moyens de partage les plus courants. Les entrepôts sont encore peu utilisés par les répondant·es (7 %). Une part significative (33 %) déclare partager leurs données « en présentiel » (poste dédié pour données sensibles, cahier de laboratoire, etc.). Un point d'attention réside sur les plateformes privées dans le cloud, que les répondant·es déclarent utiliser un peu plus (27 %) que les plateformes institutionnelles (21 %). On peut se demander comment les répondant·es ayant évoqué les réseaux sociaux (5 %) utilisent concrètement ces outils au regard des données : communication d'un lien extérieur ou partage effectif de fichiers via ces outils ?

3.4.2. Publication des données : le paysage de l'Université Paris-Saclay

Quelles pratiques ?



Un tiers des répondant·es (34 %) déclare avoir déjà publié ses données sur le web. Les pratiques ne diffèrent que très légèrement entre les domaines disciplinaires : les données sont un peu plus publiées en sciences du vivant et de l'environnement (36 % des répondant·es) qu'en sciences et technologie (34 %) et en sciences humaines et sociales (31 %).



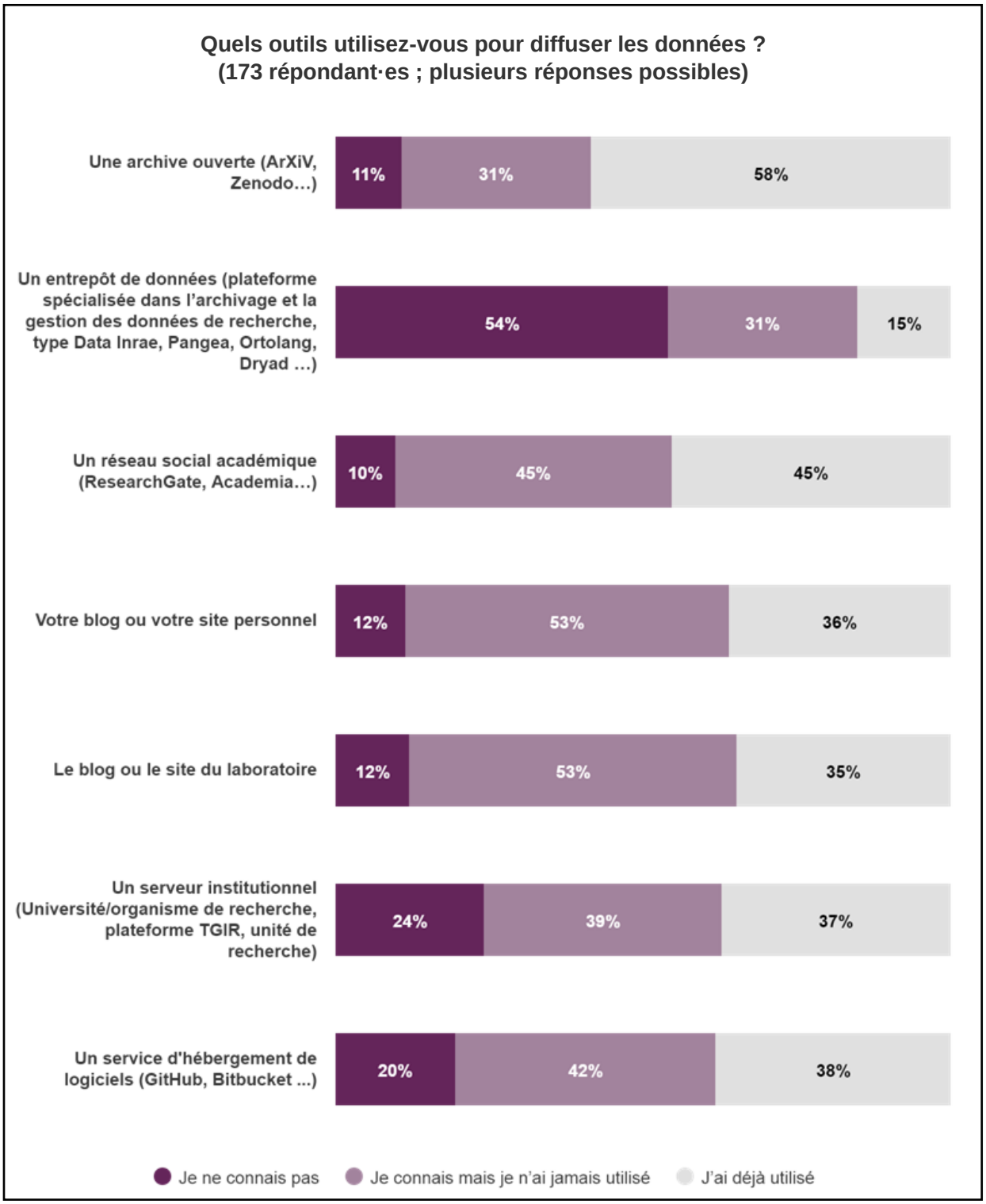
La distinction dans les pratiques de publication de données repose sur l'âge : les répondant·es de moins de 30 ans publient sensiblement moins (20 %) que les répondant·es de plus de 50 ans (35 %) et ceux de 30 à 50 ans (39 %).

Cela peut être imputé à la durée de la carrière et des opportunités de dépôts et de publication des données : les plus jeunes répondant·es ayant eu moins d'occasions de publier autant que ceux âgés de 30 à 50 ans, et les plus âgés n'ayant pas pu bénéficier des systèmes et des infrastructures permettant la publication dès le début de leur carrière. Les usages de début de carrière peuvent aussi expliquer une publication plus faible : le manque de connaissance sur les outils, la volonté de garder le contrôle sur ses données, etc. Les chiffres sur les pratiques spécifiques des doctorant·es en la matière corroborent cette observation.

Parmi les répondant·es qui ont déclaré n'avoir pas publié de données, la moitié pense néanmoins que leurs données sont publiables, en partie ou en totalité. Par contre, il ne semble pas que la nature sensible ou personnelle ou à caractère industriel des données ait une influence sur leur publication ou non.

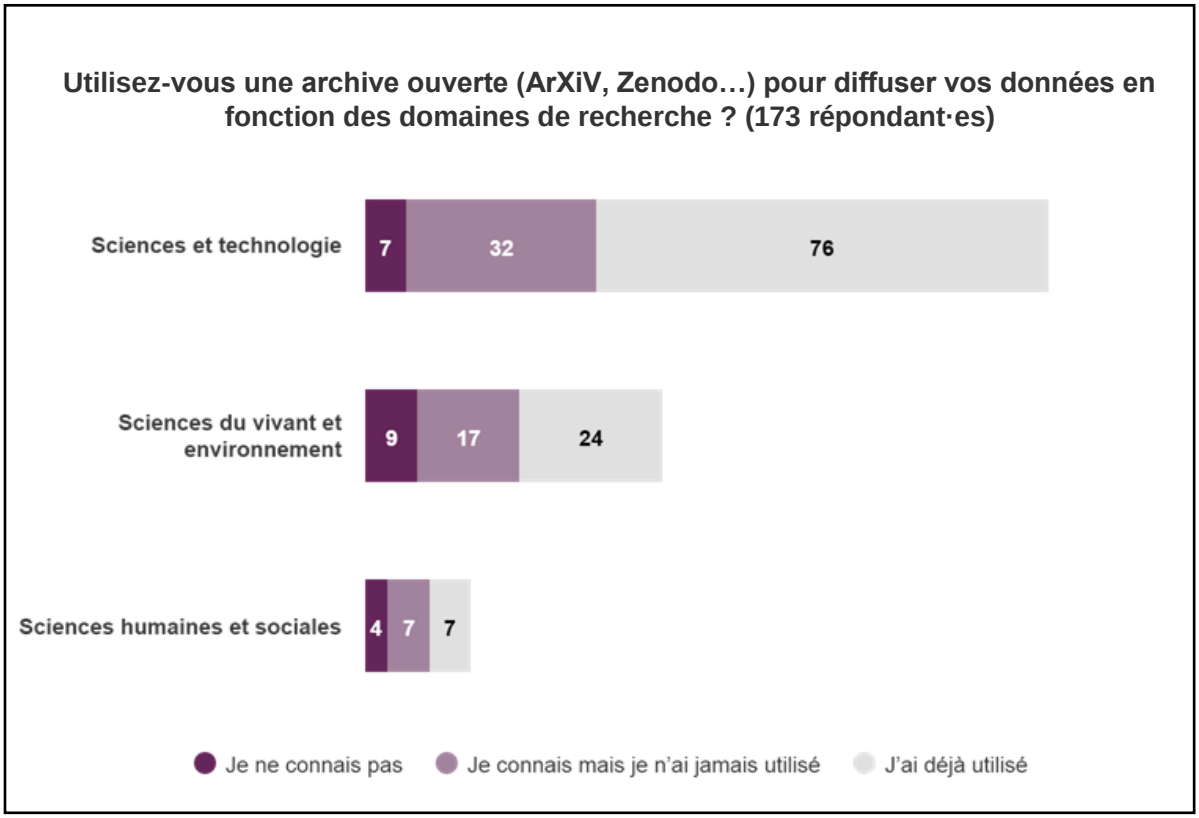
C'est le cas dans tous les domaines de recherche : un peu plus en sciences et technologie (54 %) et en sciences du vivant et de l'environnement (42 %) qu'en sciences humaines et sociales (39 %).

