Enquête quantitative sur les pratiques et les besoins des chercheurs sur la gestion des données de la recherche, algorithmes et codes sources dans les établissements du site toulousain

Danielle Brunet, Soraya Demay, Pierre Diaz, Borbala Goncz, Laure Leclerc, Flora Poupinot, Michelle Sibilla

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Enquête quantitative sur les pratiques et les besoins des chercheurs sur la gestion des données de la recherche, algorithmes et codes sources dans les établissements du site toulousain

Rapport rédigé par :

Danielle Brunet, documentaliste (Institut de Chimie de Toulouse, Université Toulouse III-Paul Sabatier)

Soraya Demay, conservatrice des bibliothèques, administratrice des données, algorithmes et codes sources (Université Toulouse III-Paul Sabatier)

Pierre Diaz, documentaliste (IMT Mines Albi-Carmaux)

Borbala Goncz, ingénieur PUD-T (Maison des Sciences de l'Homme et de la Société de Toulouse)

Laure Leclerc, bibliothécaire (Université Toulouse III-Paul Sabatier)

Flora Poupinot, ingénieur d'études (Université Toulouse III-Paul Sabatier)

Michelle Sibilla, professeur d'informatique (IRIT, Université Toulouse III-Paul Sabatier)

Septembre 2023

Table des matières

Remerciements

L'équipe qui a rédigé ce rapport remercie l'ensemble des personnes qui ont œuvré à la réalisation de cette enquête sur les pratiques et les besoins de la communauté scientifique autour des données de la recherche, algorithmes et codes sources :

Merci aux membres du groupe « Enquête » issus du GT4 Etat des lieux du CéSO : Danielle Brunet, Anne Cambon-Thomsen (directrice de recherche émérite au CNRS, CERPOP, Université Toulouse III), Soraya Demay, Jean-Luc Demonsant (ingénieur PUD-T, Maison des Sciences de l'Homme et de la Société de Toulouse), Pierre Diaz, Borbala Goncz, Jacques Py (professeur de Psychologie sociale, Université Toulouse Jean-Jaurès), Michelle Sibilla.

Un grand merci aux stagiaires Lucas Bertrand et Adrien Nicolas (master 2 Psychologie sociale, du travail et des organisations), ainsi qu'à Valentine Charmay (L3 Mathématiques Informatiques Appliquées aux Sciences Humaines et Sociales).

Merci aux répondants de l'enquête qualitative qui ont accepté de nous rencontrer et de nous accorder de leur temps.

Merci aux personnes qui ont diffusé le questionnaire de l'enquête quantitative aux membres de la communauté scientifique.

Et enfin, merci à tous ceux qui ont répondu au questionnaire de l'enquête quantitative. Leurs réponses constituent une base importante pour mieux connaître les pratiques et les besoins autour des données de la recherche, algorithmes et codes sources et proposer des services d'accompagnement.

Introduction

Les données de la recherche deviennent un enjeu majeur pour la gestion et la diffusion des connaissances scientifiques. Parmi les objectifs du 2ème Plan National pour la Science Ouverte[1], le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation incite la communauté scientifique à favoriser la réutilisation des données (axe 2) et à ouvrir les codes sources face à l'enjeu de reproductibilité des résultats de la recherche (axe 3). En pratique, où en sont les personnels de la recherche avec leurs données et leurs codes sources ?

L'ouverture et le partage des données de recherche est un mouvement international et inter-disciplines scientifiques qui s'appuie sur les principes FAIR : « les données doivent être trouvables, accessibles, interprétables et réutilisables » (Wilkinson et al., 2016)[2]. Des recommandations - qui pourraient devenir des normes - sont formulées pour faciliter un large accès aux données de la recherche (Wu et al., 2019)[3].

Le Comité de réflexion pour le partage et la valorisation des données de la recherche et la coordination de la Science Ouverte (CéSO) de l'Université de Toulouse a pour objectif de favoriser l'ouverture des données de la recherche du site et de coordonner les actions associées à la Science Ouverte. L'université de Toulouse regroupe 31 établissements dont les 143 unités de recherche sont rassemblées au sein de pôles de recherche.

Au printemps 2021, le CéSO s'est lancé dans ce projet d'enquête à destination de l'ensemble de la communauté de recherche de l'université de Toulouse. S'adressant à toutes les disciplines, elle doit permettre de produire un état des lieux des pratiques et des connaissances à l'échelle du site, ainsi que des besoins des chercheurs en matière de gestion des données de la recherche.

Partie 1. Méthodologie

1. Cadre et méthodologie

1.1 Objectifs généraux

Les objectifs généraux de l'enquête étaient d'identifier les connaissances, les pratiques, les outils et les besoins des chercheurs concernant les données de la recherche ainsi que les algorithmes et codes sources. Les résultats permettront de préciser l'offre de services proposée sur le site toulousain.

Cette enquête quantitative fait suite à une série d'entretiens menés aux mois de juin et juillet 2021 auprès d'une quarantaine de chercheurs du site toulousain. Ces entretiens ont permis d'élaborer le questionnaire de l'enquête quantitative qui a été diffusée entre juin et août 2022 auprès de la totalité des chercheurs du site.

1.2. Le questionnaire

Une définition des données de la recherche a été proposée en introduction du questionnaire. Elle est extraite de l'OCDE : « Sous forme d'enregistrements factuels (chiffres, textes, images et sons), elles sont utilisées comme sources principales pour la recherche scientifique, et sont généralement reconnues par la communauté scientifique comme nécessaires pour valider des résultats de recherche. »[4].

[1] MESRI (2021). Ouvrir la Science ! https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-pnso/
[2] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*: 160018 https://doi.org/10.1038/sdata.2016.18
[3] Wu, M., Psomopoulos, F., Jodha Khalsa, S., et de Waard, A. (2019). Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal, 18*(3), pp. 1–13. https://doi.org/10.5334/dsj-2019-003
[4] OCDE (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*. https://www.oecd.org/fr/science/inno/38500823.pdf

Par ailleurs, il est apparu essentiel d'inclure les codes sources et algorithmes dans le cadre de cette enquête. Ainsi, chaque question concerne l'ensemble des données, algorithmes ou codes sources et dans certains cas, les réponses sont collectées séparément pour les codes sources et algorithmes.

Le questionnaire comportait quatre thèmes :

1. Les données de recherche, algorithmes et codes sources utilisés : nature, formats, pratiques de réutilisation, curation des données
2. Les pratiques de gestion de données tout au long du cycle de vie de la donnée : plan de gestion de données, documentation des données, stockage, partage pendant l'activité de recherche et à son issue, archivage
3. Les motivations et les obstacles au partage des données
4. Les besoins et attentes en termes d'infrastructure, d'aide et d'accompagnement à la gestion des données

Le questionnaire contenait surtout des questions fermées, souvent inspirées par les études existantes et l'enquête qualitative[5], cependant, à la fin de chaque partie thématique une question ouverte était proposée aux répondants pour leur donner la possibilité de compléter ou expliquer leurs réponses[6].

Des informations personnelles ont été collectées à la fin du questionnaire : le statut du répondant, l'établissement, le champ disciplinaire, le pôle de recherche. L'unité de recherche était également demandée mais de façon facultative afin de respecter le droit à l'anonymat.

L'enquête a reçu un avis favorable du Comité d'Ethique de la Recherche de Toulouse et le traitement des données personnelles a fait l'objet d'une mise en conformité RGPD.

1.3. Diffusion de l'enquête

Le questionnaire anonyme, de 10 minutes environ, a été mis en ligne sur LimeSurvey du 30 mai au 15 août 2022. Il a été diffusé via les vice-présidences (VP) Recherche des établissements, qui l'ont transmis aux directions d'unités. La diffusion auprès des doctorants a été assurée par l'Ecole des docteurs de Toulouse (15 écoles doctorales). Il y a eu plusieurs vagues de rappels.

Les chercheurs, enseignants-chercheurs, doctorants, jeunes docteurs et post-doctorants, ingénieurs de recherche et ingénieurs d'études et techniciens de la recherche du site toulousain ont été sollicités pour répondre à ce questionnaire.

[5] Prost, H., Schöpfel, J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3 : Rapport final.* Lille 3. https://hal.univ-lille.fr/hal-01198379 ; Bauer, B., Ferus, A., Gorraiz, J., et al. (2015). *Researchers and their data. Results of an Austria survey – Report 2015*. https://doi.org/10.5281/zenodo.34005 ; Serres, A. (dir.), Malingre M.-L., Mignon M., et al. (2017). *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : Une enquête à l'Université Rennes 2—Archive ouverte HAL*. https://hal.science/hal-01635186v2 ; Commission européenne, Direction générale pour la Recherche et l'Innovation (2022). *European Research Data Landscape : final report*, Office des publications de l'Union européenne. https://data.europa.eu/doi/10.2777/3648 ; Questionnaire de l'enquête multi-établissements (2019-2021) financée par le GIS Réseau URFIST. http://gis-reseau-urfist.fr/enquete-donnees-de-la-recherche/

[6] Les questions ouvertes ne sont pas systématiquement analysées dans ce rapport, cependant, nous nous sommes appuyés sur ces réponses pour mieux comprendre le sujet et en citons quelques-unes pour illustrer nos propos.

2. Description de l'échantillon et méthodologie de l'analyse

2.1. Description de l'échantillon

Compte tenu du nombre très important de chercheurs sur le site universitaire toulousain (environ 4 500 chercheurs et enseignants-chercheurs et environ 3 800 doctorants), notre objectif était d'atteindre un taux de participation d'au moins 5%. Au final 547 personnes ont répondu au questionnaire (objectif légèrement dépassé).

L'échantillon n'est pas une photographie exacte des pratiques et attitudes de l'ensemble de la communauté scientifique de l'UT et de ses établissements. Néanmoins, il permet d'analyser les pratiques de gestion des données existantes et d'appréhender leurs spécificités en fonction des pôles.

La recherche au niveau de l'université de Toulouse s'articule autour de six pôles de recherche qui regroupent les 143 structures de recherche du site :

BABS : Biologie, Agronomie, Biotechnologie, Santé

DSPEG : Droit, Science politique, Economie, Gestion

H-SHS : Humanités, Sciences Humaines et Sociétés

MST2I : Mathématiques, Sciences et Technologies de l'Information et l'Ingénierie

SdM : Sciences de la Matière

UPEE : Univers, Planète, Espace, Environnement

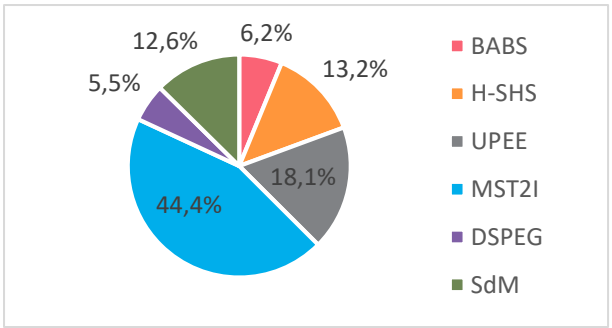

Figure 1 Répartition des répondants par pôle de recherche (%, n=547)

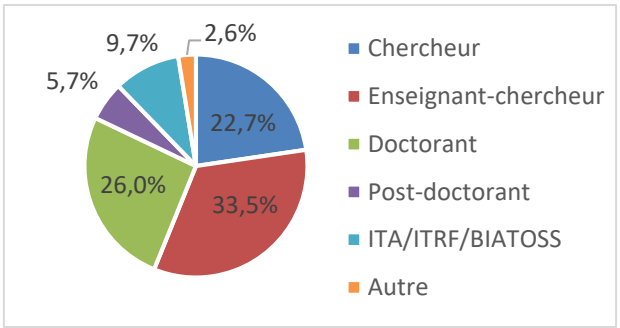

Figure 2 Répartition des répondants par statut (%, n=547)

45% des répondants appartiennent au pôle MST2I, suivi par UPEE, 18%, SdM et H-SHS 13% chacun, les pôles BABS et DSPEG représentant respectivement 6% et 5% de l'échantillon.

La représentativité varie selon les pôles de recherche. Ainsi, les pôles MST2I (28% de la population cible totale[7] vs 45% des répondants) et UPEE (13% vs 18%) sont surreprésentés parmi nos répondants. À l'inverse, le pôle BABS représente 23% de la population du site mais ne forme que 6% de notre échantillon. C'est également le cas, dans une moindre mesure, pour le pôle H-SHS (17% vs. 13% de l'échantillon).

Les chercheurs et enseignants-chercheurs représentent plus de la moitié (56%) de l'échantillon et sont, comme c'est le cas dans plusieurs autres enquêtes sur cette thématique, surreprésentés[8]. En revanche, les ingénieurs et les doctorants ont moins répondu : 26% de l'échantillon étant doctorants, 6% post-doctorants et seulement 10% ingénieurs et personnels techniques et administratifs (cf. Figure 2 Répartition des répondants par statut (%, n=547).

On peut noter que 21 répondants indiquent ne pas utiliser de données de recherche. Dans ce cas, le questionnaire renvoyait directement à la partie Freins et leviers. On peut également supposer que les moins concernés par la gestion des données ont moins répondu à cette enquête portant sur un sujet qu'ils considèrent marginal. Cet effet d'auto sélection issu de l'intérêt porté au sujet a rendu difficile d'atteindre l'ensemble de la population cible lors de la diffusion de l'enquête.

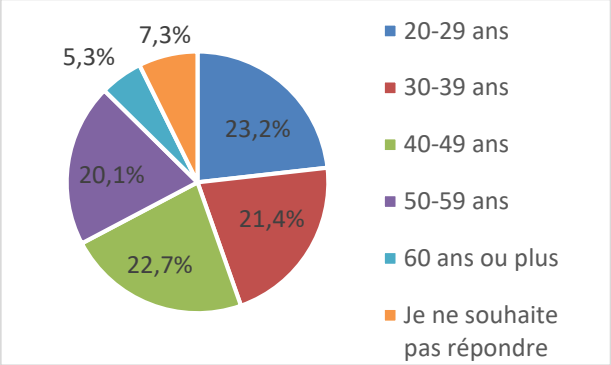

Figure 3 Répartition des répondants par âge (%, n=547)

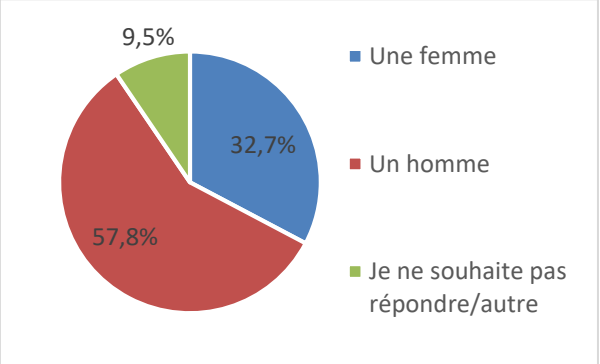

Figure 4 Répartition par genre (%, n=547)

En termes d'âge, les répondants se répartissent de manière égale entre les différentes tranches d'âge, les 20-29 ans, les 30-39 ans, les 40-49 ans et les 50 ans ou plus constituent pour chacun un peu moins d'un quart des répondants. La répartition de genre est un peu moins égalitaire avec un tiers des répondants étant femme. Cependant, on constate une différence de répartition des hommes et des femmes dans les champs disciplinaires. En effet, les femmes sont surreprésentées dans les pôles DSPEG et H-SHS par rapport à la moyenne, alors que les hommes sont en majorité dans les autres pôles. Il existe des tendances similaires par rapport à l'âge, les hommes dans notre échantillon appartiennent davantage à des classes d'âge plus jeunes que les femmes. Les mêmes différences sur l'âge des répondants se retrouvent par pôle : en H-SHS, il y a une surreprésentation des classes d'âge 30-49 ans, alors que le pôle MST2I est plus caractérisé par les 2UT0-29 ans (cf. Annexe, Tableau 4 Répartition, genre, âge, statut par pôle de recherche (%, n=547).

Les établissements et organismes de recherche de l'Université de Toulouse sont presque tous représentés dans l'enquête, bien qu'avec une répartition variable (cf. Annexe, Figure 40 Répartition des répondants par établissement (nombre de réponses, n=547). 20% de l'échantillon est issu de l'université Toulouse III Paul Sabatier (UT3), 8% de l'université Toulouse Jean Jaurès (UT2J) et 4% de l'université Toulouse Capitole (UT

[7] Les données de références sur la population cible proviennent de cette présentation : https://www.univ-toulouse.fr/sites/default/files/2019-03/Poles-de-coordination_UFTMP-2019.pdf

[8] Donati C. S. (2019). *Données de la recherche : Quelles pratiques ? Quels besoins ? Enquête à Aix-Marseille Université.* Aix Marseille Université. https://amu.hal.science/hal-02493679v1 ; Brenel M., Mercier C., Suhan S., Kassas A., Ménard C., et al. (2022). *Rapport d'analyse – Enquête : Les données de la recherche à l'université Paris Saclay, panorama et perspectives*. Université Paris-Saclay. https://universite-paris-saclay.hal.science/hal-03857804

Capitole). Le CNRS a été choisi par 18% des répondants en tant qu'établissement principal, sans autre information sur l'établissement où la personne est affectée. On note, un taux de réponse très important à l'IMT Mines Albi, les répondants de cet établissement représentant 12% de l'échantillon total, bien au-delà de leur proportion dans la population cible.