Quels outils pour diffuser les données ?



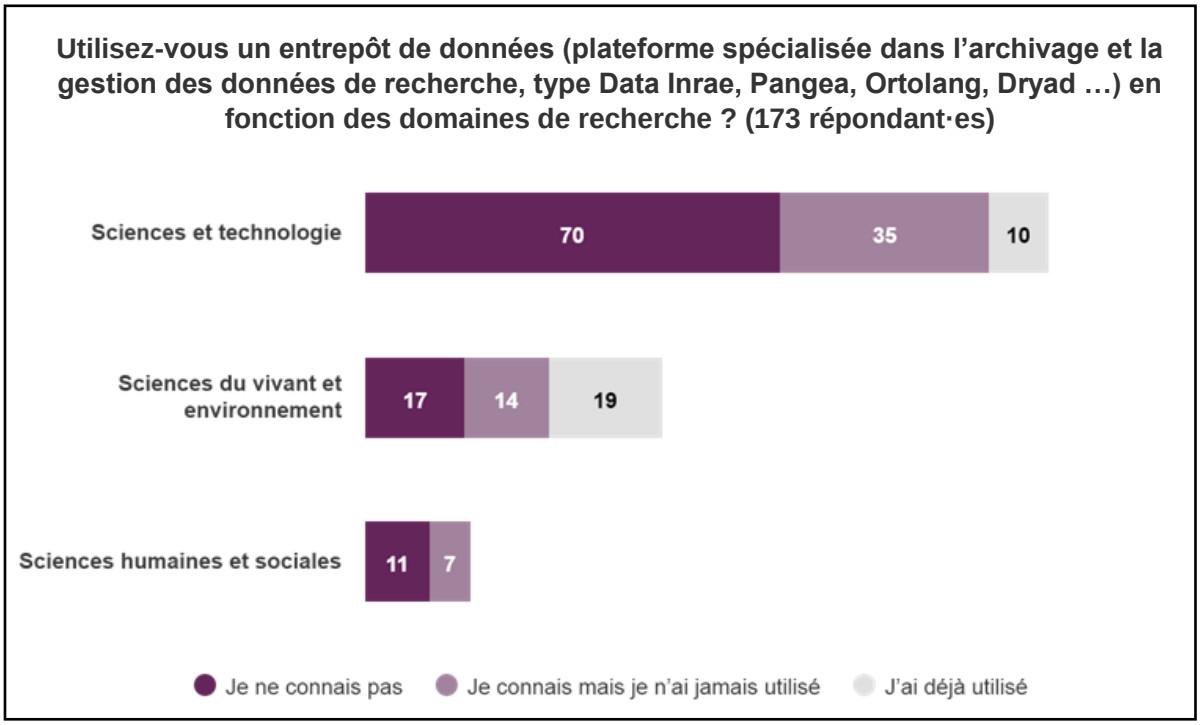
173 répondant·es à l'enquête ont affirmé publier leurs données. Leur connaissance des nombreux outils pour publier des données en ligne et l'utilisation qu'ils en font se révèlent très hétérogènes.

Globalement, les différents canaux de publication mentionnés par l'enquête (archive ouverte, blog ou site personnel ou institutionnel, entrepôt, réseau social…), s'ils ne sont pas toujours utilisés, sont connus par plus de 80 % des répondant·es, à l'exception notable des entrepôts de données (voir plus bas). Les outils les plus utilisés sont les archives ouvertes (58 % des personnes qui publient leurs données) et les réseaux sociaux académiques (45 %) ; les moins utilisés, en plus d'être les moins connus, sont les entrepôts de données. Enfin, il est à noter que quelques réponses libres mentionnent la diffusion des données dans les publications, en annexe des articles scientifiques.



L'utilisation de ces différents canaux n'est pas liée à la nature des données produites, mais semble liée à des usages disciplinaires. Les distinctions d'usage selon les domaines de recherche sont notables en particulier pour les archives ouvertes et les entrepôts de données. Les archives ouvertes sont très largement connues et utilisées en particulier dans le domaine des sciences et technologie (respectivement 94 % et 66 %) et dans celui des sciences du vivant et de l'environnement (82 % et 48 %).

Focus sur les entrepôts de données

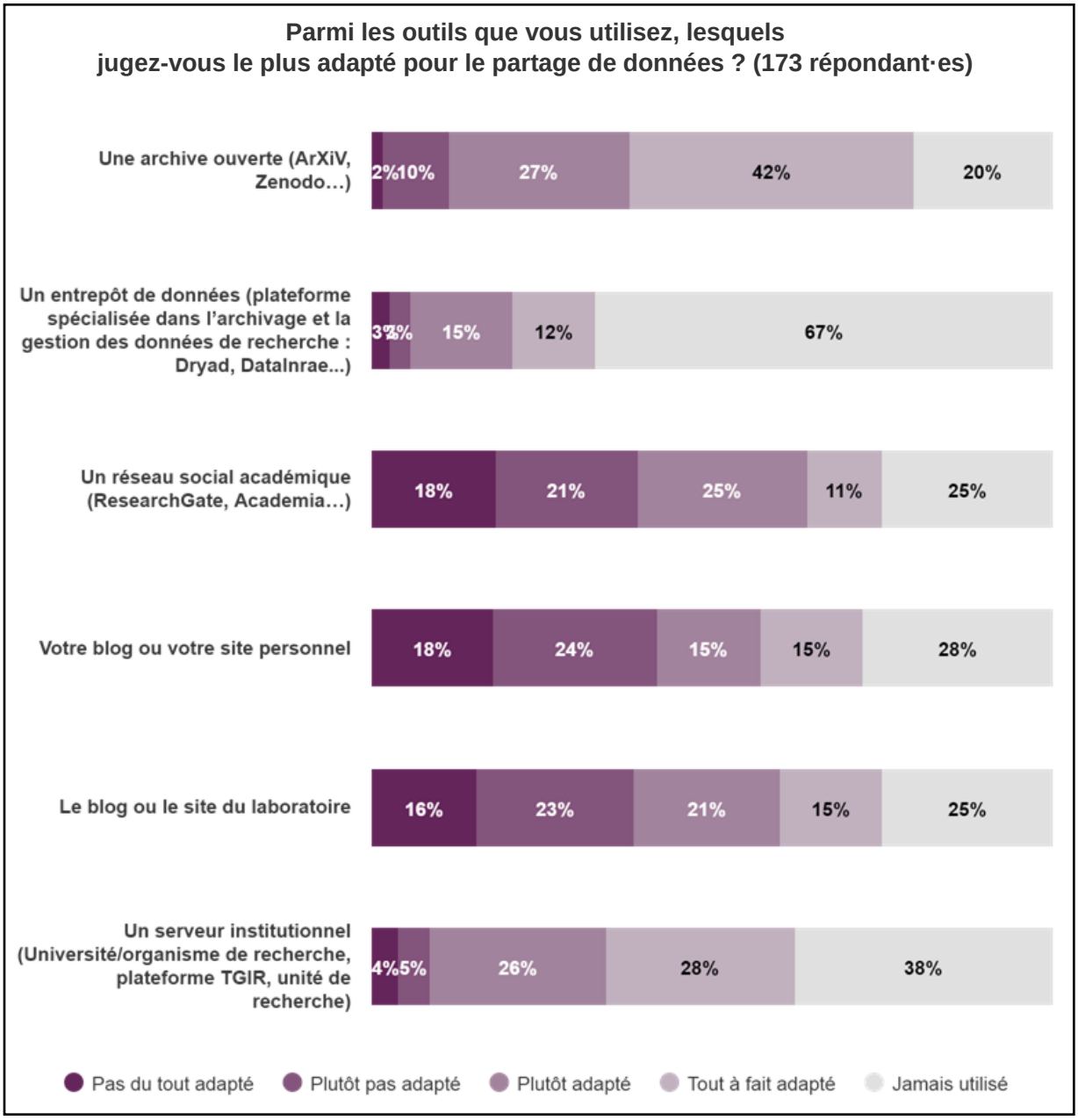


L'outil le moins connu, et par conséquent le moins utilisé, est celui qui est spécifiquement prévu pour la publication de données en ligne. Plusieurs éléments peuvent expliquer ce phénomène à première vue étonnant : le manque de connaissance de l'existence d'un tel service ou sa mise à disposition par un établissement particulier peuvent faire que les chercheur·es ne savent ou ne peuvent pas l'utiliser.

La spécialisation disciplinaire des entrepôts constitue aussi une explication. On note en particulier qu'aucun des répondant·es en sciences humaines et sociales n'a déclaré avoir utilisé un entrepôt de données (et 33 % seulement connaissent ce type de service), contrairement aux autres domaines de recherche : les répondant·es en sciences du vivant et de l'environnement sont 38 % à en avoir déjà utilisé et 28 % à les connaître.

Quels outils adaptés à la publication des données ?

Au-delà des usages, la question de savoir si ces outils sont les plus adaptés à la publication des données se pose.



L'usage des outils institutionnels dédiés au stockage et à la mise à disposition de données est plébiscité par les répondant·es qui ont déjà publié leurs données : les archives ouvertes, les serveurs institutionnels et, dans une moindre mesure parce que globalement mal connus, les entrepôts de données. Les blogs et sites personnels ou de laboratoires, qui ne sont pas faits en premier lieu pour recueillir et publier des données, et les réseaux sociaux académiques, même s'ils sont utilisés largement par les répondant·es à l'enquête, sont moins considérés par les répondant·es comme des outils adaptés.

Si on compare l'utilisation qui est faite de ces outils et ce qu'en pensent les répondant·es, il ressort clairement qu'il manque aux chercheur·es des canaux de diffusion adaptés pour les données de la recherche. Plus précisément, ceux qui sont les plus utilisés, comme les réseaux sociaux académiques ou les sites et blogs personnels, ne sont pas considérés comme les plus adaptés. Et ceux qui sont considérés comme les plus adaptés par les éventuels utilisateurs eux-mêmes, comme les entrepôts de recherche ou les serveurs institutionnels, sont les plus mal connus et les moins utilisés : soit ces services n'existent pas, soit ils ne sont pas assez connus des chercheur·es.

3.4.3. Réutilisation des données

Le téléchargement de données en ligne via un entrepôt n'est pas si marginal parmi les répondant·es, puisque la moitié déclare avoir déjà récupéré les données d'autres chercheur·es via un entrepôt de données en ligne ; en revanche, seuls 15 % déclarent le pratiquer sur une base régulière. Compte tenu de la thématique de l'enquête, on peut néanmoins se demander si cette information est représentative des pratiques de nos communautés, ou si nous bénéficions de réponses d'une partie des chercheur·es déjà sensibilisés aux données de la recherche. Compte tenu des retours réalisés sur la connaissance des entrepôts, notamment en entretiens, on peut également se demander dans quelle mesure la notion d'entrepôt a été comprise ici.

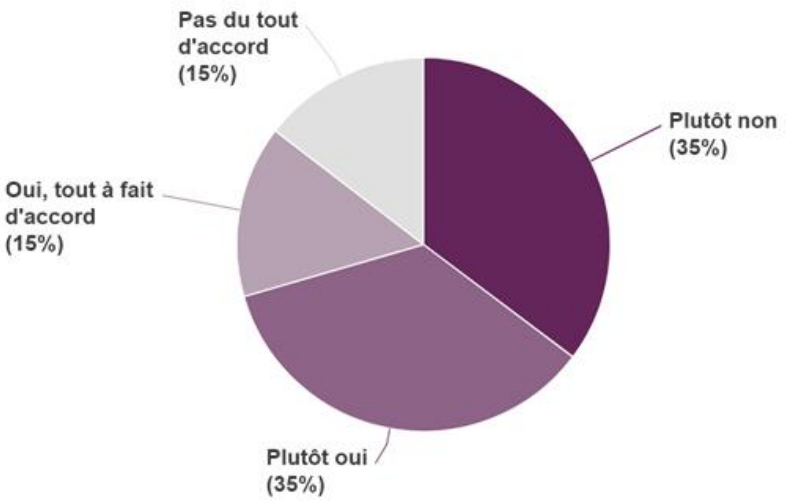
La discipline ne semble pas jouer de rôle majeur dans la pratique du téléchargement : les répondant·es en sciences et technologies et sciences du vivant (environ 50 %) sont un peu plus nombreux à déclarer avoir déjà téléchargé des données, mais la différence avec les sciences humaines et sociales (environ 45 %) reste faible.

La principale raison déclarée du non téléchargement de données est le manque d'opportunité ou de connaissance de la possibilité (94 %), ce qui souligne l'absence d'une opposition de principe marquée à une telle pratique ; la réutilisation de données publiées ne fait simplement pas encore partie des habitudes des répondant·es dans la construction des protocoles de recherche.

Le téléchargement de données est principalement utilisé à des fins de réutilisation dans ses propres recherches (89 %) ; la vérification des résultats apparaît comme plus minoritaire (37 %) pour les répondant·es.

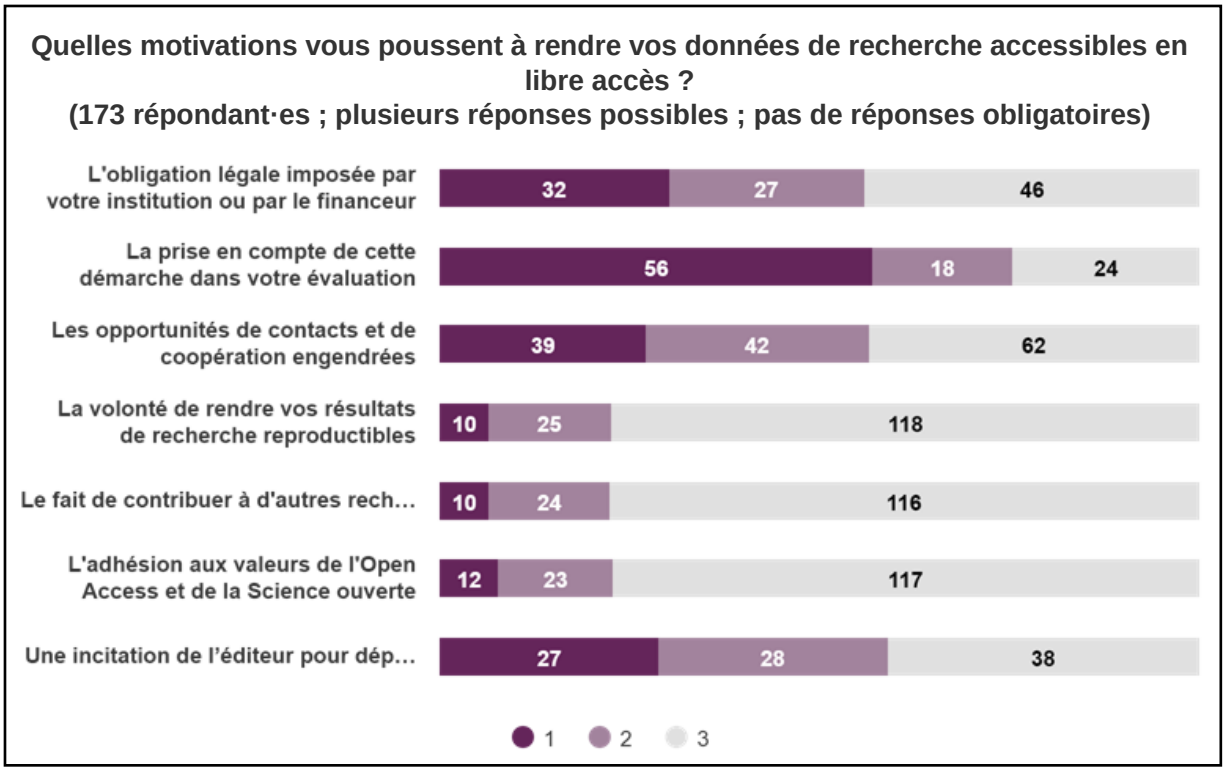
4. Motivations et freins à la publication des données

Pensez-vous que vos données de recherche puissent être diffusées tout ou partie en libre accès sur le web ? (329 répondant·es)



La moitié des répondant·es est d'accord pour publier tout ou partie de leurs données de recherche en libre accès sur le web.

4.1. Accessibilité des données : des motivations centrées autour des pratiques de la recherche et de la Science Ouverte



Les répondant·es ont été invités à évaluer entre 1 (aucun impact) et 3 (fort impact) les motivations qui les encouragent à publier leurs données en libre accès. Trois raisons se détachent nettement :

- Le fait de contribuer au développement d'autres recherches en permettant l'exploitation des jeux de données (77 %)
- La volonté de rendre les résultats de recherche reproductibles (77 %)
- L'adhésion aux valeurs de l'Accès Ouvert et de la Science Ouverte (77 %)

Suivent :

- L'obligation légale imposée par l'institution ou par le financeur (44 %)
- Les opportunités de contacts et de coopération engendrées (43 %)
- L'incitation de l'éditeur pour déposer en libre accès les données associées à une publication (41 %)
- La prise en compte de cette démarche dans l'évaluation de l'activité scientifique (25 %)

D'autres motivations ajoutées par les répondant·es dans la rubrique « Autre » se rapportent à la nécessité de transparence, accessibilité et visibilité de la recherche, par exemple : « rendre mes travaux plus visibles et accessibles » (enseignant/enseignante –

chercheur/chercheuse, 30-50 ans) ou « la recherche doit être en libre accès par définition » (chercheur/chercheuse, moins de 30 ans). Un des répondants souligne que les données de la recherche produites sur des fonds publics doivent être rendues accessibles au public. Il note : « Je pense que mon envie de partager les données va plus loin que l'adhésion à la Science ouverte. Je ne vois même pas comment on peut trouver normal que de l'argent public serve à produire des données inaccessibles par le public » (chercheur/chercheuse, 30-50 ans). Une autre personne met l'accent sur la demande des revues de la discipline de mettre à disposition des données lors de la publication : « Cette démarche d'archivage et de mise à disposition des données transférables est rendue obligatoire pour publication dans les meilleures revues de la discipline [...] Ces revues prennent en charge l'archivage » (enseignant/enseignante-chercheur/chercheuse, 30-50 ans).