L'appartenance disciplinaire était recueillie par une question basée sur la nomenclature ERC (Conseil Européen de la recherche)[9]. Après avoir indiqué s'ils appartenaient aux domaines SHS, Sciences et Technologies ou Vie et Santé, les répondants pouvaient préciser en cochant une ou plusieurs disciplines (1,3 en moyenne). Ainsi, 76% de l'échantillon a déclaré appartenir au champ des disciplines des Sciences et Technologies, 19% des Sciences Humaines et Sociales et seulement 6% des disciplines Vie et Santé.

2.2 Méthodologie de l'analyse

Un travail d'analyse a été réalisé sur la correspondance disciplines/pôles. Il y a globalement une très forte association entre les appartenances disciplinaires des répondants et les pôles de recherches auxquelles ils sont affiliés (cf. Annexe, Tableau 5 Nombre de répondants par discipline et appartenance aux pôles de recherche).

Parmi les disciplines des Sciences et Technologies, le pôle MST2I se compose de mathématiciens, physiciens, informaticiens et ingénieurs. Les physiciens (ceux impliqués dans la science physique de la matière condensée : structure, propriétés électroniques, fluides, nanosciences, physique biologique et de la science physique des particules, nucléaire, des plasmas, atomique, moléculaire, des gaz et optique) sont également présents dans le pôle SdM, qui regroupe les chimistes (chimie de synthèse et matériaux, chimie analytique, théorie chimique, chimie physique/physico-chimie). Finalement, on retrouve les répondants des sciences du Système Terre (géographie physique, géologie, géophysique, sciences de l'atmosphère, océanographie, climatologie, cryologie, écologie, changements environnementaux globaux, cycles biogéochimiques, gestion des ressources naturelles) et ceux des sciences de l'Univers (astro-physique/chimie/biologie; système solaire; astronomie stellaire, galactique et extragalactique, systèmes planétaires, cosmologie, science de l'espace, instrumentation) dans le pôle UPEE.

Les répondants dans les pôles DSPEG et H-SHS ont indiqué une ou plusieurs disciplines en Sciences Humaines et Sociales, comme sociologie, psychologie sociale, anthropologie sociale, démographie, éducation, communication, sciences cognitives, psychologie, linguistique, littérature, philologie, études culturelles, étude des arts, philosophie, archéologie et histoire pour le pôle H-SHS, ou économie, finance et management pour le pôle DSPEG. Les répondants qui ont choisi science politique, droit, science de la durabilité, géographie, étude et aménagement du territoire appartenaient, à part égale, aux pôles DSPEG et H-SHS.

Les répondants, bien que peu nombreux, du pôle BABS ont majoritairement indiqué des disciplines en Vie et Santé, comme neurosciences, biologie moléculaire, biochimie, biologie structurale et biophysique moléculaire, génétique moléculaire, génétique quantitative, épidémiologie, écologie des écosystèmes, biologie de l'évolution, écologie comportementale, ou écologie microbienne.

Au-delà de quelques exceptions, chaque discipline pouvait être majoritairement associée à un pôle de recherche.

Ainsi, les pôles de recherche se situent à un niveau de granularité qui permet de représenter la diversité disciplinaire et peuvent être utilisés pour identifier les pratiques propres à chaque communauté.

Au vu de ces résultats, les réponses ont été analysés par pôle de recherche, catégorie structurante au niveau de l'UT. En effet, les conclusions de l'analyse sont pertinentes à cette échelle.

Concernant l'âge et le statut, nous avons essayé de créer des catégories qui soient à la fois pertinentes du point de vue de l'interprétation des résultats et qui contiennent des effectifs suffisants pour effectuer des analyses

[9] « Nomenclature ERC » (MAJ 2022, 15 mars). Cat Opidor. https://cat.opidor.fr/index.php/Nomenclature_ERC

statistiques. Ainsi, les doctorants et les post-doctorants ont été regroupés, en raison de pratiques de gestion des données et d'attitudes par rapport à la Science Ouverte assez proches. De même, la catégorie « Autre » du statut a été regroupée avec la catégorie ITA/ITRF/BIATOSS en raison de pratiques similaires. Pour les catégories d'âge, les 60 ans et plus n'étant pas très nombreux, après une première inspection de leurs pratiques, ils ont été regroupés avec les 50-59 ans en créant une catégorie de 50 ans et plus.

Ainsi, l'analyse des résultats se fait en premier lieu en interprétant la totalité des réponses par question, qui est ensuite décomposée par les effets du pôle de recherche, de l'âge et le statut en ne rendant compte de différences par rapport à ces catégories seulement si elles sont statistiquement significatives[10]. L'effet du genre, de ce point de vue peut être fallacieux : les différences entre les pratiques des hommes et des femmes sont vraisemblablement moins dues à leur genre qu'à leur répartition dans les champs disciplinaires.

[10] Les tests d'association de Chi2 et du V de Cramer ont été utilisés pour déterminer si deux variables sont indépendantes, ainsi que la significativité et l'intensité du lien, avec un seuil de signification de 0,05 - soit 5 % de risque de conclure à tort qu'il existe une association entre deux variables.

Partie 2. Analyse des résultats

3. Données de recherche, algorithmes et codes sources utilisés

Dans cette première partie du questionnaire, nous avons souhaité dresser un aperçu sur les données et codes sources utilisés par les personnels de recherche du site. Les cinq questions posées abordent la classification, le type, la réutilisation, la curation des données et l'usage de format ouvert.

3.1 Typologies des données utilisées pendant les recherches

La diversité des données de la recherche et la difficulté de les classer ont été mises en lumière lors de l'enquête qualitative. Ainsi, un participant indiquait :

> « […] la première chose qu'on a sous les yeux, ce serait le **cahier de laboratoire** […] dans lequel on reporte toutes les expériences qu'on fait […] dans un deuxième temps, il y a les **analyses** qu'on fait et donc qui sont donc très souvent sur des appareils connectés à des ordinateurs qui nous permettent d'enregistrer **les données d'une expérience.** » (SdM)

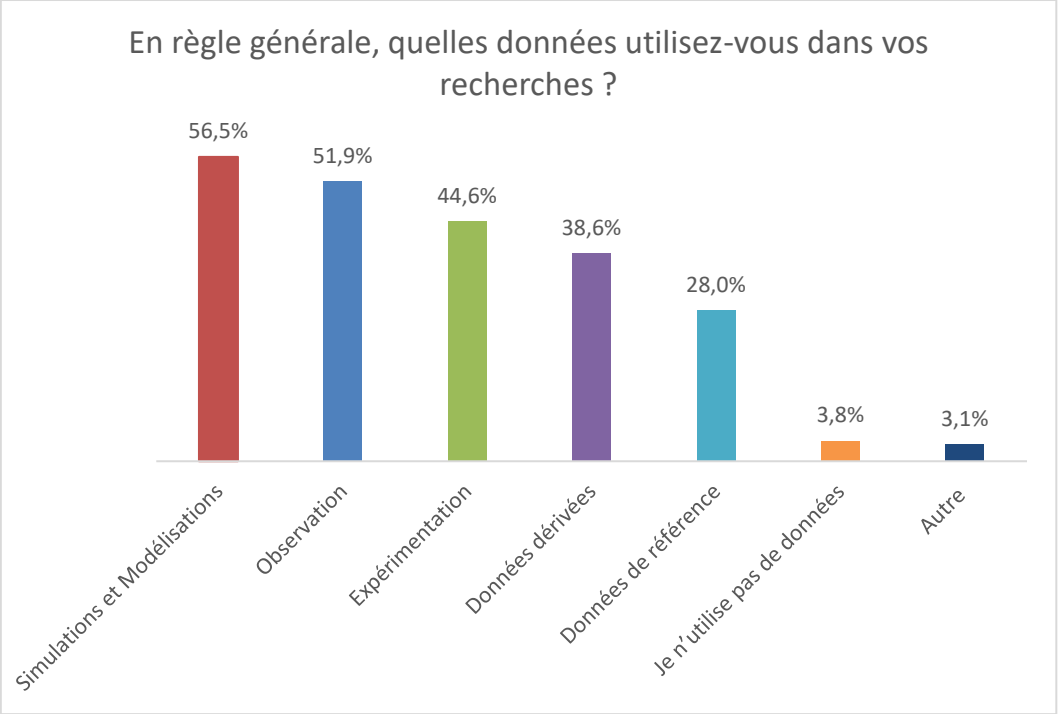

Figure 5 Nature des données utilisées

Les choix de réponses proposés pour cette question ont été extraits de la classification des données de la recherche de Francis André : observation, simulation, expérimentation, données dérivées et de référence[11]. Cette classification est relative aux finalités et procédures de la génération des données.

On note une prédominance des données de simulation et modélisation concentrées au sein des trois pôles Univers-Planète-Espace-Environnement (79 %), Mathématiques, Sciences et Technologies de l'Information et de l'Ingénierie (68 %) et Sciences de la Matière (61 %). Elle s'explique tout d'abord par la surreprésentation de ces trois pôles dans notre échantillon (75 %) (cf. Annexe, Tableau 4 Répartition, genre, âge, statut par pôle de recherche (%, n=547) mais également par leurs pratiques scientifiques générant ce type de données (à titre

[11] André, F. (2015). Déluge des données de la recherche ? Dans Calderan, L., Laurent, P., Lowinger, H., & Millet, J. (dir.), *Big data : nouvelles partitions de l'information. Actes du Séminaire IST Inria*, octobre 2014, pp. 77-95. De Boeck; ADBS, Louvain-la-Neuve.

d'exemples pour le pôle UPEE : modélisations océaniques, sismiques, climatiques ; pour le pôle MST2I : simulation d'aérodynamisme ; et pour le pôle SdM : modélisation de structure et propriétés physiques de systèmes moléculaires).

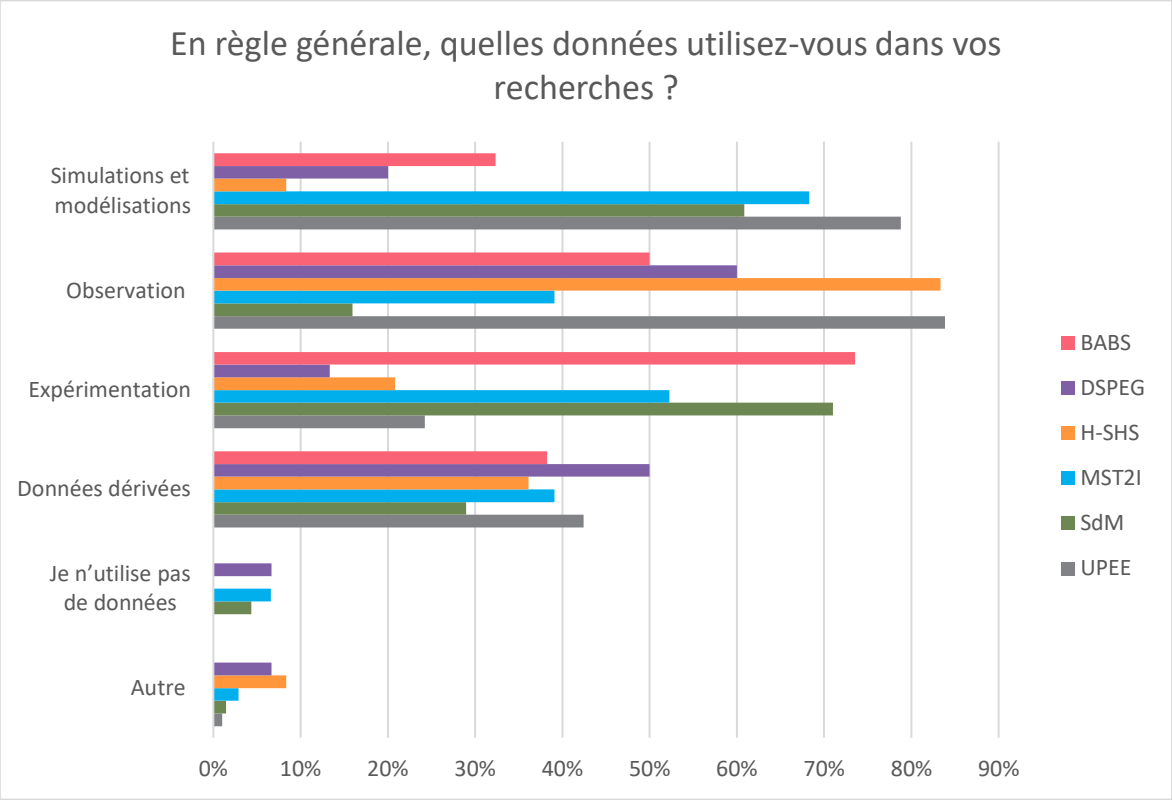

Figure 6 Nature des données utilisées par pôle de recherche

Les données d'observation sont fortement citées par les répondants appartenant aux pôles Univers-Planète-Espace-Environnement (84 %), Humanités-Sciences Humaines et Sociétés (83 %), Droit, Science Politique, Economie, Gestion (60 %) et Biologie, Agronomie, Biotechnologie, Santé (50 %). Elles peuvent être respectivement des mesures d'instruments embarqués dans des satellites, des données d'enquêtes ou d'entretiens, des images ou bien encore des données de capteurs.

Les données d'expérimentation atteignent des taux importants pour les pôles BABS (73 %), SdM (71 %) et MST2I (52 %). Elles sont généralement obtenues en laboratoire à l'aide d'équipements de recherche tels que des machines de caractérisation, de microscopes optiques par exemple, ou de plateformes plus spécifiques.

Les données dérivées (issues de traitement ou de la compilation de données brutes) représentent environ 39% des données utilisées. Elles sont plus fortement citées par les pôles DSPEG (50 %) et UPEE (42 %).

Les données de référence (telles que les banques de données de séquences ADN, de structures chimiques, les portails de données spatiales, ou de données INSEE...) atteignent un taux global de 28 % et sont plus présentes dans les pôles DSPEG (47 %) et BABS (41 %).

Seulement 3,8 % des répondants (appartenant aux pôles DSPEG, MST2I et SdM) ont signalé ne pas utiliser de données.

72 % des répondants ont choisi plusieurs réponses (2 réponses en moyenne), témoignant de la diversité des données présentes au sein des différents pôles dans l'activité de recherche (également relevé lors de l'enquête qualitative).

Dans la suite du questionnaire, il n'était pas demandé de préciser la provenance des données mais il serait intéressant d'établir le lien avec les grandes plateformes d'observation, d'acquisition, d'expérimentation et de simulation du site[12].

La question suivante visait à préciser les types de données.

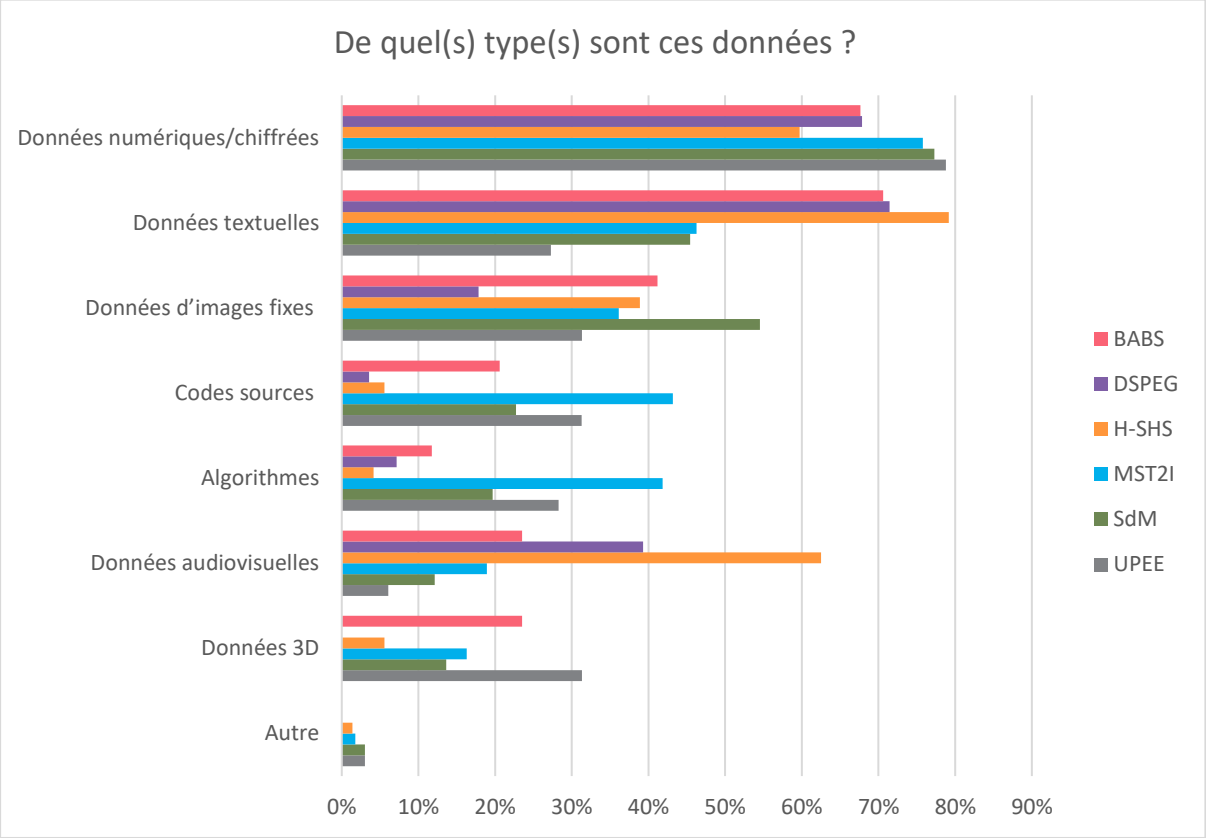

Figure 7 : Type de données utilisées par pôle de recherche

Nous pouvons relever que les données **numériques** prédominent globalement (73 %) et ce dans tous les pôles de recherche (pourcentages variant de 60 % (pôle H-SHS) à 79 % (pôle UPEE)).