Il est intéressant de souligner que les réponses diffèrent légèrement selon les postes et les responsabilités lorsque les répondant·es ont à analyser les raisons qui les poussent à publier leurs données ; la différence significative se trouve dans l'ordre et l'importance donnée au trio de tête noté plus haut.

Si on examine ces raisons en fonction du poste occupé au sein de l'Université Paris-Saclay, on observe que les chercheur·es citent surtout l'obligation légale imposée par l'institution ou le financeur, le fait de contribuer à d'autres recherches et la volonté de rendre les résultats reproductibles. Pour les enseignants-chercheur·es, tout comme pour les doctorant·es, les ingénieur·es, les personnels IST et bibliothèques, l'adhésion aux valeurs de l'Open Access et de la Science Ouverte se situe en première place, suivie par l'importance de participer à d'autres recherches et la vue des résultats reproductibles et réutilisables. C'est la seule différence qui s'entrevoit.

L'analyse en fonction d'autres responsabilités exercées au sein de l'Université Paris-Saclay révèle que l'obligation légale de la part de l'institution ou financeur est le premier choix des directeurs et directrices de thèse. Les personnes qui ont des responsabilités de direction ou chargé de projet déclarent comme raisons principales l'adhésion aux valeurs de l'Open Access, le fait de contribuer à d'autres recherches, l'obligation légale de l'institution ou de financeur. Pour les directeurs et directrices de laboratoire, c'est la volonté de contribuer à d'autres recherches qui compte le plus.

L'adhésion aux valeurs de l'Open Access se situe en première position dans les Sciences et Technologie et dans les Sciences Humaines et Sociales. Le fait de contribuer à d'autres recherches prévaut dans les Sciences du vivant et environnement.

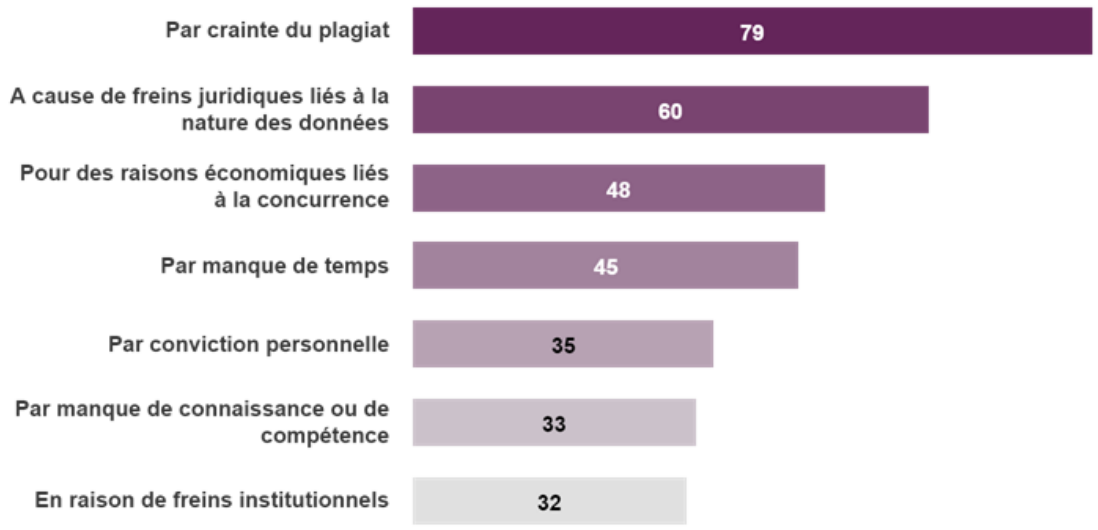
Il n'est pas observé de différences substantielles en fonction du domaine de la recherche, de la nature des données, ni en fonction de la connaissance des PGD.

4.2. Des freins importants : la contradiction inhérente à la publication des données

Les entretiens font apparaître des opinions plus mesurées. Personne ne semble être foncièrement contre la publication des données, mais beaucoup d'interrogations subsistent, notamment sur l'utilité d'ouvrir des données brutes. La lisibilité de ces données est questionnée, nombreux sont ceux qui n'y voient pas d'intérêt, qui ne souhaitent pas y consacrer plus de temps ou, même, qui se méfient de cet aspect. Les valeurs d'accessibilité et de publication sont comprises, mais ne semblent pas suffire face aux problématiques compétitives dans certaines disciplines. La publication et la « rentabilisation » des données de la recherche sont pour certains deux injonctions qui se contredisent et qui freinent les initiatives sur la question. Pour celles et ceux qui par contre adhèrent aux discours de publication, les problématiques des logiciels et codes sources, du développement du libre et de la science participative sont également des voies à explorer.

> *« Vu la compétition qu'on a en tant que chercheur, on n'est pas du tout sur ce genre de paradigme. On est sur vraiment de la privatisation de données que j'ai pu vivre moi avec carrément une volonté, dans certains sujets, d'empêcher l'accès à ces données alors qu'elles sont publiques ou de rendre difficile l'accès. »*
> *(doctorant en sciences humaines et sociales)*

Pour quelle(s) raison(s) ne souhaitez-vous pas diffuser en libre accès vos données de recherche ? (164 répondant·es ; plusieurs réponses possibles)



S'agissant de ce qui limiterait les répondant·es dans la publication de leurs données, ils citent en premier lieu la crainte du plagiat ; viennent ensuite les freins juridiques, les raisons économiques et le manque de temps.

La crainte du plagiat est le premier choix de toutes les catégories d'acteurs de la recherche, tant des directeurs ou directrices de thèse ou de laboratoires, que de doctorant·es ou d'enseignants-chercheur·es.

Lors des entretiens, un chercheur pointe l'intérêt perçu comme anecdotique de disposer des données brutes. Et un deuxième chercheur avance le fait que c'est l'interprétation de ces données qui aurait principalement du sens.

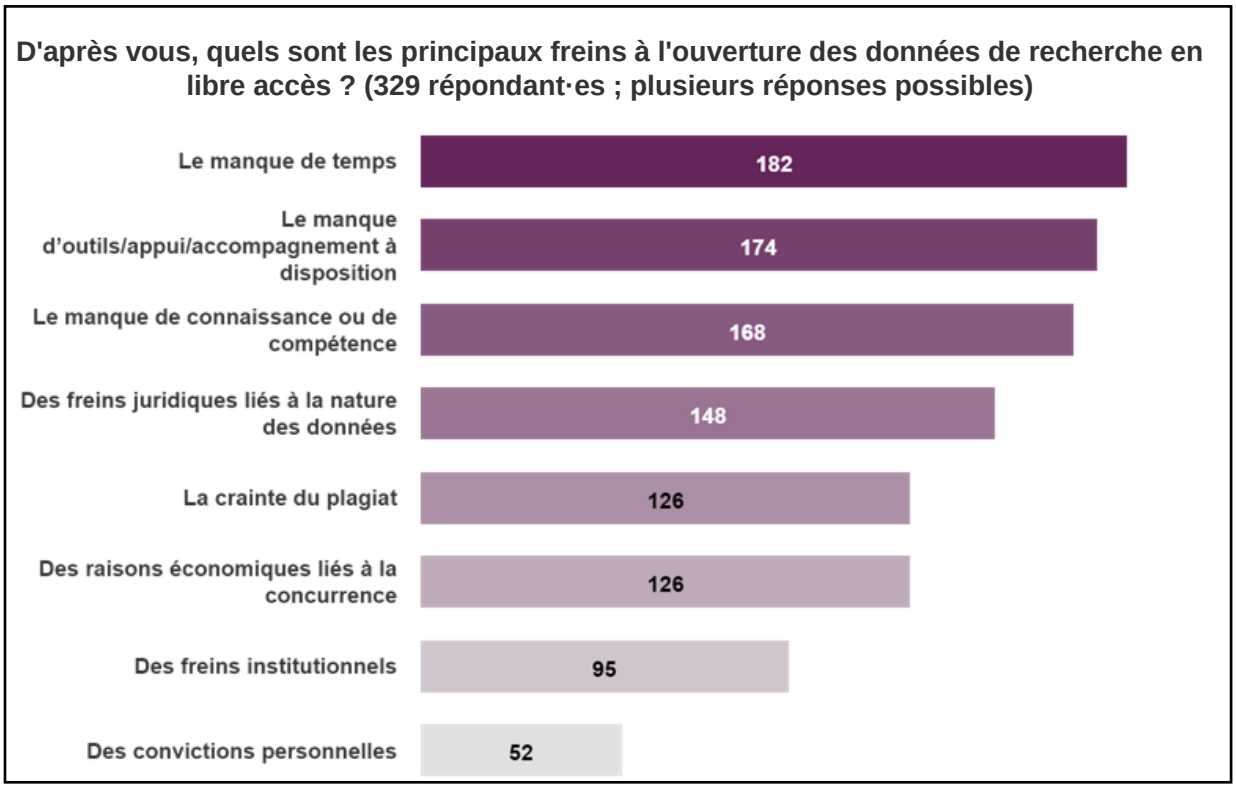
> *« Comment les données qu'on a recueillies pour des expériences tellement spécifiques, même si on les partageait, est-ce qu'elles pourraient vraiment être comprises par d'autres ? […] Comment trouver un langage commun, finalement, qui soit compréhensible à un plus grand nombre ? »* (chercheur en science du vivant et environnement)

> *« Quand vous voulez construire un modèle interprétatif, parce que c'est ça qui a du sens : si c'est pour donner des données au lecteur, il ne va pas lire. Donc les interprétations sont un travail excessivement difficile et je pense que c'est notre boulot. »* (chercheur en sciences humaines et sociales)

Cette opinion est largement partagée par les répondant·es de l'enquête. Le manque de temps pour organiser et structurer les données afin de les rendre compréhensibles aux non-initiés constitue un obstacle important pour leur publication. Un des répondant·es note : « Elles sont trop 'brutes', pas mises en forme. Je sais « décoder » mes propres données au sens de : retrouver quoi est quoi (p. ex. dans ce fichier le temps est en première colonne, la température, en Kelvin en deuxième, la pression en Pascal pour la troisième colonne, etc.), et ce avec une organisation parfois fluctuante, mais dont je sais retrouver les étapes grâce à des notes manuscrites écrites dans une sorte de 'cahier de manips'. Ce serait un très gros travail que de les uniformiser et de les rendre décodables par des non-initiés. Cela m'est pourtant arrivé d'en partager avec des collègues bien identifiés » (chercheur/chercheuse, plus de 50 ans).

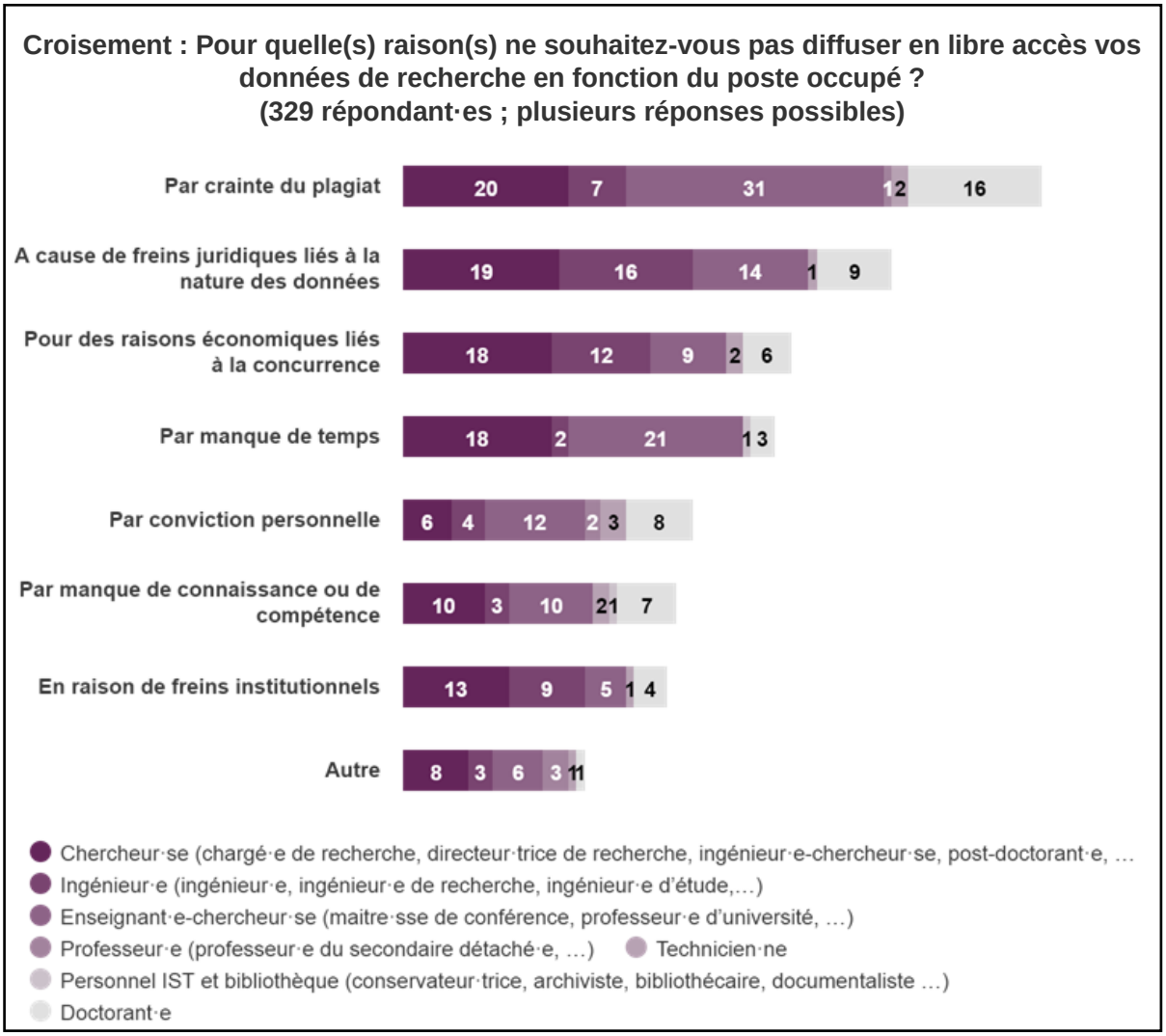
Certains enquêtés ne voient pas, tout simplement, l'intérêt et l'utilité de donner accès libre à leurs données de recherche, surtout aux données brutes : « Généralement, on donne accès sur demande aux données brutes après leur publication ou au moment de leur publication. Mais je ne vois pas l'intérêt de laisser libre accès aux données non publiées » (chercheur/chercheuse, plus de 50 ans) ; « Je pense mettre suffisamment de descriptions et données dans mes publications, l'accès à mes données brutes ne me paraît pas utile » (chercheur/chercheuse, plus de 50 ans). Ces opinions sont plus récurrentes dans la catégorie des plus de 50 ans.

Il est intéressant de voir que les répondant·es n'ont pas répondu exactement de la même façon à la question leur demandant d'identifier les freins à la publication des données (et non pas leurs propres raisons pour ne pas le faire).



Arrivent en tête les freins suivants :

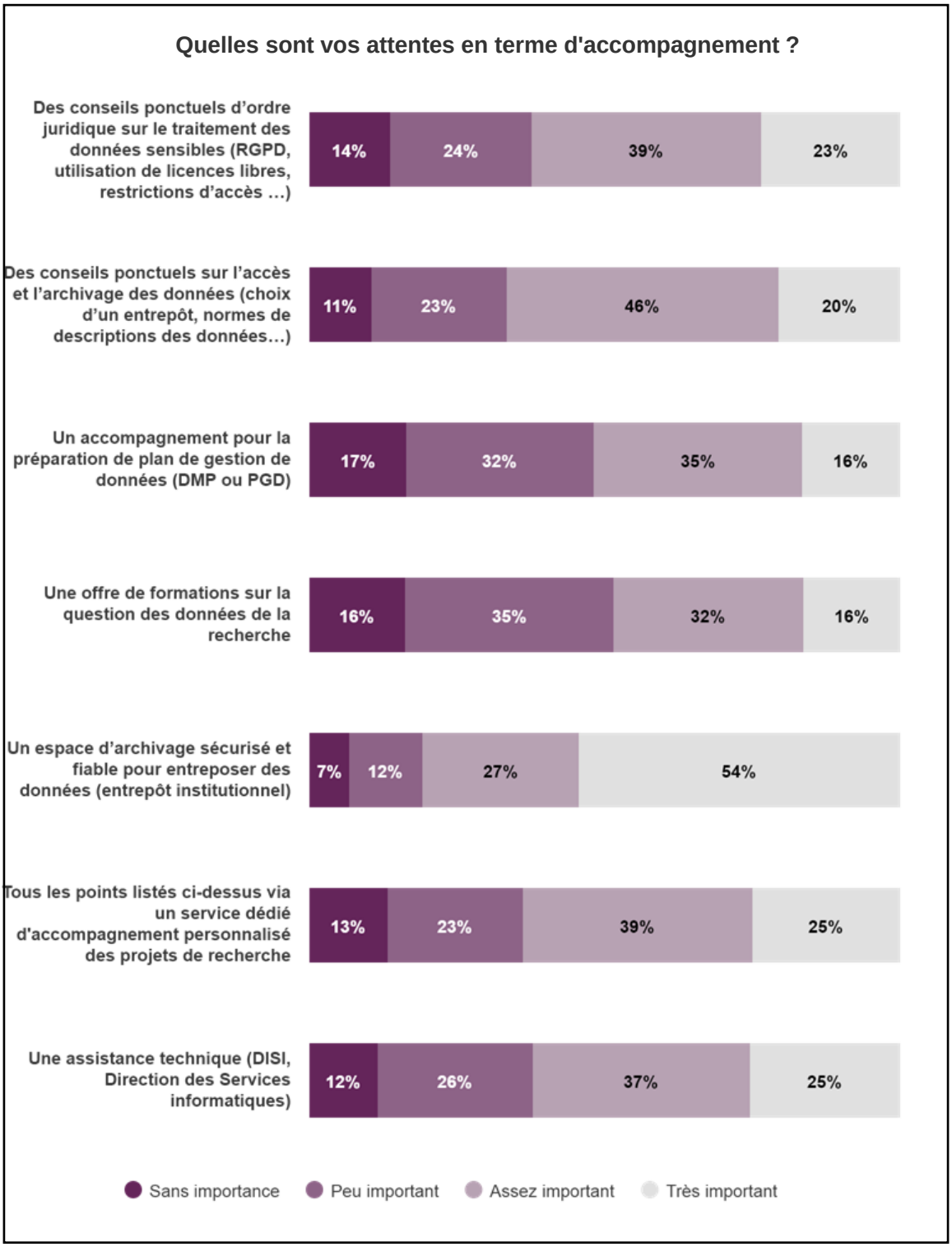
- Le manque de temps (54 %)
- Le manque d'outils/appui/accompagnement à disposition (51 %)
- Le manque de connaissance ou de compétence (50 %).