Viennent ensuite les données **textuelles** (50 %), fortement présentes au sein des pôles H-SHS (79 %), BABS (71 %) et DSPEG (71 %). Les données d'**images fixes** ont un pourcentage important dans les pôles SdM (54 %) et BABS (41%) et toutefois remarquable pour H-SHS (39 %), MST2I (36 %) et UPEE (31 %).

Les **codes sources** et **algorithmes** atteignent respectivement un taux de 30 % et 28 % et sont plus conséquents au sein des pôles MST2I (43 %) et UPEE (31 %).

[12] CCRRDT (2019). *Politique régionale de la donnée et du numérique : Bilan sur l'activité Recherche et Innovation et Plan d'actions 2020-2025 en région Occitanie*. https://www.laregion.fr/IMG/pdf/atelier_6_ccrrdt_politique_re_gionale_donne_e_nume_rique.pdf

Les **données audiovisuelles** occupent une place importante au sein du pôle H-SHS (62 %) après les données textuelles (79 %).

77 % des répondants ont choisi plusieurs réponses (2,8 réponses en moyenne).

3.2 Réutilisation de données de recherche et/ou algorithmes et codes sources

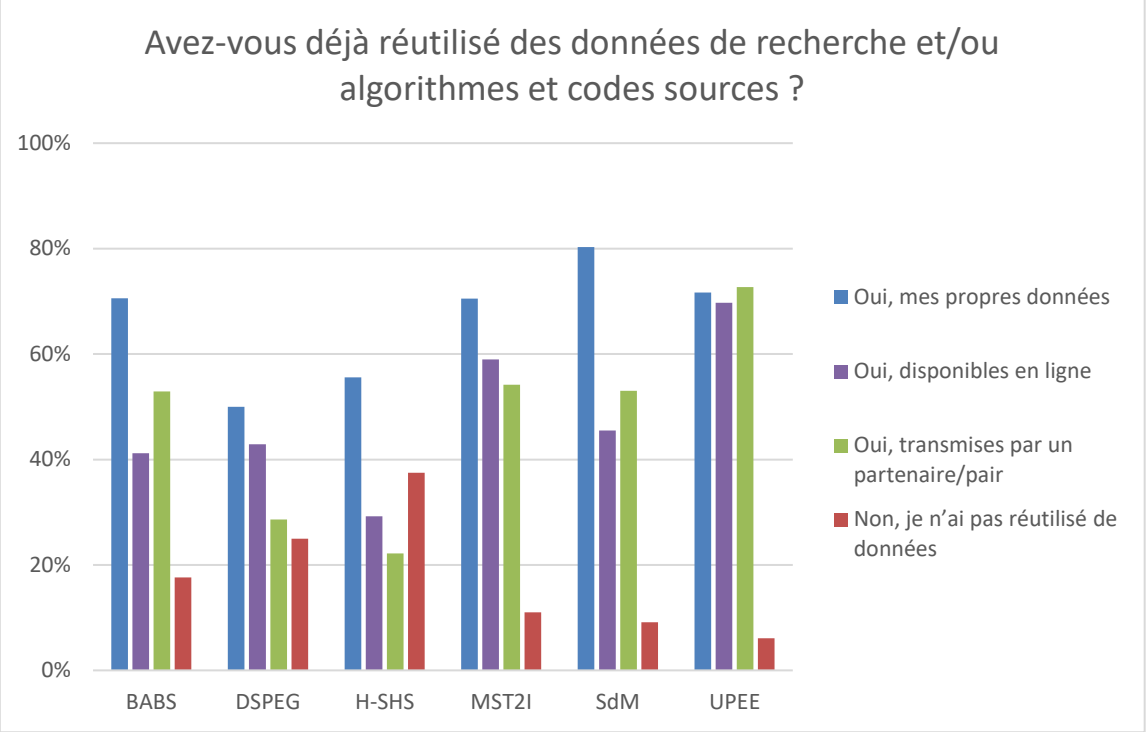

Figure 6 Réutilisation de données et/ou algorithmes et codes sources

Près de 69 % des répondants, indépendamment de leur appartenance à un pôle, réutilisent leurs propres données. La majorité des répondants réutilise des données obtenues en ligne (53 %) ou transmises par des partenaires/pairs (52 %).

Relevons les taux élevés de réutilisation des données pour le pôle UPEE (propres données 72 %, données disponibles en ligne 70 %, données transmises par un partenaire ou un pair 73 %), qui témoignent de pratiques de partage adoptées depuis longtemps par la communauté des Sciences de la Terre et de l'Univers.

Dans l'ensemble, 15 % des enquêtés ont répondu ne pas réutiliser des données, algorithmes ou codes sources (38 % pour le pôle H-SHS).

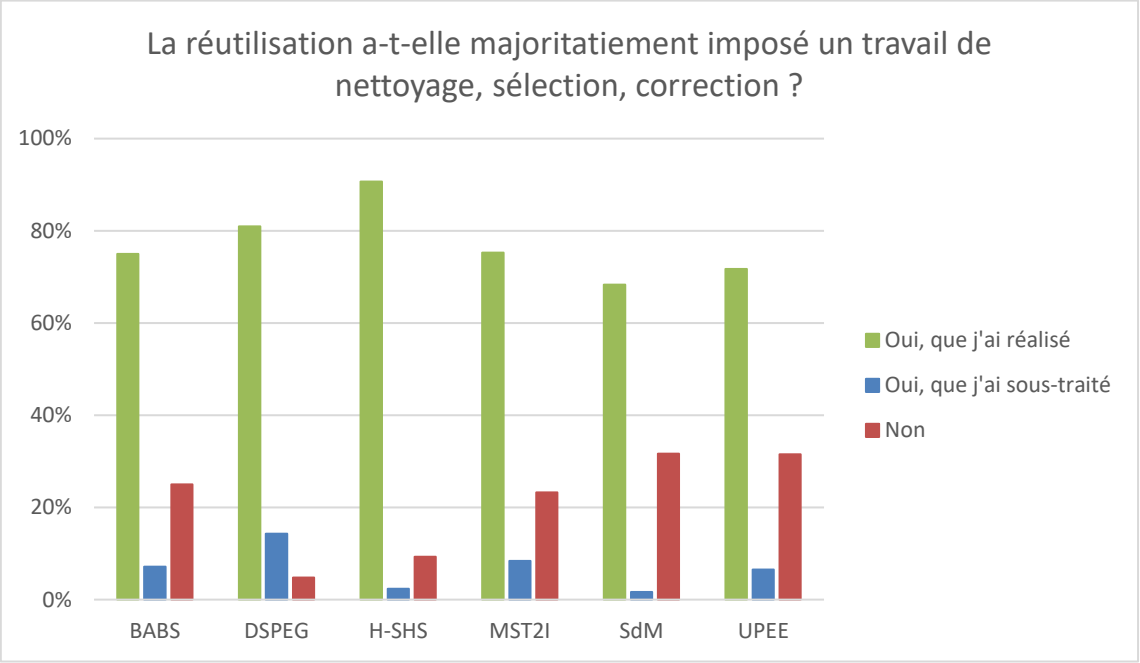

Figure 7 Réutilisation et curation

Deux tendances ressortent quel que soit le pôle :

- La réutilisation des données a majoritairement imposé un travail de nettoyage que les répondants ont réalisé eux-mêmes : de 68 % pour le pôle SdM à 91 % pour H-SHS
- Il y a peu de pratique de sous-traitance de cette tâche : de 2 % pour le pôle SdM à 14 % pour DSPEG

On distingue des pourcentages relativement faibles de réutilisation qui n'ont pas nécessité un traitement en amont (entre 5 % dans le pôle DSPEG et 32 % en SdM).

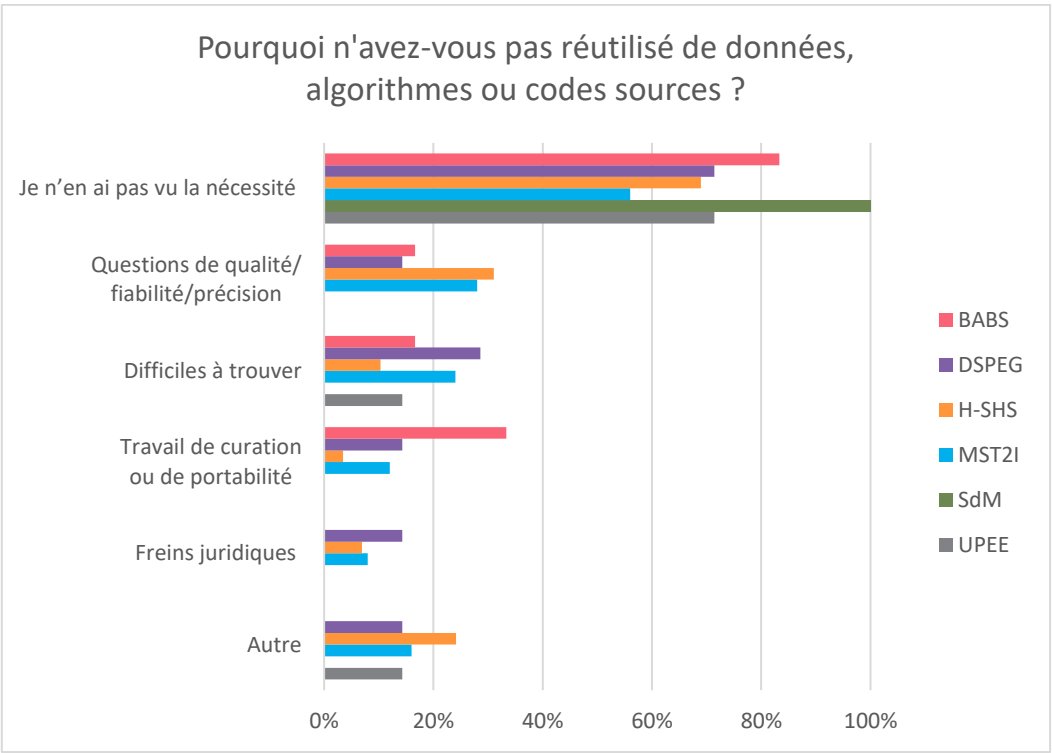

Figure 8 Raisons de "ne pas réutiliser de données, algorithmes et codes sources"

Parmi les 14 % qui ont répondu « ne pas réutiliser de données, algorithmes ou codes sources » à la question « Avez-vous déjà réutilisé des données, algorithmes ou codes sources ? », la motivation majoritaire (69 %) semble être qu'ils n'en ont pas vu l'utilité (cf. Figure 8). Viennent ensuite la question de la qualité (22 %), la difficulté de les trouver (16 %), le travail de curation (9 %) et les freins juridiques (6 %).

Un exemple de difficulté liée au droit données est précisé par un répondant en commentaire de cette question :

> « Plusieurs sont issues de l'aspiration de sites ou réseaux socionumériques effectuées par des organismes publics tiers (organismes d'archivage du web notamment qui possèdent les droits sur ces données, ou interfaces type "crowdtangle" (Meta) » (DSPEG)

3.3 Utilisation de formats ouverts

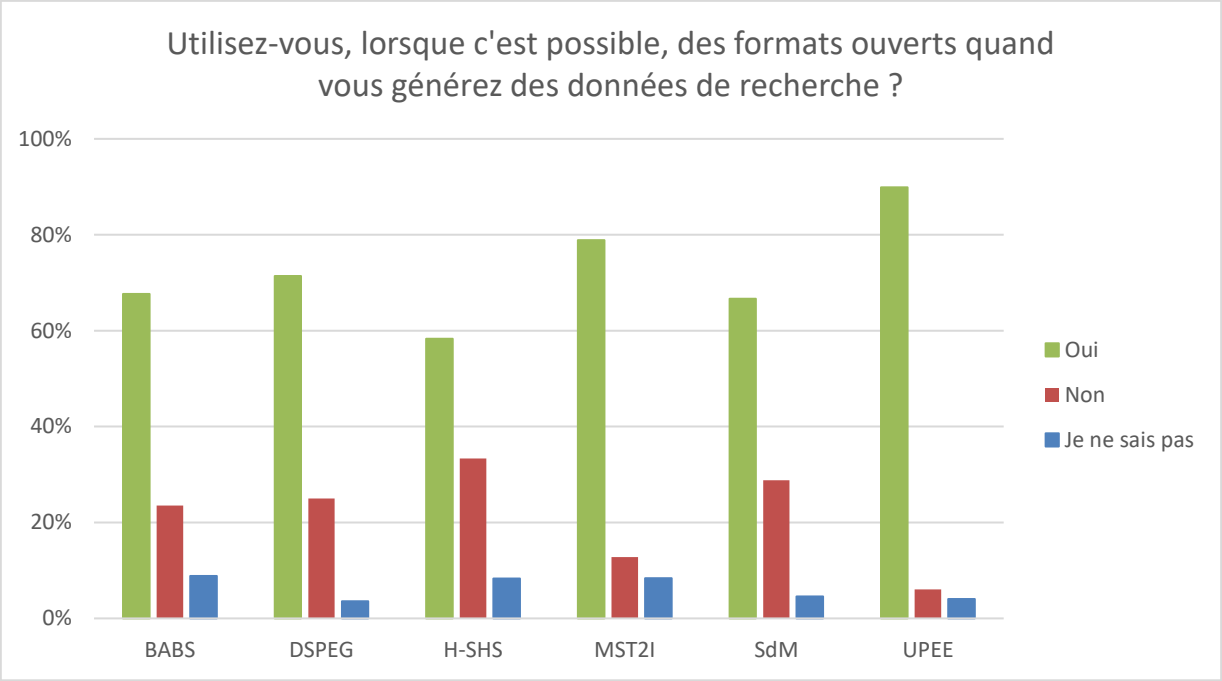

Figure 9 Utilisation de formats ouverts

Les pourcentages obtenus pour la réponse « Oui », tous pôles confondus, laissent penser que la communauté est sensible à l'usage de formats ouverts. Le taux de réponses « Non » est légèrement plus élevé pour le pôle H-SHS (33 %).

Les commentaires de l'enquête ont permis aux répondants de préciser certains standards et formats utilisés dans leur communauté. Ces formats ou standards peuvent être libres :

> « … netcdf (standard dans la communauté "climat"), grib (standard pour les données issues des modèles météorologiques), fa (format natif du modèle météo, principalement interne au service météo) » (UPEE)

Ou propriétaires (notamment issus de machines de caractérisation) :

> « Beaucoup d'équipement de caractérisation sortent les données dans des formats brut liés au logiciel de la machine, exporter les données dans un format ouvert fait souvent perdre une partie des informations sur les conditions d'analyses » (SdM)

> « Je réalise des scans tomographiques, donc je génère des données bruts en .tiff et en .raw à partir d'un logiciel propriétaire » (MST2I)

Outre le format, la structuration des jeux de données est également importante :

> « Concernant les données expérimentales voire aussi issues de simulation (à usage général), il me semble qu'il y a encore beaucoup à expliquer puis faire concernant le bon usage du 'formatage' i.e., l'adoption d'une structuration (et une garantie de pérennisation aussi, sans doute) plus fiable, disons, que des feuilles de tableur, de l'ascii ou du (pseudo, parfois) csv (souvent trop pratiqués). Au-delà du FITS de l'astronomie, le HDF5 p.ex. devrait à terme être utilisé plus souvent, ainsi qu'avec du formatage de nom de fichier lui-même (en utilisant des métadonnées d'importance : lieu et/ou date, nature de l'objet etc.) également. Cela faciliterait sans doute des usages multiples voire à long terme. » (UPEE)

4. Pratiques de gestion de données

Dans la deuxième partie du questionnaire, nous souhaitions connaître les habitudes et les pratiques de gestion de données. Les questions posées abordent tout le cycle de vie de la donnée depuis la planification et la description des données, algorithmes et codes sources jusqu'à l'archivage et la diffusion après le projet de recherche.