La concurrence et la compétition universitaire sont évoquées dans les commentaires, ainsi que la méfiance envers les usages des données mises en accès libre : « pas la crainte du plagiat, mais la crainte de voir d'autres aller plus vite que soi dans l'exploitation des même données » (chercheur/chercheuse, plus de 50 ans) ; « se faire doubler à la publication, parfois sans citation des auteurs de la donnée » (doctorant/doctorante, moins de 30 ans). D'autres freins sont relevés, comme le coût des publications en libre-accès si l'on choisit un modèle auteur-payeur.

Ainsi, un état contradictoire se dégage de l'enquête : les répondant·es dépeignent un climat globalement favorable à la publication des données au nom des usages de la recherche (reproductibilité des résultats, diffusion des résultats de la recherche, science ouverte), mais se montreraient plutôt réticents dans la pratique. Si le manque de temps et d'accompagnement sont cités dans les raisons objectives, les impressions personnelles font plutôt apparaître une crainte marquée du plagiat, de la compétitivité de la recherche et la réticence face au partage des données brutes.

5. Besoins exprimés par les répondant.es



On relève des attentes globales sur la formation, la mise à disposition de supports, la mise en place d'une plateforme mutualisée. Il semble y avoir un besoin d'avoir un interlocuteur,

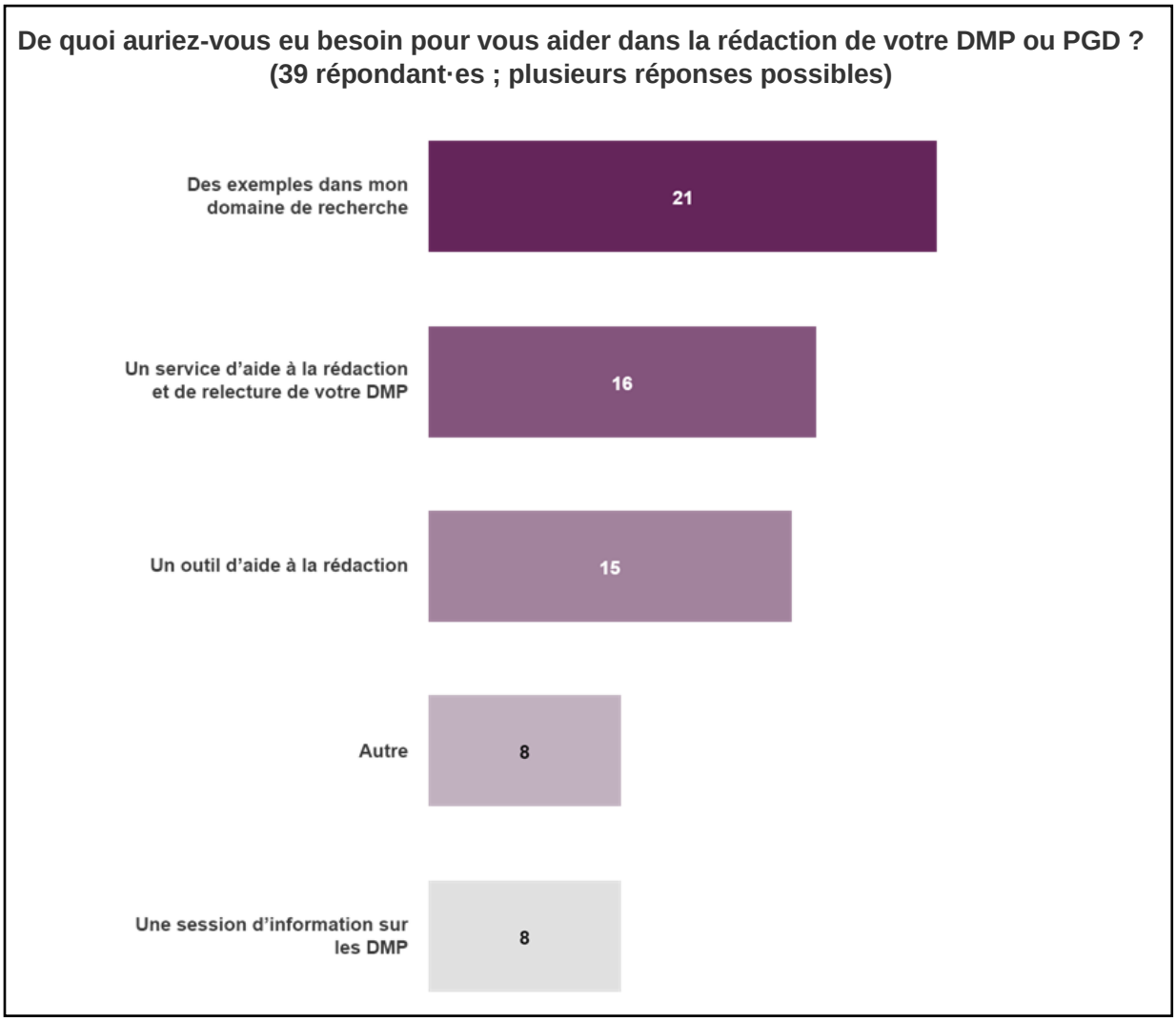
au plus près de la discipline, de l'unité de recherche, pour prendre en charge la totalité des dossiers sur cette question. Les personnes interviewées tout comme les répondant·es identifient également un besoin de formation, pour elles, les doctorant·es, les étudiant·es de master, même s'il n'y a pas vraiment de consensus sur la forme que devrait prendre un éventuel dispositif de formation. Le constat est le même pour la question d'une plateforme : on ressent un besoin sur le stockage, l'archivage, les métadonnées, et toutes les problématiques que l'on pourrait y agréger, mais aucune forme réelle ne se dessine. Un des interviewés résume en entretien :

> *« Pour moi, le premier point, ce serait la formation, la formation des doctorants, la formation des chercheurs. Que faire pour que ce soit une préoccupation au laboratoire de comment on gère des données ? Pour que ça devienne un sujet, déjà. Ensuite sur les dispositifs, s'il y avait un entrepôt de données mieux identifié par le laboratoire où on pourrait déposer nos données. […] Et puis oui, après, je pense qu'il manque des exemples, des expériences sur lesquelles on puisse s'appuyer. »* (chercheur en sciences et technologie)

La moitié des répondant·es reconnaît des difficultés dans la gestion des données, liées principalement au manque de temps et au manque de connaissances sur le sujet. Le lien entre manque de connaissances et besoin en formation n'est pas fait systématiquement, puisque seuls 35 % des répondant·es disent manquer d'accompagnement et de formation sur le sujet. L'absence de ressources matérielles est une difficulté pour 29 % des répondant·es, bien que soit majoritairement exprimé par ailleurs le besoin d'une infrastructure d'archivage des données (voir plus bas).

Il n'y a pas de désintérêt exprimé pour la gestion des données de la recherche puisque seuls 16 % disent éprouver un manque d'intérêt pour la question, à rebours de l'idée que cette gestion serait perçue comme une contrainte éloignée des préoccupations quotidiennes des chercheur·es.

5.1. Autour du Plan de Gestion des Données



En ce qui concerne les besoins en matière de rédaction d'un Plan de Gestion des Données, 54 % de l'ensemble des répondant·es indiquent qu'ils ont besoin d'exemples dans leur domaine de recherche. 41 % souhaitent bénéficier d'un service à la rédaction et relecture de leur PGD. 38 % déclarent avoir besoin d'un outil d'aide à la rédaction et 21 % une session d'information sur les PGD.