4.1. Planification et description

4.1.1. Plan de gestion des données

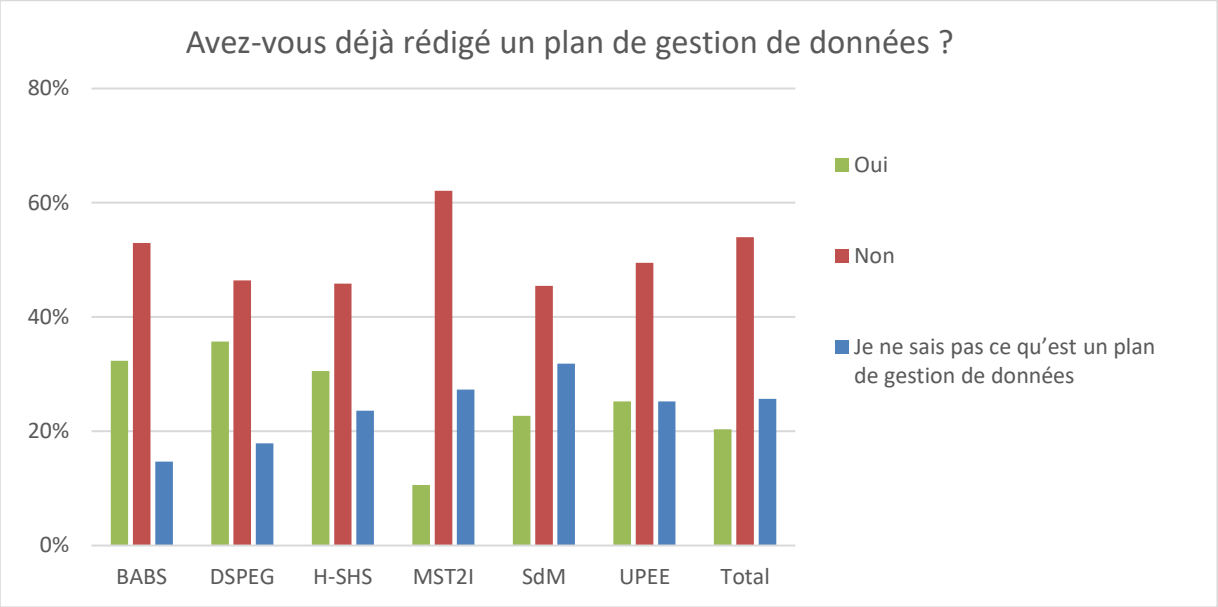

Figure 10 Rédaction de plan de gestion de données

20 % Oui
54 % Non
26 % Je ne sais pas ce qu'est un plan de gestion de données

Le pourcentage de répondants ayant participé à la rédaction d'un plan de gestion de données (PGD) est globalement peu élevé (20 %).
Le statut influe logiquement sur le fait d'avoir déjà pris part à la rédaction de ce type de document. Ainsi, seuls 7 % des doctorants et post-doctorants indiquent y avoir déjà participé. Ce pourcentage est plus élevé pour les chercheurs (24 %) et les enseignants-chercheurs (29 %).

On note également que le PGD n'est pas encore largement connu : plus d'un quart des répondants a déclaré ne pas savoir de quoi il s'agit. La non-connaissance du PGD est plus particulièrement marquée chez les doctorants et post-doctorants (43 %) ainsi que chez les répondants appartenant aux pôles SdM (32 %) et MST2I (27 %).

4.1.2. Documentation des données, algorithmes et codes sources

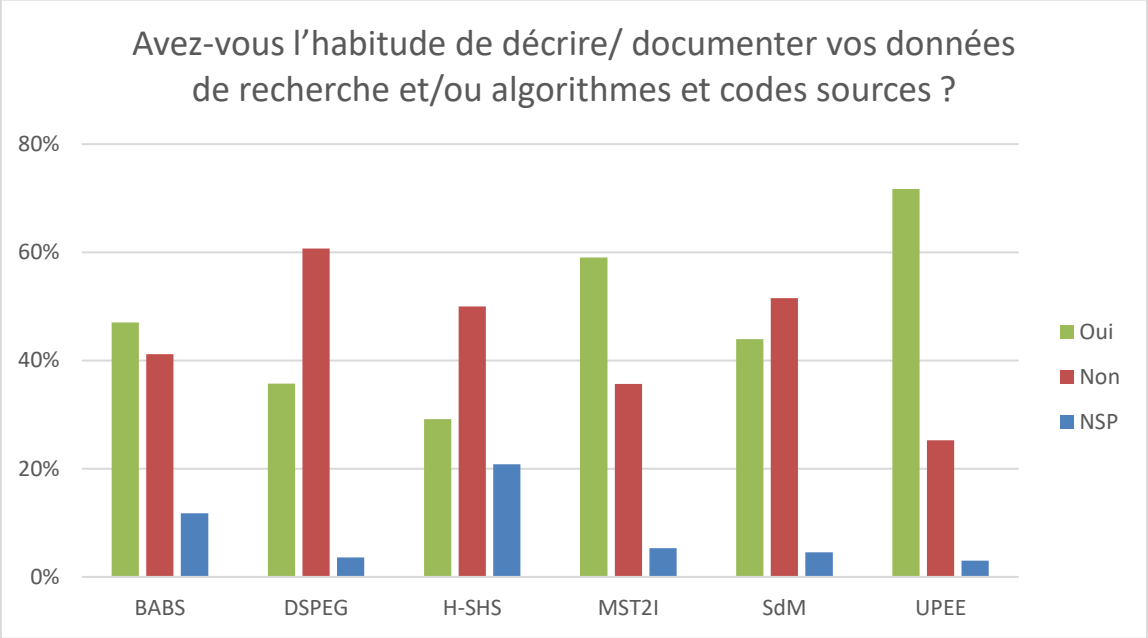

Figure 11 Habitude de documentation des données de recherche et/ou algorithmes et codes sources

53 % des répondants déclarent avoir l'habitude de décrire/documenter leurs données. La documentation des données est une pratique particulièrement importante dans les pôles UPEE (72 %) et MST2I (59 %). Elle est plus faible dans le pôle H-SHS (29 %) mais c'est également dans ce pôle que 21 % des enquêtés ont répondu « Je ne sais pas » ce qui laisse penser que la question n'a pas été comprise pour une partie d'entre eux.

Quel que soit leur statut, les répondants documentent leurs données de manière relativement homogène.

Néanmoins si les doctorants et post-doctorants décrivent moins leurs données (46 %) que les autres (53 % des enseignants-chercheurs ; 61 % des chercheurs et personnels ITA/TRF/BIATOSS), ils sont aussi moins familiers avec cette pratique (13 % ont répondu « Je ne sais pas »).

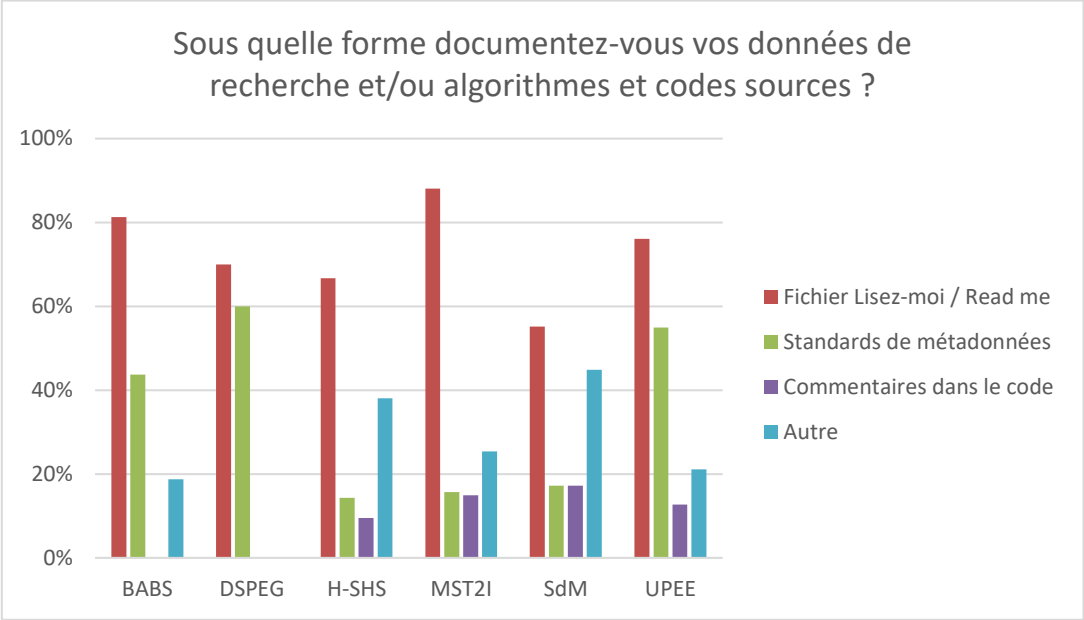

Figure 12 Forme de documentation des données de recherche et/ou algorithmes et codes sources

La documentation des données, algorithmes et codes sources peut prendre différentes formes en fonction des disciplines ou des objets décrits. Cette question à choix multiple témoigne de l'utilisation de plusieurs moyens par les répondants pour les documenter.

Accompagner les données, algorithmes et codes sources d'un fichier Lisez-moi/Read me est la pratique de documentation la plus répandue dans tous les pôles. Elle apparaît plus utilisée dans les pôles MST2I (88 %) et BABS (81 %).

Cette forme de documentation est également citée par les répondants qui ont choisi « Autre » et ont précisé leur réponse telle que, par exemple, « document texte », « documents textes/notes donnant des détails sur les données », « Je décris textuellement le contexte dans lequel elles ont été recueillies ».

Parmi ceux qui documentent leurs données, 29 % utilisent des standards de métadonnées. La grande majorité d'entre eux (85 %) travaille dans les disciplines de Science et technologies. Le recours aux standards de métadonnées varie selon la discipline. Ils sont plus utilisés en sciences du système Terre, en sciences informatiques ainsi qu'en sciences de l'Univers. Par ailleurs, le pourcentage élevé de personnes ayant recours aux métadonnées dans les pôles DSPEG (60 %) et BABS (44 %) ne peut pas être considéré représentatif des pratiques de l'ensemble du pôle en raison du faible nombre de répondants (10; 16). En DSPEG, les répondants qui documentent leurs données utilisent à la fois un fichier Lisez-moi/ Read me et des standards de métadonnées.

La réponse « Autre » permettait aux répondants de spécifier les autres formes de documentation utilisées. Ainsi, 13 % ont recours aux commentaires intégrés pour documenter du code. Ce pourcentage était important, une catégorie à part a été créée. Parmi les autres formes de documentation citées, il y a le cahier de laboratoire (BABS, SdM) et les articles (SdM, MST2I).

4.1.3. Données de recherche et RGPD

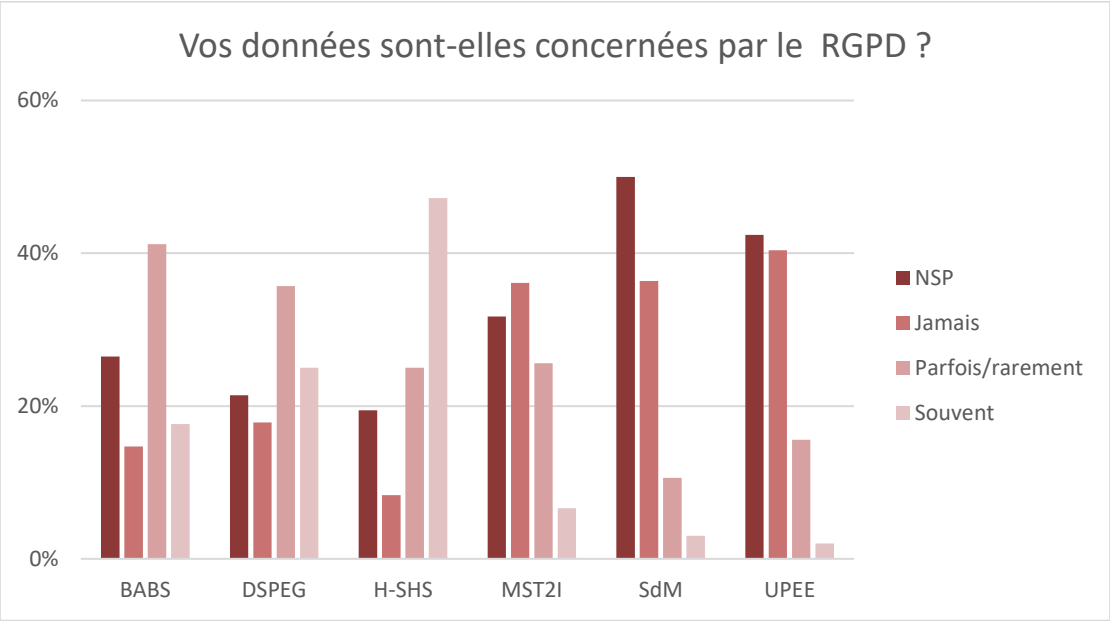

Figure 13 Données concernées par le RGPD par pôle de recherche

Concernant le Règlement général sur la protection des données (RGPD), les répondants se répartissent en trois groupes équilibrés :

- 36 % ont des données concernées par le RGPD dont 13 % déclarent avoir des données souvent concernées.
- 31 % ont des données qui ne sont pas concernées
- 34 % ne savent pas si leurs données sont concernées

Les personnes ayant des données concernées par le RGPD appartiennent principalement aux pôles H-SHS (72 %), DSPEG (61 %) et BABS (59 %). Parmi ces dernières, les personnes du pôle H-SHS sont plus nombreuses à déclarer avoir des données « souvent » concernées (47 %).
Dans les pôles MST2I, SdM et UPEE, le pourcentage de répondants ayant des données concernées par le RGPD est moins élevé (entre 14 % et 32 %). Ils déclarent également être moins fréquemment concernés que les autres pôles.
Cette répartition est cohérente avec les objets étudiés dans les disciplines recouvertes par les pôles.

Un tiers des répondants indiquent ne pas savoir si leurs données sont concernées par le RGPD. Le pourcentage élevé dans les pôles SdM (50 %) et UPEE (42 %) peut être expliqué par la nature des objets étudiés dans les disciplines de ces pôles.

Enfin, 43 % des doctorants et post-doctorants ne savent pas s'ils ont des données concernées par le RGPD. Ces derniers appartiennent principalement aux pôles MST2I (44 %) et UPEE (21 %) et ont très majoritairement déclaré « Sciences et technologies » pour champ disciplinaire principal (78 %).

4.2 Stockage, sauvegarde et accès aux données au cours d'une activité de recherche

4.2.1. Stockage

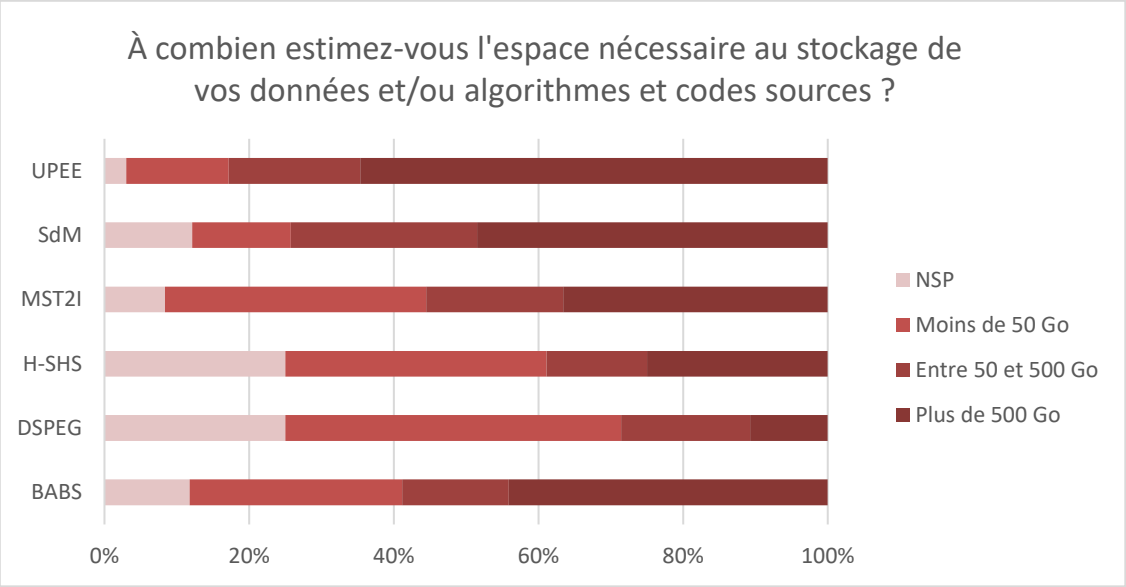

Figure 14 Espace nécessaire au stockage des données et/ou algorithmes et codes sources par pôle de recherche

Au sein de chaque pôle, il y a une nécessité de disposer d'espace de stockage de capacité élevée.

Les réponses « moins de 2 Go » et « moins de 10 Go » recueillies étant très faibles, elles ont été regroupées avec les « moins de 50 Go ». Les réponses « entre 500 Go et 1 To » et « plus d'un To » ont aussi été rassemblées.

Parmi les enquêtés des pôles SdM et MST2I, respectivement 49 % et 37 % déclarent avoir besoin de plus de 500 Go pour le stockage de leurs données tandis qu'ils sont 65 % dans le pôle UPEE. Néanmoins, dans ce dernier, les répondants déclarent avoir plus besoin d'une capacité de stockage de plus d'un To (48 %) que située entre 500 Go et 1 To (17 %).

À noter que dans les pôles regroupant les disciplines de sciences humaines et sociales (H-SHS et DSPEG), les données de recherche peuvent exiger des espaces avec une capacité de stockage relativement importante : 39 % des répondants du pôle H-SHS et 29 % des répondants du pôle DSPEG indiquent avoir besoin d'un espace de stockage d'au moins 50 Go.

Cependant, pour 11 % des répondants, il est difficile d'évaluer l'espace nécessaire au stockage de leurs données et/ou algorithmes et codes sources. Cette difficulté touche 25 % des répondants dans les pôles DSPEG et H-SHS.

La difficulté de disposer d'une capacité de stockage adaptée est un des aspects abordés dans les commentaires libres de l'enquête :

> « Difficulté à trouver la bonne méthode pour stocker et avoir des données diffusable (fichier de plusieurs Go, dataset de plusieurs To..) » (MST2I)
>
> « Notre serveur CNRS Owncloud avec 100 Go est insuffisant pour stocker mes données de recherche. Il me faudrait 300 Go. Je suis donc obligée d'utiliser un disque dur externe, ce qui est très stressant pour moi. » (H-SHS)

Tableau 1 Solutions de stockage utilisées par pôle

	BABS	DSPEG	H-SHS	MST2I	SdM	UPEE
Ordinateur professionnel	**70,6 %**	**64,3 %**	**65,3 %**	**82,4 %**	**78,8 %**	**75,8 %**
Serveur de ma structure	**76,5 %**	28,6 %	36,1 %	**61,7 %**	**59,1 %**	**78,8 %**
Support externe (disque dur externe, clé USB…)	**52,9 %**	**50,0 %**	**77,8 %**	**50,2 %**	**66,7 %**	**50,5 %**
Cloud institutionnel	29,4 %	32,1 %	27,8 %	28,2 %	22,7 %	18,2 %
Ordinateur personnel	14,7 %	**53,6 %**	**47,2 %**	14,1 %	18,2 %	15,2 %
Cloud privé	5,9 %	39,3 %	29,2 %	11,0 %	4,5 %	5,1 %
Autre	0,0 %	0,0 %	2,8 %	3,5 %	10,6 %	8,1 %

Tous pôles confondus, les supports de stockage les plus utilisés sont l'ordinateur professionnel (77 %), un support externe (clé USB, disque dur, etc.) et le serveur géré par la structure (60 %).

Néanmoins, il existe des différences de pratiques entre les pôles :

- Le serveur de la structure : au sein des pôles BABS, MST2I, SdM et UPEE, le serveur de la structure est utilisé par 59 % à 79 % des répondants tandis que pour les pôles DSPEG et H-SHS, il est respectivement utilisé par 29 % et 36 % des répondants.
- L'ordinateur personnel : au sein des pôles DSPEG et H-SHS, il est respectivement utilisé comme lieu de stockage par 54 % et 47 % des répondants tandis que dans les pôles BABS, MST2I, SdM et UPEE, moins de 20 % des répondants disent l'utiliser pour stocker leurs données.
- Le support externe : si dans chaque pôle la majorité de répondants déclare utiliser cette solution de stockage, leur pourcentage est plus élevé en H-SHS (78 %).

Cette question à choix multiple montre que la multiplication des supports de stockage est courante.
Aussi, le stockage « en réseau » (= sur un serveur géré par la structure) n'écarte pas l'utilisation de supports externes. Dans le pôle SdM, le stockage sur un serveur géré par la structure (60 %) coexiste avec une utilisation importante de supports externes (67 %).

Les solutions de stockage de type cloud, qu'elles soient institutionnelles ou privées, sont peu utilisées. 26 % des répondants utilisent un cloud institutionnel tandis qu'ils sont 13 % à se tourner vers un cloud privé. Le cloud privé est beaucoup plus utilisé dans les pôles DSPEG (39 %) et H-SHS (29 %) que dans les autres pôles (entre 5 et 11 %).

On peut s'interroger sur les motivations de cette utilisation (absence de solution de stockage dans la structure ?, besoin « simple » et « rapide » de partager des données au cours d'un projet ?, etc.).

20 % des doctorants et post-doctorants utilisent un cloud. Ils sont par ailleurs 37 % à utiliser leur ordinateur personnel et 56 % à utiliser un ordinateur professionnel.

Lors des entretiens, les participants ont dit parfois raisonner en rapport coût/bénéfices pour déterminer leur choix de solution de stockage :

> « Un endroit où je pourrais mettre, un endroit sécurisé les données qui sont actuellement sur les disques externes pour être sûr qu'elles sont sauvées pendant peut-être 15 ans et l'assurance qu'on n'a pas un coût démentiel pour les établissements et pour la planète, parce que je voudrais savoir ça je veux bien faire cet effort là mais il faut me montrer l'analyse coûts bénéfices par rapport à mon stockage sur un disque dur externe » (SdM)

La taille de l'espace nécessaire pour le stockage des données donne-t-elle lieu à des usages spécifiques de solution de stockage ? Pour tenter de répondre à cette question, nous avons croisé les réponses concernant l'espace nécessaire au stockage avec les réponses des solutions de stockage utilisées. Les résultats donnent quelques pistes.

Tableau 2 Solution de stockage utilisée en fonction de l'espace nécessaire au stockage des données

	Moins de 50 Go	Entre 50 et 500 Go	Entre 500 Go et 1 To	Plus de 1 To
Sur mon ordinateur professionnel	**81,2%**	75,5%	74,2%	73,8%
Sur un serveur géré au sein de ma structure	49,4%	54,1%	**66,7%**	**82,0%**
Sur un support externe	42,2%	**56,1%**	**69,9%**	**60,7%**
Sur un cloud institutionnel	27,9%	**35,7%**	26,9%	19,7%
Sur mon ordinateur personnel	**27,3%**	23,5%	18,3%	**11,5%**
Sur un cloud privé	14,3%	14,3%	15,1%	9,0%
Autre	3,9%	1,0%	4,3%	10,7%

Plus l'espace de stockage nécessaire est important, plus les répondants utilisent le serveur géré par leur structure ou un support externe. Pour les besoins d'espace de stockage entre 50-500 Go et 500 Go-1 To le recours au support externe est équivalent (56 % ; 70 %) à l'utilisation du serveur géré par la structure (54 % ; 67 %). Au-delà d'un To, la tendance s'inverse. Les répondants qui ont besoin de plus d'un To utilisent plus le serveur géré au sein de leur structure (82 %) qu'un support externe (61 %).

Le stockage sur l'ordinateur professionnel reste homogène. Il oscille entre 74 % et 78 % pour les répondants qui ont besoin de moins de 2 Go, moins de 10 Go, entre 50 et 500 Go, entre 500 Go et 1To et plus d'un To. Cependant,

chez les répondants qui ont besoin de plus de 10 Go mais de moins de 50 Go, il est utilisé pour stocker des données par 88% d'entre eux.

Enfin, plus l'espace nécessaire au stockage des données est faible, plus le pourcentage de répondants ayant recours à l'ordinateur personnel est élevé. L'ordinateur personnel est plus utilisé par les personnes ayant besoin d'un espace de stockage de moins de 50 Go (27%) et entre 50 et 500 Go (24%). Au-delà de 500 Go, le pourcentage de répondants qui déclarent y stocker des données est plus faible.

La taille de l'espace nécessaire pour le stockage des données ne semble pas avoir un grand impact sur l'utilisation d'un cloud privé.

4.2.2. Sauvegarde pendant l'activité de recherche

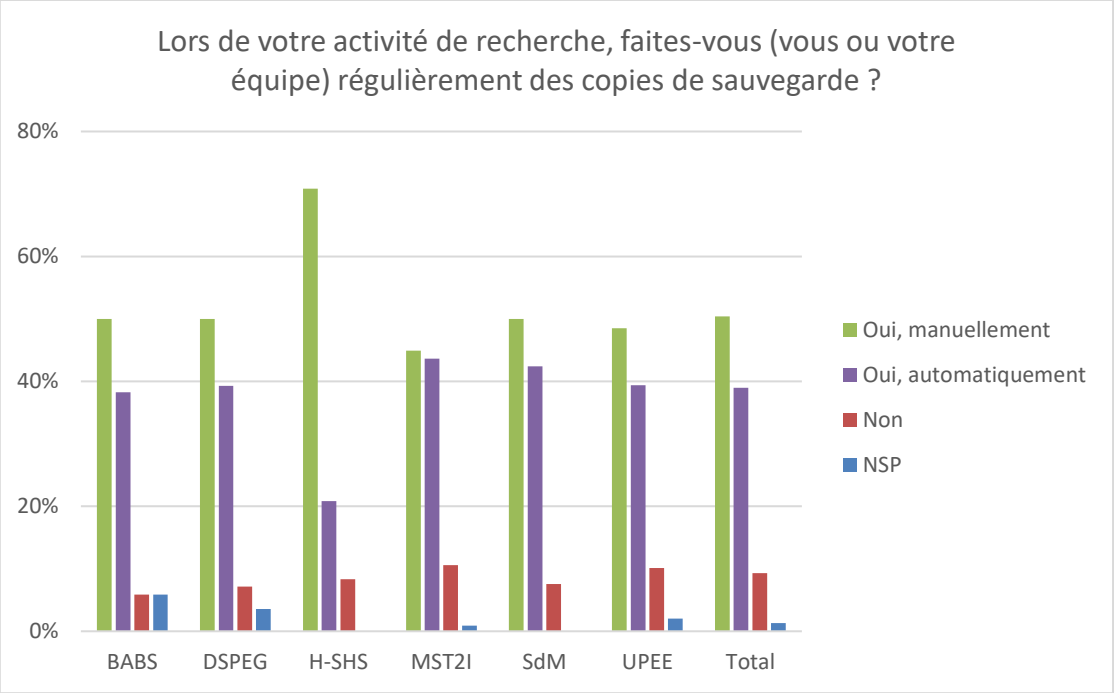

Figure 15 Sauvegarde pendant l'activité de recherche

La sauvegarde des données est une pratique très répandue. Moins de 10 % des répondants indiquent ne pas faire de sauvegarde de leurs données liées à des activités de recherche.

La sauvegarde manuelle est la plus répandue pour l'ensemble des pôles de recherche. Elle se retrouve plus particulièrement dans le pôle H-SHS (71 %).

La sauvegarde automatique est pratiquée dans tous les pôles (sauf H-SHS) par près de 40 % des répondants.

La pratique de sauvegarde n'est pas corrélée au statut : tous font des sauvegardes mais le type de sauvegarde diffère selon le statut. Ainsi, les sauvegardes manuelles sont plus pratiquées par les enseignants-chercheurs (55 %) que les chercheurs (39 %). Les doctorants et post-doctorants sauvegardent en majorité de façon manuelle (57 % contre 21 % de manière automatisée).
Parmi les répondants, 9 % indiquent ne pas faire de copie de sauvegarde. Il s'agit de personnels ITA/TRF/BIATOSS (14 %) et de doctorants et post-doctorants (13 %) tandis que les enseignants-chercheurs et les chercheurs sont 6 %. Néanmoins, le nombre de répondants n'effectuant pas de sauvegarde est faible.

Le stockage et la sauvegarde pendant l'activité sont corrélés au volume des données. Ainsi, en croisant la volumétrie avec la solution de stockage, on observe que les données de petite taille sont sauvegardées sur l'ordinateur professionnel (81 %), tandis que les gros volumes (supérieur à 1 To) sont en très grande majorité sauvegardés sur un serveur de la structure (82 %).

La question de la sauvegarde automatisée ou manuelle est en lien avec les solutions de stockage utilisées. Ainsi, les répondants des pôles BABS, MST2I, SdM et UPEE qui utilisent plus le serveur géré par leur structure pour le stockage font plus de sauvegarde automatique tandis que ceux du pôle H-SHS qui stockent principalement sur un support externe, leur ordinateur professionnel ou personnel, font plus de sauvegarde manuelle. 83 % des répondants du pôle H-SHS qui font une sauvegarde manuelle stockent leurs données et/ou algorithmes et codes sources sur un support externe.

4.2.3. Accès aux données

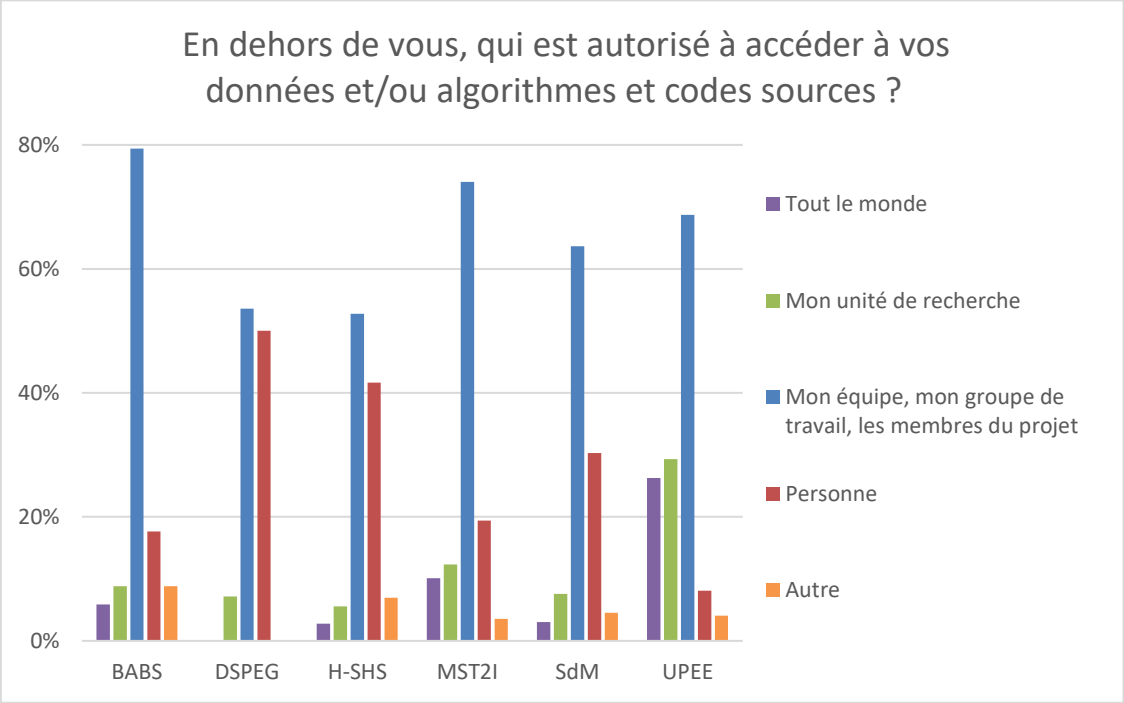

Figure 16 Accès aux données et/ou algorithmes et codes sources pendant l'activité de recherche

Cette question avait pour objectif de connaître les pratiques de partage de données, algorithmes et codes sources pendant une activité de recherche.
Quel que soit le pôle, la majorité des répondants déclare donner accès à ses données à son équipe, son groupe de travail ou aux membres du projet. Il existe néanmoins des disparités dans les pratiques : 79 % et 74 % des répondants des pôles BABS et MST2I permettent à leur équipe d'accéder à leurs données pendant leur recherche contre 54 % et 53 % des répondants dans les pôles DSPEG et H-SHS. Dans les pôles DSPEG et H-SHS, 50 % et 42 % des répondants déclarent que personne n'a accès à leurs données, algorithmes et codes sources au cours du projet contre 8 à 30 % dans les autres pôles. Ceci peut relever d'une pratique de la recherche plus individuelle dans les disciplines de ces pôles.
Le pôle UPEE a une pratique de partage des données plus ouverte durant le projet de recherche. Au-delà de l'équipe de recherche (69 %), les données sont ouvertes à l'ensemble de l'unité de recherche (29 %) et à tous (26 %).

Il est également observé dans notre échantillon de répondants que les chercheurs partagent plus avec leur unité de recherche (24 %) que les enseignants chercheurs (5 %).