Les croisements des besoins avec le poste, d'autres responsabilités et le domaine de recherche ne montrent pas de différences importantes dans les besoins liés à la rédaction d'un PGD. Toutefois, on note que les enseignants-chercheur·es donnent la priorité à un service d'aide à la rédaction et de relecture de leur PGD. Dans les commentaires, un des répondant·es souligne la nécessité d'une assistance concernant les métadonnées : « une assistance pratique, par exemple, au choix ou à la conception et à la mise en place de

modèles de métadonnées » (ingénieur, plus de 50 ans). Un écho se retrouve dans les entretiens :

> *« Finalement ce n'était pas si compliqué. On a utilisé le truc, Opidor là. Je pense que si on avait eu un service d'appui qui aurait pu relire par exemple, on aurait apprécié. »* (chercheuse en sciences du vivant et environnement)

Quant à la sensibilisation et à la formation des doctorant·es en matière de gestion des données de la recherche, 42 % des répondant·es déclarent les encourager à se renseigner sur la gestion des données de recherche, que ce soit par leurs propres moyens (59 %), les formations proposées (52 %) ou la documentation fournie par le laboratoire (17 %). Les croisements avec le poste/ autres responsabilités/ domaine de recherche ne montrent pas de différences significatives, si ce n'est que les directeurs/directrices de laboratoire donnent la préférence aux formations organisées.

D'autres moyens de formation signalés dans la rubrique « Autre » sont la discussion, la participation et la pratique au sein des projets de recherche menés dans les laboratoires. Par exemple, une des répondant·es note : « en pratiquant, via les compétences de l'équipe » (enseignant-chercheur, 30-50 ans).

Pour ce qui concerne les difficultés liées à la gestion des données de la recherche, les répondant·es mentionnent essentiellement le manque de connaissances sur le sujet (52 %) et le manque de temps (50 %). Le manque d'accompagnement et de formation (35 %) et l'absence des ressources matérielles nécessaires (29 %) suivent. Certains répondant·es déclarent n'avoir pas eu de difficultés particulières (25 %). D'autres indiquent le manque d'intérêt pour la question (16 %).

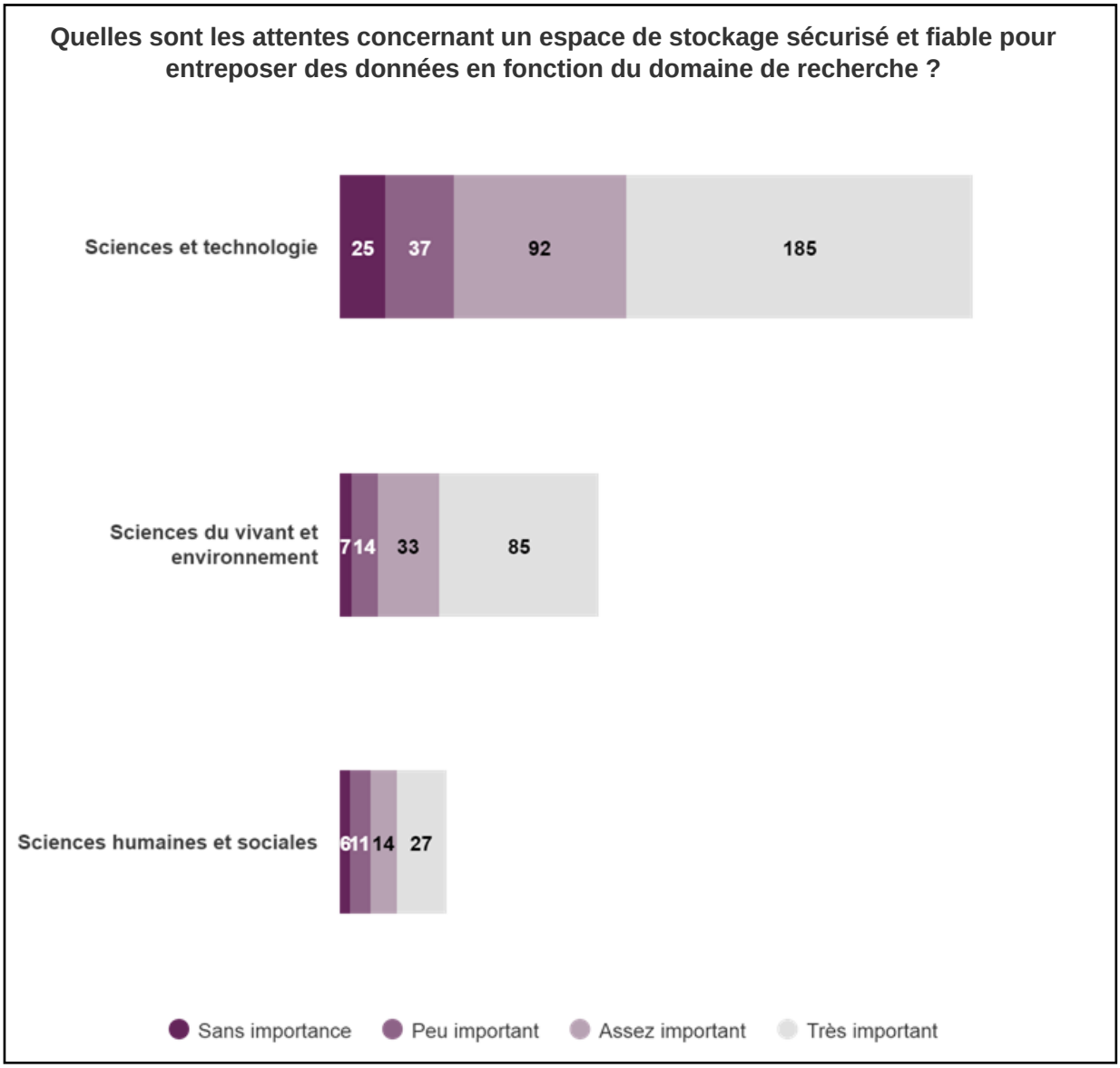
En somme, les acteurs de la recherche sont en demande de cadres, d'outils et de méthodes pour penser le plan de gestion de données. Un chercheur résume la situation en entretien :

> *« J'aurais aimé avoir les outils pour y penser avant, parce que du coup, je me suis retrouvé à partager des données et à réfléchir après au plan de gestion et en fait ce n'est pas comme ça qu'il faut faire quoi. »* (chercheur en sciences et technologie)

5.2. Autour des outils de diffusion et du stockage

Les attentes en termes d'accompagnement à la gestion et à la publication des données sont plutôt fortes dans l'ensemble. Un espace d'archivage sécurisé et fiable pour entreposer des données, tel qu'un entrepôt de données institutionnel, est le service le plus plébiscité par les répondant·es. Plus de la moitié des répondant·es le juge même « très important », ce qui montre en creux l'absence de solutions de stockage et de publication de leurs données à laquelle peuvent être confrontés de nombreux chercheur·es.

Des conseils sur l'archivage des données et les normes de description des données sont également largement attendus : les deux tiers des répondant·es jugent ce service « assez important » ou « très important ». Il est donc intéressant de constater que l'attente concerne non seulement la mise à disposition d'espaces d'archivage mais aussi l'accompagnement à l'usage de ces solutions logicielles. Ces résultats apportent un éclairage intéressant sur le peu de difficultés exprimées concernant un éventuel manque de ressources matérielles (matériel informatique et serveur) mentionné ci-dessus : les équipements existent mais la couche logicielle qui constituerait un véritable espace d'archivage sécurisé et fiable manque encore.



Les attentes en termes de services d'accompagnement apparaissent comme peu corrélées au domaine de recherche, même si l'espace d'archivage sécurisé est davantage attendu en sciences et technologies et sciences du vivant et de l'environnement (plus des trois quarts des répondant·es) que par les chercheur·es en sciences humaines et sociales

(les deux tiers des répondant·es). Ces résultats doivent cependant être interprétés avec prudence du fait d'un nombre de répondant·es significativement plus faible en sciences humaines et sociales dans l'enquête qualitative.

5.3. Un accompagnement aux multiples facettes

Les répondant·es sont une large majorité à attendre des conseils d'ordre juridique sur le traitement des données sensibles ou sur l'archivage des données. L'attente est également forte concernant une offre de formations sur les données de la recherche ou les services d'accompagnement pour la préparation de plans de gestion de données, jugée comme étant « assez important » ou « très important » par environ une moitié des répondant·es. Ce dernier besoin en particulier peut ne pas être identifié par tous comme étant prioritaire, compte tenu du fait qu'un nombre encore important de chercheur·es n'a pas été confronté à la rédaction de plans de gestion de données.