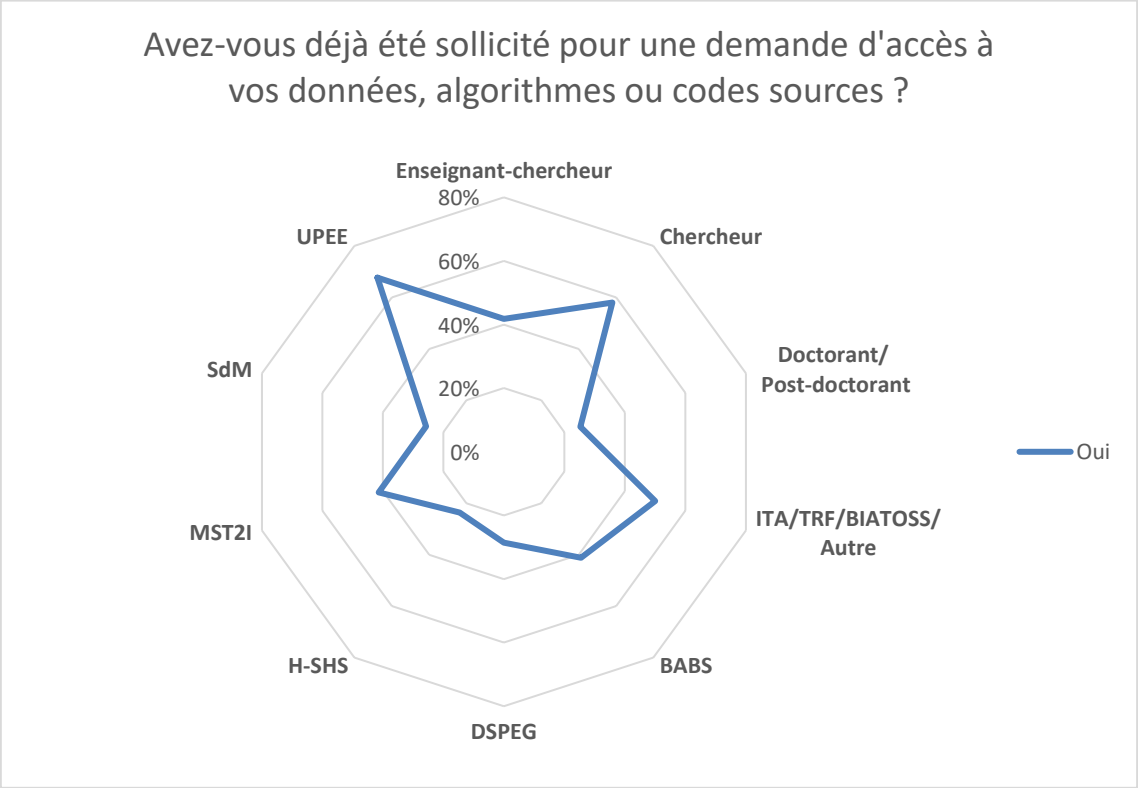

Figure 17 Sollicitation pour une demande d'accès aux données, algorithmes et codes sources

41 % des répondants ont déjà été sollicités pour une demande d'accès à leurs données. Cela concerne principalement les chercheurs (58 %) et dans une moindre mesure les enseignants-chercheurs (42 %). Les doctorants et post-doctorants sont moins sollicités (25 %).

Le pôle UPEE est le plus concerné (68 %), ce qui peut s'expliquer par leur activité de production de données, viennent ensuite les pôles BABS et MST2I (41 % chacun).

4.3 Archivage et diffusion

Cette partie s'intéresse au devenir des données, algorithmes et codes sources à l'issue du travail de recherche.

4.3.1. Archivage

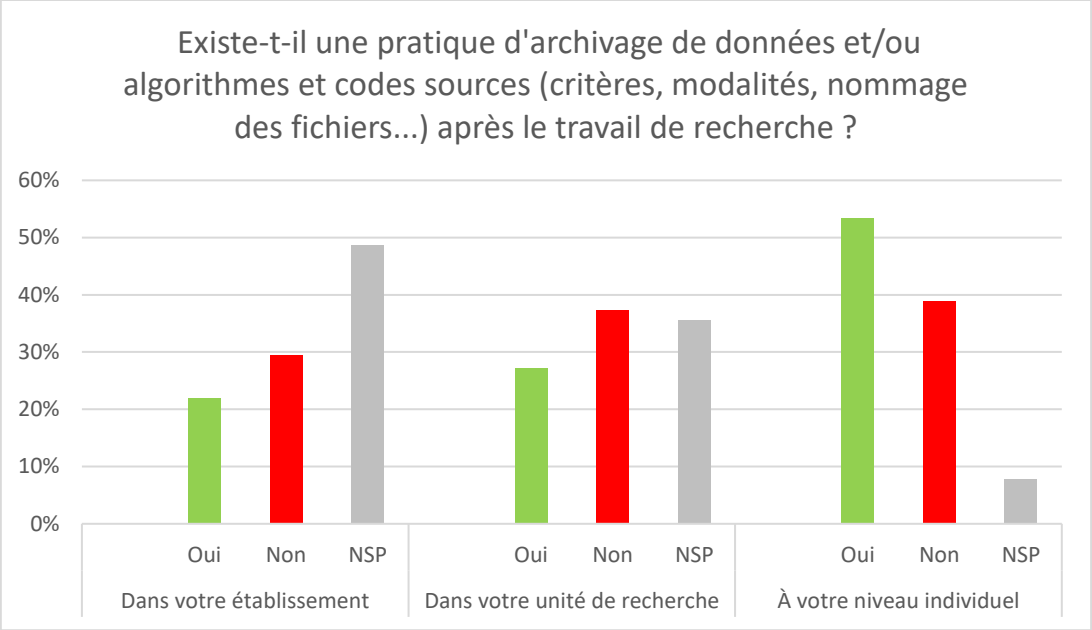

Figure 18 Archivage des données et/ou algorithmes et codes sources après le travail de recherche

Cette question avait pour objectif d'appréhender les pratiques d'archivage de données de recherche, algorithmes et codes sources à différentes échelles.

L'archivage de données de recherche, algorithmes et codes sources apparaît plus comme une pratique individuelle (53 %) que d'établissement (22 %) ou d'unité de recherche (27 %), quel que soit le profil des répondants.

Néanmoins, dans les pôles BABS et UPEE, les pratiques d'archivage au sein de l'établissement (32 % et 34 %) ou de l'unité de recherche apparaissent plus élevées (44 % pour les deux).
Le pourcentage important de réponses « Je ne sais pas » pour le niveau établissement (49 %) et le niveau unité de recherche (36 %) témoigne soit de l'absence de pratique d'archivage au niveau concerné soit de la méconnaissance des pratiques existantes. On observe globalement une meilleure connaissance des pratiques mises en place au niveau de l'unité de recherche.
Le niveau d'information sur les pratiques d'archivage varie selon le statut des répondants. Les personnels ITA/TRF/ BIATOSS en ont une meilleure connaissance que ce soit au niveau établissement (66 %) ou unité de recherche (83 %). De la même manière, les chercheurs sont plus informés de ces pratiques dans leur unité de recherche (77 %) que les enseignants-chercheurs (69 %). Les doctorants et post-doctorants répondent majoritairement qu'ils ne savent pas s'il existe des pratiques d'archivage au niveau établissement (63 %) et unité de recherche (57 %).

Lors des entretiens, la question de l'archivage pérenne des données avait été abordée. Une chercheuse du pôle MST2I a notamment souligné la nécessité d'archiver de manière pérenne les données, d'avoir une politique d'archivage systématique et d'être accompagné aux bonnes pratiques dès le doctorat.

4.3.2. Diffusion

4.3.2.1. Expérience de la diffusion

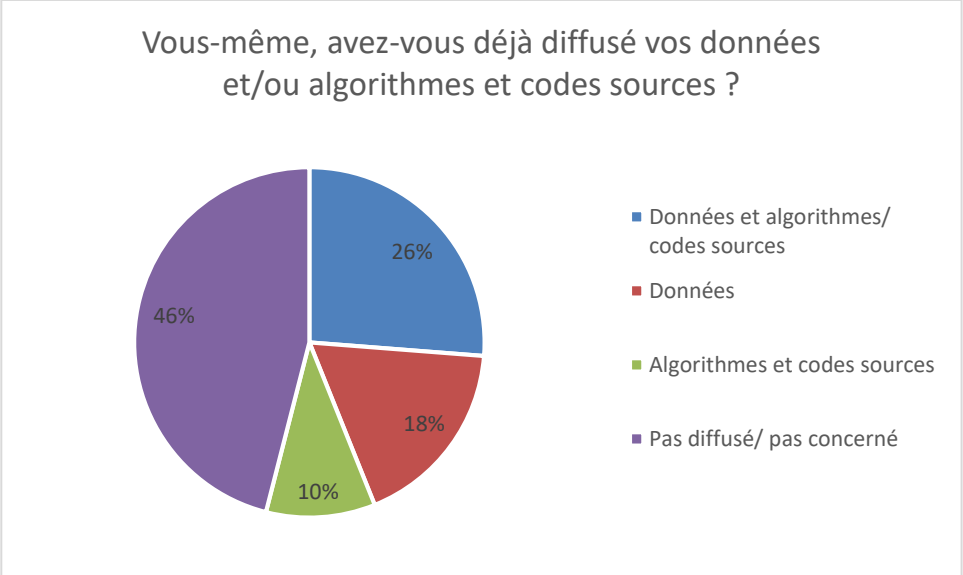

Figure 19 Diffusion des données et/ou algorithmes et codes sources

Tableau 3 Taux de diffusion par pôle

	BABS	DSPEG	H-SHS	MST2I	SdM	UPEE
Données	17,6 %	10,7 %	26,4 %	12,3 %	21,2 %	23,2 %
Algorithmes et codes sources	2,9 %	3,6 %	1,4 %	12,3 %	12,1 %	14,1 %
Données et algorithmes et codes sources	32,4 %	17,9 %	5,6 %	27,3 %	16,7 %	45,5 %

54 % des enquêtés ont déjà diffusé leurs données et/ou algorithmes et codes sources à l'issue de leur travail de recherche. Leur répartition au sein des pôles de recherche permet de dégager trois tendances :

- La diffusion est une pratique largement répandue : 83 % des répondants du pôle UPEE ont déjà diffusé leurs données et/ou algorithmes et codes sources.
- La diffusion est courante : dans les pôles BABS, MST2I et SdM, la moitié des répondants (53 %, 52 % et 50 %) ont déjà diffusé leurs données et/ou algorithmes et codes sources.
- La diffusion est peu pratiquée : dans les pôles DSPEG et H-SHS, seul un tiers des répondants (32 % et 33 %) ont déjà diffusé leurs données et/ou algorithmes et codes sources.

46 % des enquêtés ne diffusent pas ou ne sont pas concernés par la diffusion. La non-diffusion de données et/ou algorithmes et codes sources ainsi que le sentiment de ne pas être concerné par le sujet peut s'expliquer en partie par le caractère des données utilisées. Dans les pôles DSPEG, H-SHS et BABS, la majorité des répondants ont des données concernées par le RGPD[13]. Le droit du vivant a certainement aussi une incidence sur les pratiques du pôle BABS. De plus, le caractère sensible et confidentiel des données est également susceptible d'impacter

[13] Cf. 4.1.3. Données de recherche et RGPD

les pratiques de diffusion des répondants dans les autres pôles (propriété intellectuelle et industrielle, PPST, etc.).

> « Cadre d'enquête complexe qui allie des données de nature très diverse, ce qui engendre des protocoles d'anonymisation et d'ouverture des données tout aussi diversifiés donc d'autant plus chronophages [14]». (H-SHS)

L'appartenance disciplinaire peut jouer sur les pratiques individuelles. En effet, le croisement discipline/pratiques de diffusion met en lumière une culture du partage et de la diffusion des données et/ou algorithmes et codes sources dans les disciplines Informatique et Sciences du système Terre.

Aussi, la pratique de la diffusion semble contrainte par le temps que les répondants peuvent y consacrer. Un commentaire ajouté par un répondant en témoigne :

> « La diffusion prend du temps (pour le faire correctement, bien documenter, etc.) alors je ne le fais pas systématiquement mais quand [le] temps [le] permet (ce que je regrette) » (MST2I).

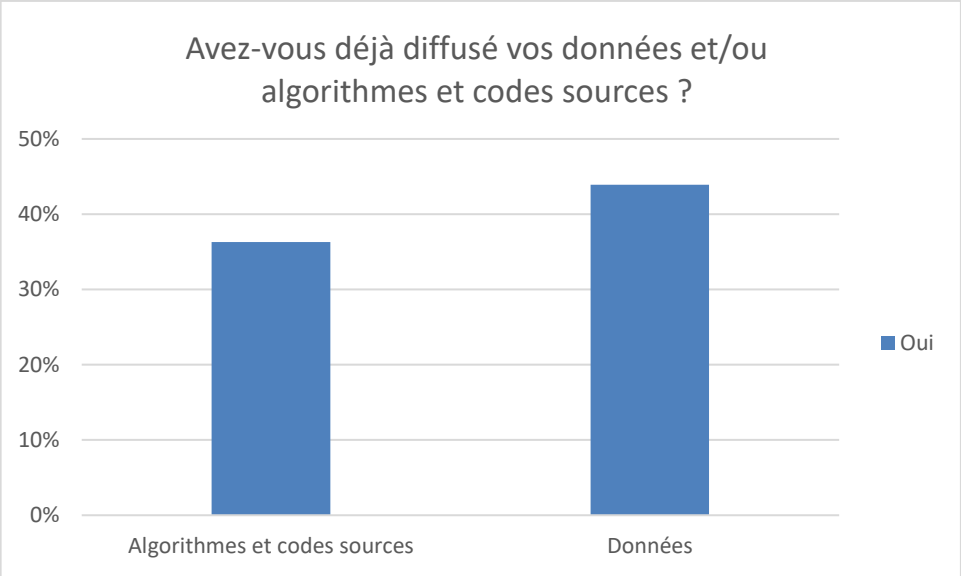

Figure 20 Pourcentage de répondants ayant déjà diffusé des données et/ou algorithmes et codes sources

44 % des répondants ont déclaré diffuser des données tandis que 36 % d'entre eux diffusent des algorithmes et codes sources. Dans le pôle MST2I, il y a un équilibre entre la diffusion de données et celle d'algorithmes et codes sources. Les répondants de ce pôle qui diffusent des données travaillent en sciences informatiques (35 %), ingénierie des produits et des procédés (33 %), mathématiques (18 %) et ingénierie des systèmes et de la communication (18 %). Pour les algorithmes et codes sources, il s'agit des mêmes disciplines mais la proportion est différente : sciences informatiques (51 %), mathématiques (24 %), ingénierie des produits et des procédés (23 %), ingénierie des systèmes et de la communication (18 %).

La diffusion de données et/ou algorithmes et codes sources est intrinsèque aux domaines de recherche des répondants. Le partage d'algorithmes et codes sources à hauteur de 21 % des répondants du pôle DSPEG s'explique par le faible taux de répondants. En conséquence, les répondants qui proviennent des disciplines économie et aménagement du territoire sont surreprésentés.

[14] Commentaire recueilli au cours de l'enquête quantitative

Qu'il s'agisse du partage de données ou d'algorithmes et codes sources, les chercheurs diffusent plus (65 % ; 53 %) que les enseignants-chercheurs (48 % ; 33 %). Les personnels ITA/TRF/BIATOSS (53 % ; 45 %) sont également impliqués dans la diffusion tandis que les doctorants et post-doctorants (22 % ; 24 %) le sont moins.

4.3.2.2. Diffusion et licence

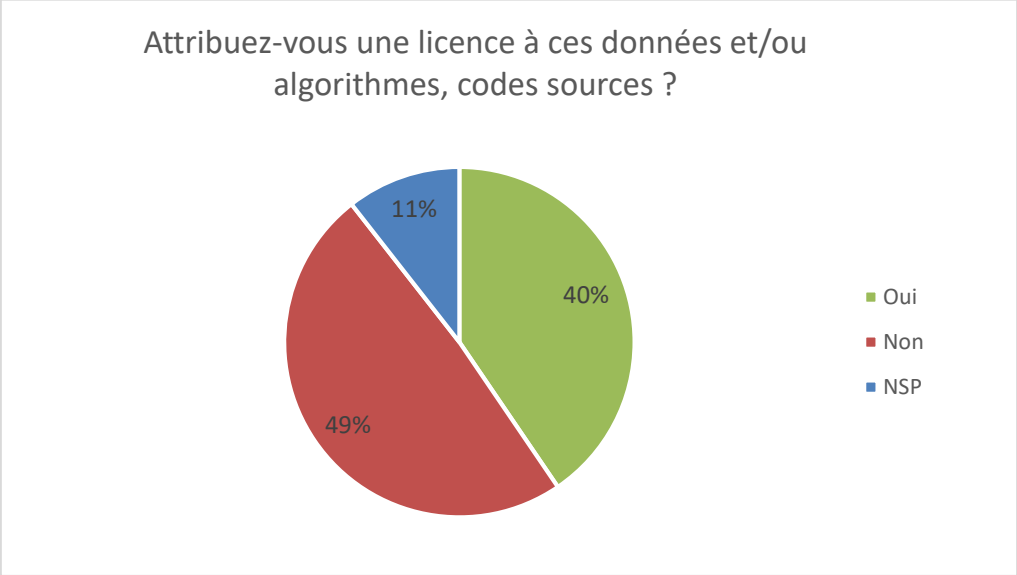

Figure 21 Attribution d'une licence aux données et/ou algorithmes et codes sources lors de la diffusion

Les données, algorithmes et/ou codes sources diffusés ne sont pas systématiquement associés à une licence. 40 % des personnes qui ont déclaré diffuser des données, algorithmes et/ou codes sources leur attribuent une licence. Cette pratique apparaît plus développée dans les pôles UPEE (49 %), MST2I (42 %) et H-SHS (42 %) que dans les pôles DSPEG (33 %), SdM (24 %) et BABS (22 %).

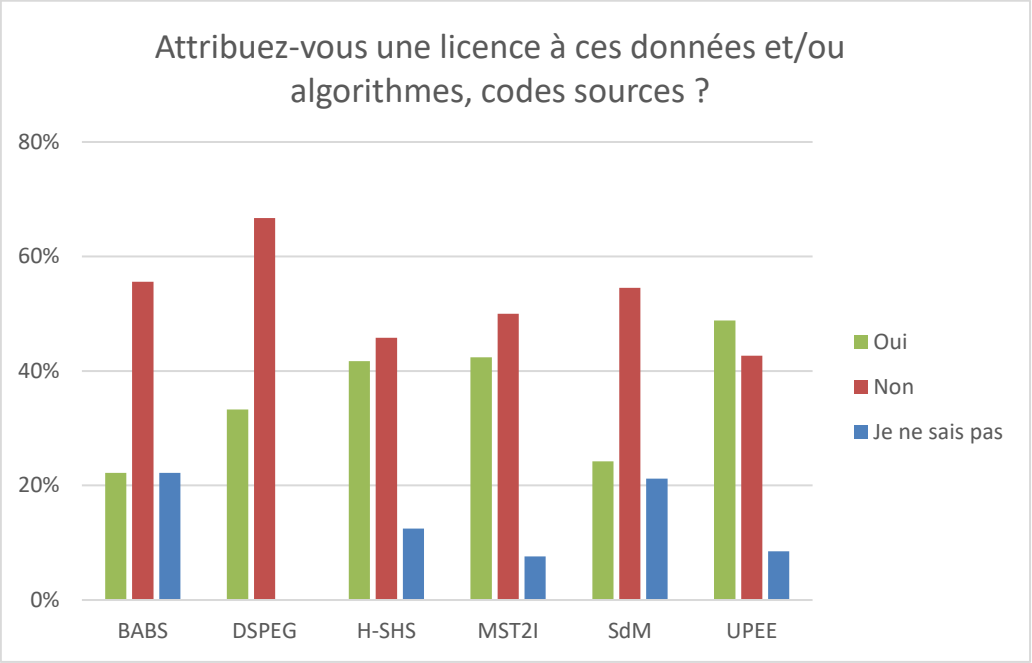

Figure 22 Attribution d'une licence aux données et/ou algorithmes et codes sources par pôle de recherche

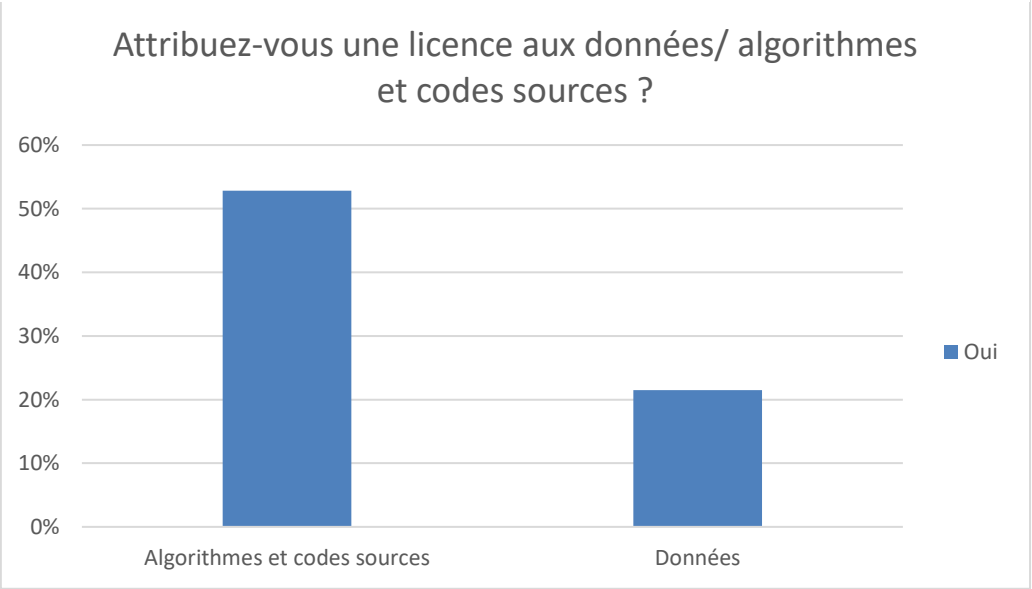

Figure 23 Pourcentage de répondants attribuant une licence lors de la diffusion de données/ algorithmes et codes sources

Attribuer une licence est une pratique plus commune chez les personnes qui diffusent des algorithmes et codes sources (53 %) que chez celles qui diffusent des données seules (22 %). Au sein du pôle UPEE, 71 % des répondants qui diffusent **seulement** des algorithmes et codes sources leur attribuent une licence contre 35 % de ceux qui diffusent **seulement** des données. On observe la même chose dans les pôles MST2I et SdM tandis que pour les pôles BABS, DSPEG et H-SHS le très faible nombre de personnes qui diffusent seulement des algorithmes ne permet pas de faire ce comparatif.

La question « Attribuez-vous une licence […] précisant les droits et obligations par rapport à la réutilisation ? » a recueilli 11 % de réponse « Je ne sais pas ». Ces répondants appartiennent principalement aux pôles BABS (22 %), SdM (21 %) et H-SHS (13 %). On observe également dans les pôles SdM et BABS une forte proportion de personnes qui n'attribuent pas de licences lors de la diffusion (55 % et 56 %).

Cette absence de pratique est probablement en lien avec différents facteurs tels que la méconnaissance des licences, de leur utilité, de la démarche pour leur application ou encore de la difficulté à en choisir une.
Lors d'un entretien mené dans le cadre de l'enquête qualitative, un participant a déclaré :

> « Nous ne sommes pas bien formés sur les notions de licence allant de pair avec l'utilisation de code open source » (MST2I)

4.3.2.3. Moyens de diffusion

Plusieurs moyens de diffusion peuvent être utilisés pour disséminer les données, algorithmes et codes sources. Le questionnaire proposait neuf réponses dont « Autre » afin que les répondants puissent préciser d'autres outils. Cette question à choix multiple témoigne de l'utilisation de plusieurs canaux pour la diffusion des données.

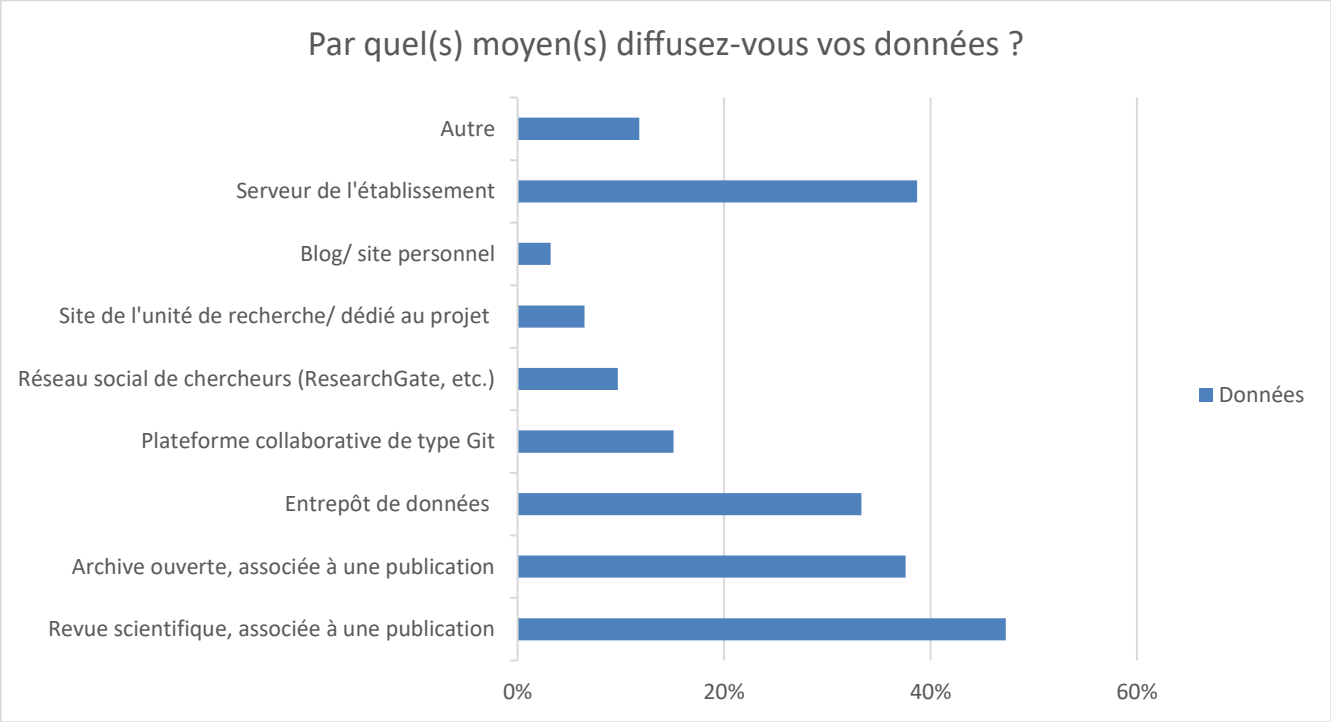

Figure 22 Moyens utilisés pour diffuser des données

Tous pôles confondus, les modes de diffusion les plus utilisés sont :

- associée à une publication dans une revue scientifique (47 %) ou dans une archive ouverte (38 %)
- le serveur de l'établissement (39 %)
- un entrepôt de données (33 %)

Associer des données à une publication est de loin le moyen le plus utilisé pour les diffuser. Néanmoins, les entretiens qualitatifs ont montré que la notion de données de la recherche n'est pas toujours bien cernée. Pour certains, les publications peuvent en faire partie [15]:

> « Alors encore une fois il faudrait définir les données de la recherche, s'il s'agit d'articles, de publications, d'algorithmes, de données analysées, de données extraites, d'information, enfin ça reste très vague pour moi » (MST2I)

[15] Propos recueillis dans le cadre de l'enquête qualitative (juillet 2021)

> « […] deux choses principalement, la première c'est sur les articles scientifiques qui sont issus de nos recherches, et la deuxième, c'est ce qu'on appelle nous dans notre domaine les traces qui sont laissées par les étudiants sur les plateformes d'apprentissage » (MST2I)

Les enseignants-chercheurs diffusent plus leurs données associées à une publication que les chercheurs que ce soit dans une revue ou dans une archive ouverte. En effet, 67 % des enseignants-chercheurs qui diffusent des données le font avec une publication dans une revue scientifique et 43 % dans une archive ouverte contre 42 % et 29 % des chercheurs qui en diffusent.

Si la diffusion dans un entrepôt de données est pratiquée par un tiers des enseignants-chercheurs, chercheurs et des doctorants et post-doctorants qui publient des données, 50 % des ITA/TRF/BIATOSS qui diffusent leurs données le font via ce canal.

Contrairement à la diffusion des données qui se fait par plusieurs moyens de manière presque équivalente, la diffusion des algorithmes et codes sources s'appuie sur un canal principal.

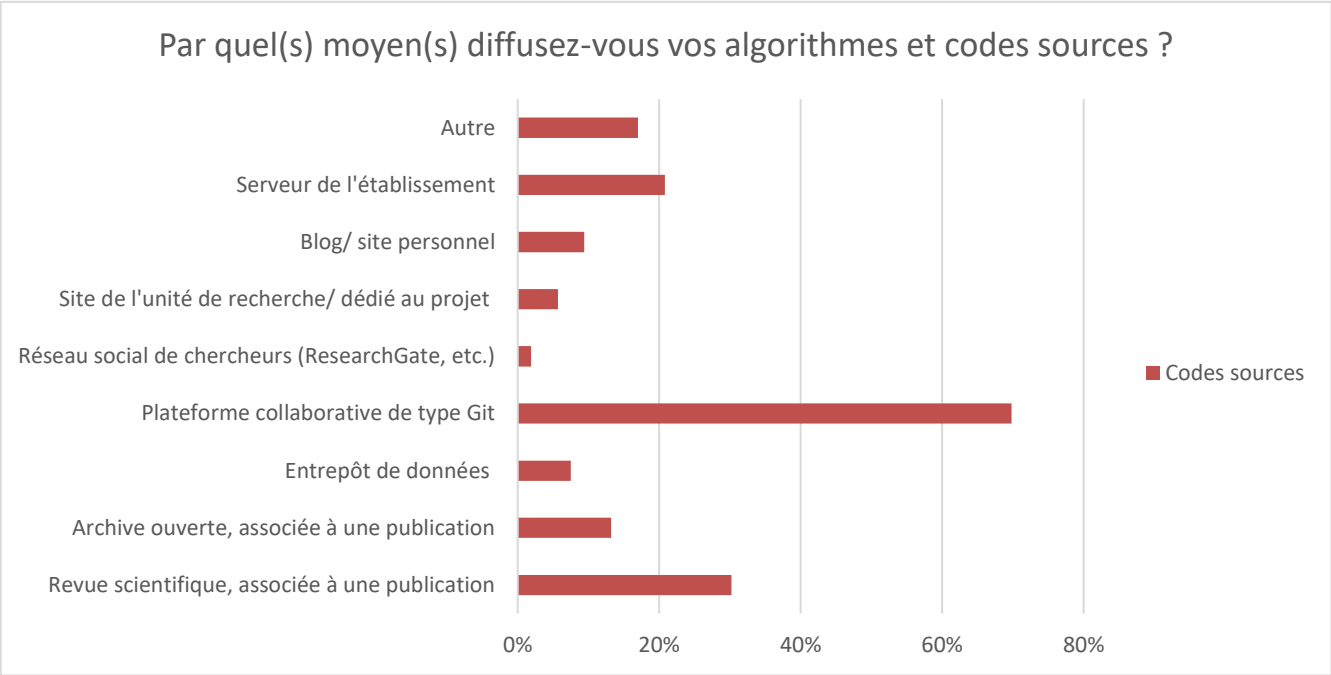

Figure 23 Moyens utilisés pour diffuser des algorithmes et codes sources

Le moyen très majoritairement utilisé pour partager des algorithmes et codes sources est une « plateforme collaborative de type Git » (70 %). D'autres modes sont également utilisés, mais de façon moins fréquente, pour diffuser les algorithmes et codes sources : associée à une publication dans des revues scientifiques (30 %) et sur le serveur de l'établissement (17 %).

14 % des enquêtés ont répondu « Autre ». Ils devaient alors obligatoirement détailler leur réponse dans un champ texte libre. Si certains ont précisé les entrepôts de données ou les plateformes Git utilisés, la majorité a indiqué des réponses telles que « envoi direct par mail » (MST2I), « sur demande » (SdM) ou encore « échange direct avec des collègues » (SdM). Cette pratique ainsi que la forte utilisation de serveur d'établissement pour la diffusion signifient que, pour une partie des participants de l'enquête, la diffusion des données, algorithmes et codes sources après le projet de recherche s'opère à plusieurs échelles de la même manière que le partage de l'accès aux données au cours du projet de recherche[16]. La diffusion après le projet de recherche n'est pas systématiquement une diffusion ouverte et formalisée. Le contact direct entre pairs est également un mode d'accès aux données.

[16] Cf. 4.2.3. Accès aux données

5. Leviers et freins par rapport à la science ouverte

Dans la troisième partie du questionnaire, nous voulions appréhender les leviers en faveur de la science ouverte ainsi que les freins rencontrés par la communauté scientifique vis-à-vis de celle-ci.

5.1 Motivations à la diffusion des données

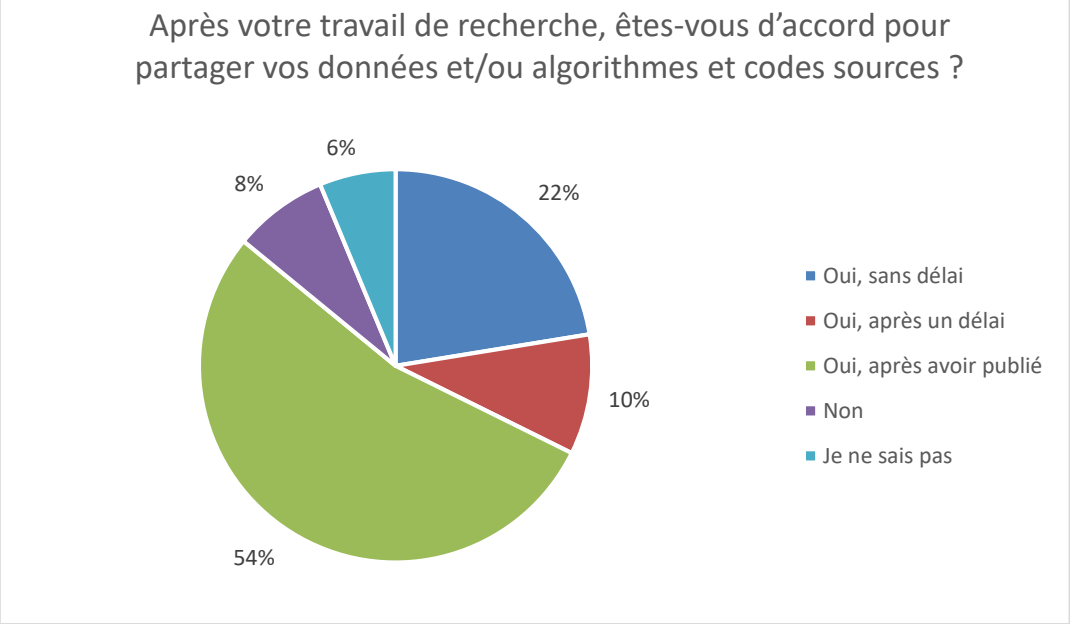

Figure 24 Adhésion au principe de partage des données et/ou algorithmes et codes selon les principes de la science ouverte

Il y a une véritable volonté de partager les données de recherches et/ou algorithmes et codes sources puisque 86 % des personnes interrogées répondent « Oui ».