La notion de service d'accompagnement personnalisé est davantage plébiscitée par les répondant·es, ce qui montre la nécessité de développer des services d'accompagnement qui soient adaptés aux spécificités des communautés de recherche, et si possible de proposer des réponses sur-mesure aux besoins des chercheur·es, au niveau même du projet de recherche. Près des deux tiers des répondant·es attendent un service d'accompagnement à la fois personnalisé et global, couvrant l'ensemble des dimensions de la gestion des données.

Enfin, pour toutes les questions, les directeurs et directrices de laboratoires jugent à chaque fois légèrement plus importants (10 à 15 % de réponses « assez important » ou « très important » en plus) les accompagnements proposés que les autres catégories de répondant·es.

Le rôle de la bibliothèque et de la proximité

Les services des bibliothèques sont reconnus pour leur position d'expertise sur les enjeux des données de la recherche, comme l'explique cet enseignant-chercheur en sciences et technologie. Paradoxalement, ils ne sont pas pour autant clairement identifiés comme des interlocuteurs potentiels.

> *« Que ce soit sur l'Open Access ou la gestion des données, effectivement, moi ce que je constate, c'est le décalage entre les services des bibliothèques qui sont super à la page, super en avance sur tout ça et le laboratoire où on a l'impression qu'on est encore il y a dix ans, il y a quinze ans et que finalement ça met du temps à changer, ça met du temps à en discuter. »* (chercheur en sciences et technologie)

L'option préférée évoquée par de nombreux interviewés est un interlocuteur de proximité, de préférence au sein-même du laboratoire, rassemblant en une même personne des compétences propres à la pratique disciplinaire de la recherche et des problématiques liées à l'archivage et la publication des données.

« Non, je pense qu'il faut que ce soit quelqu'un qui prenne en charge l'intégralité de la problématique, c'est-à-dire à la fois sensibilisation, formation, mais aussi aide au référencement, au dépôt, à la valorisation, à la promotion. » (chercheur en sciences humaines et sociales

« Moi ce que j'aime bien, c'est le travail de proximité. Et c'est pour moi [...] extrêmement efficace. Effectivement, il faudrait quelqu'un qui soit dans l'équipe qui puisse effectivement interagir en live pour pouvoir effectivement faire ce genre de choses. » (ingénieur en sciences et technologie)

Le développement du travail en réseau des services d'appui à la recherche, notamment le développement de référents recherche, pourrait venir étancher ce besoin de proximité et d'interlocuteur immédiat.

5.4. Parole libre

La dernière partie de l'enquête, intitulée « À vous la parole », invitait les répondant·es qui le souhaitaient à exprimer librement leurs remarques. 90 répondant·es sur 513 ont laissé un commentaire. Ces remarques sont principalement de trois ordres : elles concernent soit des besoins, soit des problématiques concrètes rencontrées dans la gestion des données, soit l'expression de limites à la politique de gestion et de publication des données de la recherche.

Un besoin plusieurs fois cité par les répondant·es est celui de la formation des doctorant·es et des étudiant·es à la gestion et à la publication des données. Un répondant insiste par exemple sur le fait de sensibiliser très tôt les étudiant·es, alors qu'un autre remarque que les données brutes des doctorant·es sont difficilement exploitables. Cette exigence trouve son écho dans les explications d'un doctorant en sciences et technologie durant son entretien :

« Oui, j'ai déjà discuté un peu "données de la recherche" avec d'autres doctorants, c'est pareil, ils ne savaient rien et il n'y avait rien de formalisé. »

Certaines contributions réclament davantage de lisibilité dans l'offre de services, notamment via une mutualisation plus forte (mise en place d'un site web de documentation technique et juridique sur les données de la recherche, mise en place d'un outil de stockage institutionnel géré par des personnels de l'université, centralisation de l'offre d'accompagnement). Dans le même temps, les contributions soulignent également directement ou non un besoin d'adaptation de l'offre, en envisageant différentes manières de traiter les données selon les communautés.

Les problématiques mentionnées dans cette partie sont de différents ordres. Le manque de consigne et d'accompagnement sur la conservation et l'archivage des données revient plusieurs fois. Un répondant mentionne des difficultés à différencier les données pouvant être publiées librement des données pour lesquelles il existe des restrictions, comme les données personnelles. Enfin un répondant souligne l'éparpillement, au sein même de

l'Université Paris-Saclay, des ressources et outils sur le sujet, là où une autre contribution fait remarquer que le manque de plateforme globale et internationale sur les données est un des principaux freins à une publication efficace. Quelques contributions soulignent la nécessité de disposer de compétences informatiques adaptées à la gestion des données, par exemple via des postes-soutien dans les laboratoires. Plusieurs contributeurs alertent sur le manque de ressources institutionnelles ou jugent que les clouds proposés pour la publication de données sont souvent moins performants que des services tiers, notamment ceux des GAFAM.

Cette partie a enfin été utilisée par les répondant·es pour exprimer un certain nombre de doutes sur le processus de publication des données. Plusieurs soulignent la contrainte que représenteraient ces nouvelles tâches « chronophages » de gestion et de publication des données avec des gains parfois perçus comme faibles, car les données ouvertes ne sont pas toujours intelligibles pour d'autres chercheurs. Les répondant·es sont plusieurs à souligner une tension entre cette politique de publication et une forme de concurrence entre les équipes de recherche, exacerbée notamment par une évaluation de la recherche fondée sur le prestige des revues dans lesquelles les résultats sont publiés. En outre, plusieurs contributions pointent une difficulté particulière en cas de recherches menées dans le cadre de partenariat industriels, la logique de publication des données étant étrangère aux entreprises. D'autres limites sont mentionnées de façon moins récurrente comme, par exemple, le « risque » de voir les données de la recherche produite en France utilisées par d'autres pays, ou des consignes contradictoires selon les tutelles. Enfin, plusieurs répondant·es en mathématiques ont souligné que l'enquête leur paraissait peu appropriée pour leur discipline.

Conclusion

Il est possible de tirer plusieurs enseignements de cette enquête, dans la perspective d'accompagner au mieux la publication des données de la recherche produites au sein de l'Université Paris-Saclay.

Un important travail de sensibilisation aux enjeux et méthodes de la publication de données reste à mener auprès des différentes communautés scientifiques de l'université : les deux tiers des répondant·es n'ont jamais entendu parler de plans de gestion de données. Pour ceux qui y ont déjà eu affaire, la perception du plan de gestion de données est plutôt négative et chronophage, sans que l'intérêt de ce document soit clairement perçu. Un effort particulier est à fournir auprès des doctorant·es, qui apparaissent comme les moins informés sur les enjeux liés aux données.

Le levier de la sensibilisation aux enjeux reste fort au regard des principales motivations affichées pour la publication des données. Celles-ci sont liées à l'adhésion aux principes de la science ouverte, et non à la contrainte que représenterait une meilleure prise en compte de cette pratique dans l'évaluation. Le principal frein à la publication des données, la crainte du plagiat, témoigne des contre-sens à lever.

Le partage des données est une pratique courante dans les laboratoires, mais cantonnée à l'échelle d'une équipe de recherche. La publication des données via un entrepôt dédié reste l'exception. Une politique de la donnée davantage formalisée au niveau institutionnelle et efficacement relayée dans les laboratoires aurait sans doute un important effet d'entraînement pour la publication des données.

L'enquête montre également que la sensibilisation seule ne suffira pas à donner les moyens d'une gestion et d'une publication efficace des données : des solutions informatiques adaptées doivent être proposées. La principale demande des répondant·es concerne en effet la mise à disposition d'espaces numériques fiables et sécurisés pour gérer les données. A défaut, le risque est grand que les GAFAM ou les grands éditeurs scientifiques, via des outils dédiés, captent l'essentiel des données produites. Le travail engagé avec l'ouverture de la plateforme nationale fédérée des données de la recherche *Recherche Data Gouv* et, à l'échelle de l'Université Paris-Saclay, avec le développement de services adossés au mésocentre, doit permettre de répondre à ces attentes.

Le développement de ces services numériques doit pouvoir s'accompagner de formations et de conseils pour les équipes de recherche. Au-delà de la sensibilisation aux enjeux de la science ouverte et de la publication des données, l'enquête a montré l'existence de besoins d'accompagnement sur les aspects techniques, archivistiques et juridiques de la gestion des données. Les services personnalisés et en proximité avec les équipes de

recherche sont les plus plébiscités par les répondant·es. Une meilleure communication sur les services déjà existants reste à mener, car ceux-ci ne semblent pas toujours bien identifiés.

L'enquête confirme donc que, pour être pertinents, les services proposés aux équipes de recherche doivent couvrir l'ensemble des dimensions liées à la gestion des données, tout en s'adaptant aux spécificités des disciplines et des projets de recherche. La structuration en cours des services d'accompagnement à l'échelle du périmètre élargi de l'université, via l'atelier de la donnée de l'Université Paris-Saclay, entre en résonance forte avec les conclusions de cette enquête. Le déploiement des services projetés dans le cadre de l'atelier de la donnée devra permettre de répondre aux attentes exprimées, en garantissant à la fois une cohérence institutionnelle dans le niveau des services proposés à chaque laboratoire, et une nécessaire proximité avec les équipes de recherche.