Plusieurs réponses étaient possibles sur le délai d'ouverture des données. Ainsi, 22 % sont favorables pour partager sans délai dès la fin de leur travail de recherche. Le délai lié à la publication est important car 54 % des répondants sont favorables à l'ouverture de leurs données mais après celle-ci. 89 % des enseignants-chercheurs sont pour un partage et 62 % le sont après la publication. Chez les doctorants et post-doctorants, 84 % sont favorables à l'ouverture de leurs données et 51 % le sont après avoir publié leurs travaux.

Par pôle de recherche, 91 % des répondants du pôle BABS sont pour le partage de données, il n'y a aucune réponse négative. 74 % d'entre eux préfèrent attendre un délai ou la publication de leurs recherches. Les réponses sont plus contrastées dans les autres pôles. Par exemple, en SdM, 23 % des personnes interrogées ne se prononcent pas ou ne souhaitent pas partager leurs données. A contrario, 73 % souhaitent partager après un délai ou après publication.

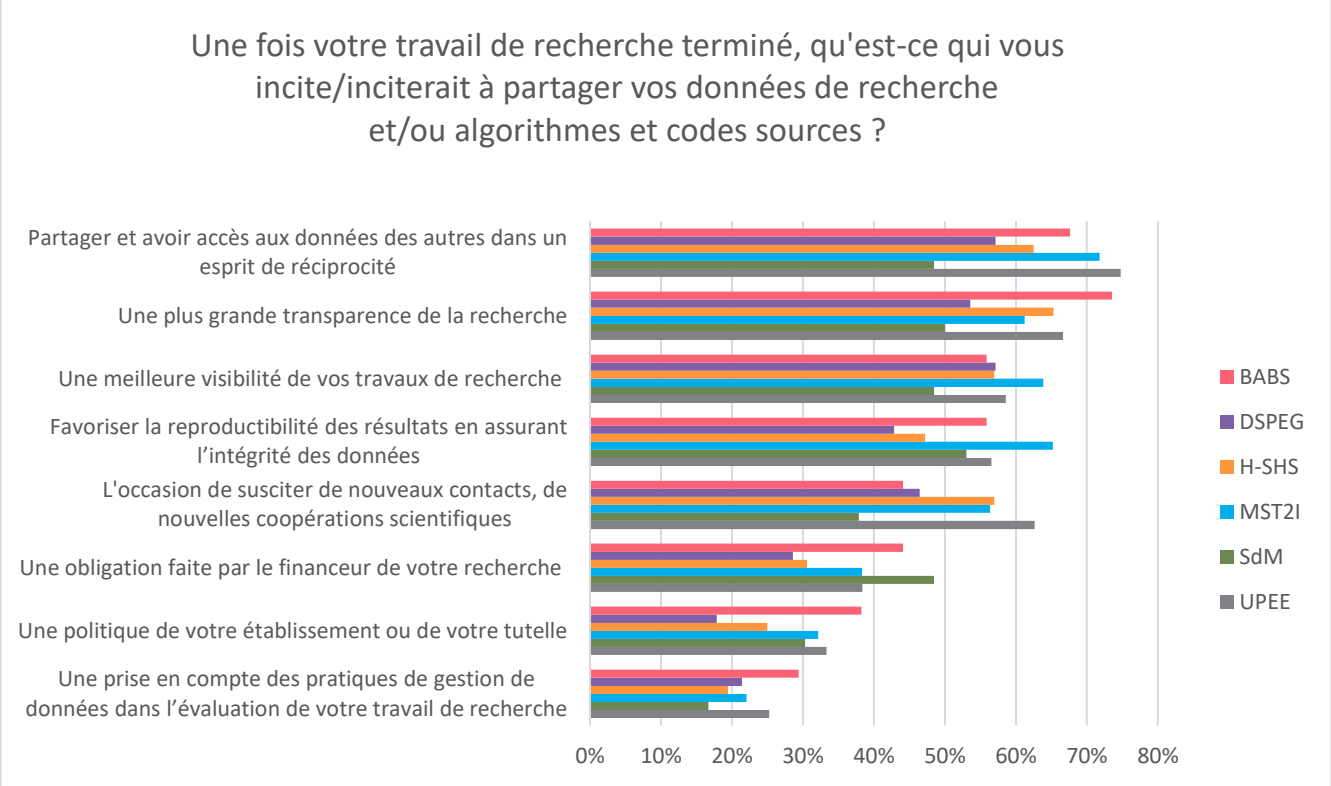

Figure 25 Motivations au partage des données et/ou algorithmes et codes sources

De manière générale, 67 % d'entre eux partagent/partageraient leurs données de recherche afin d'avoir accès eux aussi aux données des autres chercheurs dans un soucis de réciprocité. Une plus grande transparence de la recherche pour la science et la société ainsi qu'une meilleure visibilité des travaux de ces chercheurs sont les réponses sélectionnées en deuxième et troisième choix (62 % et 59 %). Plus de la moitié d'entre eux souhaitent également partager afin de favoriser la reproductibilité des résultats en assurant l'intégrité des données et de susciter de nouveaux contacts et coopérations scientifiques (58 % et 54%).

Les répondants de BABS se distinguent des autres pôles en souhaitant une plus grande transparence de la recherche pour la science et la société (74 %). À l'inverse, la prise en compte des pratiques de gestion de données de la recherche dans l'évaluation du travail de recherche inciterait moins les répondants des pôles BABS, H-SHS, MST2I, SdM et UPEE à partager leurs données de recherches. Pour le pôle DSPEG, c'est la politique de leur établissement ou tutelle qui recueille le moins de suffrages.

5.2 Freins au partage des données

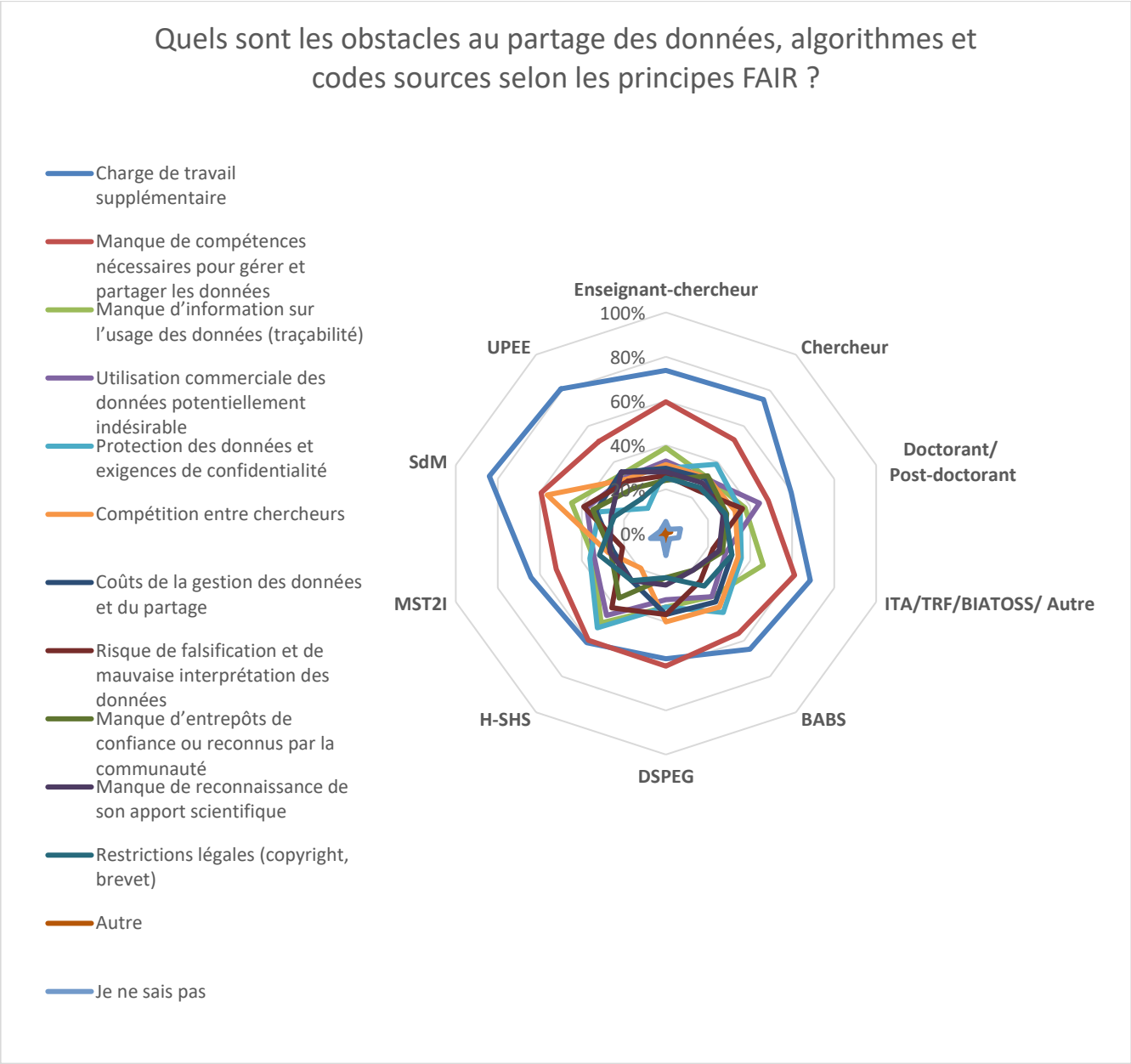

Figure 26 Obstacles au partage des données, algorithmes et codes sources selon les principes FAIR

Si l'enquête qualitative soulignait une position des participants globalement favorable au partage des données sur le principe, elle faisait aussi apparaître de fortes réserves quant à son application sur le terrain. Les entretiens révélaient des réticences liées aux tâches supplémentaires de gestion de données jugées inutiles et pas du ressort des chercheurs pour beaucoup d'entre eux. Ces entretiens témoignaient aussi de la crainte vis-à-vis d'un manque de garde-fous dans un environnement de compétition, et de plagiat. Ce fort niveau de défiance se confirme à travers les résultats de l'enquête quantitative, les freins les plus importants recueillant jusqu'à 80 % de réponses par pôle[17].

[17] Sauf mention contraire, l'ensemble des citations ci-dessous est issu des commentaires recueillis dans l'enquête quantitative.

Concrètement, le partage des données, algorithmes et codes sources inspire des inquiétudes quant à la charge de travail supplémentaire.

Cela concerne une large majorité de répondants (70 %) quel que soit leur pôle d'appartenance ou leur statut. Les pourcentages oscillent entre 84 % pour le pôle SdM et 56 % pour le pôle DSPEG. Les répondants confrontés à la gestion des données sont particulièrement sensibles à cette question, avec une augmentation à mesure que l'on avance dans la carrière (60 % des 20-29 ans contre 76 et 72 % des 40-49 ans et des 50 et plus). La gestion des données selon les principes FAIR est vécue comme une contrainte supplémentaire, chronophage sur du temps de recherche :

> « L'obstacle majeur est le temps nécessaire à -mettre en forme les données ; -remplir les métadonnées ; -les publier (sur des plateformes que nous ne connaissons souvent pas du tout) il peut manquer de l'incitation (non-réciprocité et non-valorisation) et potentiellement un manque d'aide à se former à tout un tas d'outils différents […] » (BABS)

Un consensus s'établit autour du manque de compétences nécessaires pour gérer et partager les données.

C'est le deuxième frein cité : il dépasse la valeur de 50 % dans tous les pôles et culmine à 60 % pour le pôle DSPEG, 59 % dans les pôles H-SHS et SdM, et 55 % pour le pôle BABS. Toutes les catégories se sentent concernées, et particulièrement les personnels ITA/TRF/BIATOSS (61 %), les enseignants-chercheurs (59 %), les chercheurs (52 %) et enfin les doctorants et post-doctorants (48 %). Ces nouvelles compétences impactent les métiers de la recherche et nécessitent de former aux nouvelles tâches les jeunes arrivants en amont et en aval d'accompagner le changement au cours de la carrière.

> « je trouve qu'il existe plein de guides et directives pour les bonnes pratiques pour la gestion de données, mais tous manquent d'exemple pratique, d'outils et d'exemple d'utilisation : les chercheurs n'étant pas sachant, il y a une certaine improvisation qui au final ne répond pas aux besoins réels tout en complexifiant le travail. » (MST2I)

Le manque de personnel d'appui à la recherche dédié à la gestion des données est cité de façon récurrente dans les commentaires de l'enquête.

Les répondants questionnent aussi la qualité et la lisibilité de ces données, algorithmes et codes sources pour des raisons multiples (formats propriétaires, contraintes liées au recueil et au traitement des données, etc.) et donc l'intérêt de passer du temps pour préparer le partage.

> « Pas tous les codes sources/algos sont d'une qualité qui mérite une publication ouverte. C'est de même pour les données. Avoir tout ouvert dans un esprit de transparence, c'est une cible noble. Mais il ne faut pas non plus polluer l'internet/la sphère académique-scientifique avec des données seulement parce qu'on se sent mieux quand c'est publiquement accessible. Il y a déjà un tas de données et codes de qualité variable sur l'internet et cela mène principalement au constat que beaucoup de monde (des jeunes ou pas) sont perdus dans le choix des données et algos. La traçabilité et de la documentation pour assurer et prouver un certain niveau de qualité est donc indispensable. Sinon c'est "garbage in, garbage out" - souvent dit dans l'informatique... » (MST2I)

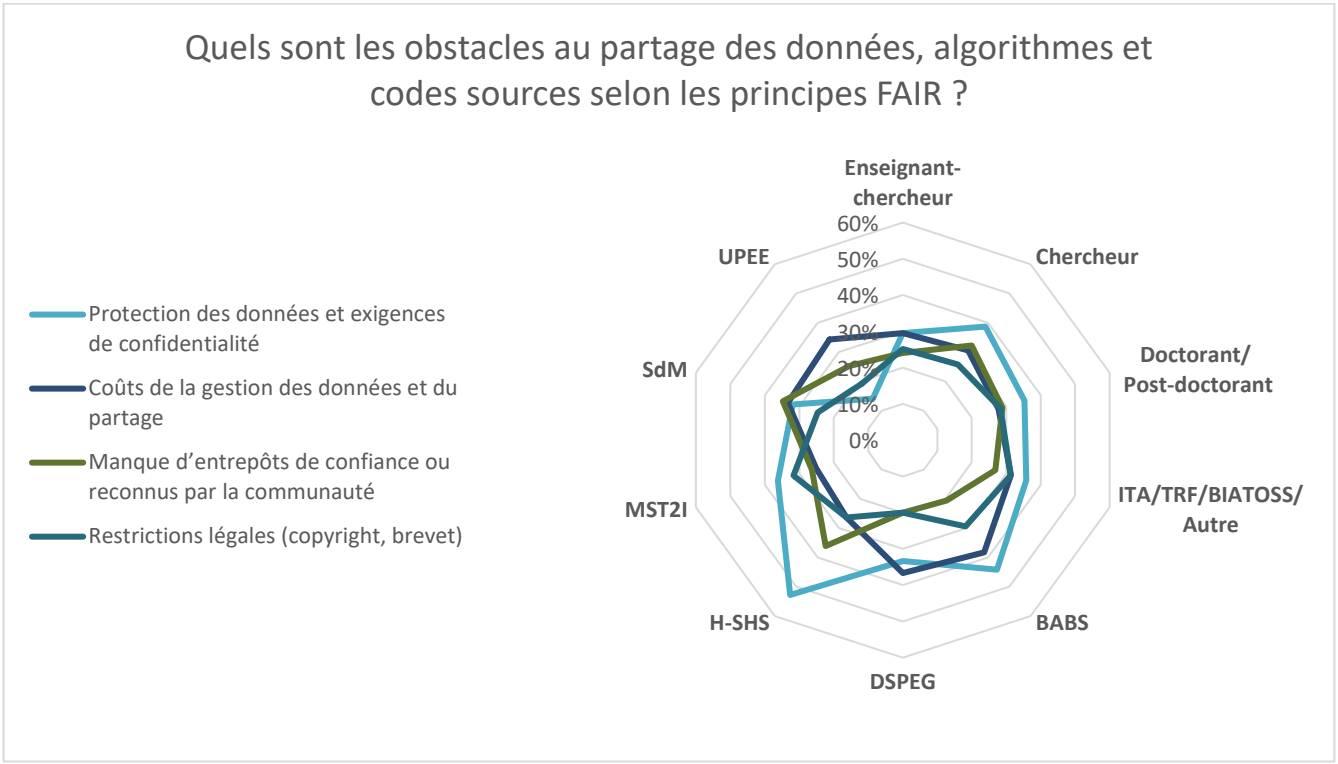

Figure 27 Obstacles au partage des données, algorithmes et codes sources selon les principes FAIR (2)

En revanche, on observe des particularités en matière de freins selon le contexte disciplinaire. La protection des données et l'exigence de confidentialité est une problématique pour la majorité des répondants des pôles H-SHS (53 %) et BABS (44 %). Par ailleurs, les restrictions légales (copyright, brevet) constituent un obstacle au partage des données en particulier pour les répondants des pôles MST2I (32 %) et BABS (29 %).

> « Beaucoup de mes données ne sont simplement pas partageables de manière publique parce que confidentielles. Je ne sais jamais comment faire un accord de confidentialité quand je veux partager ces données. » (H-SHS)

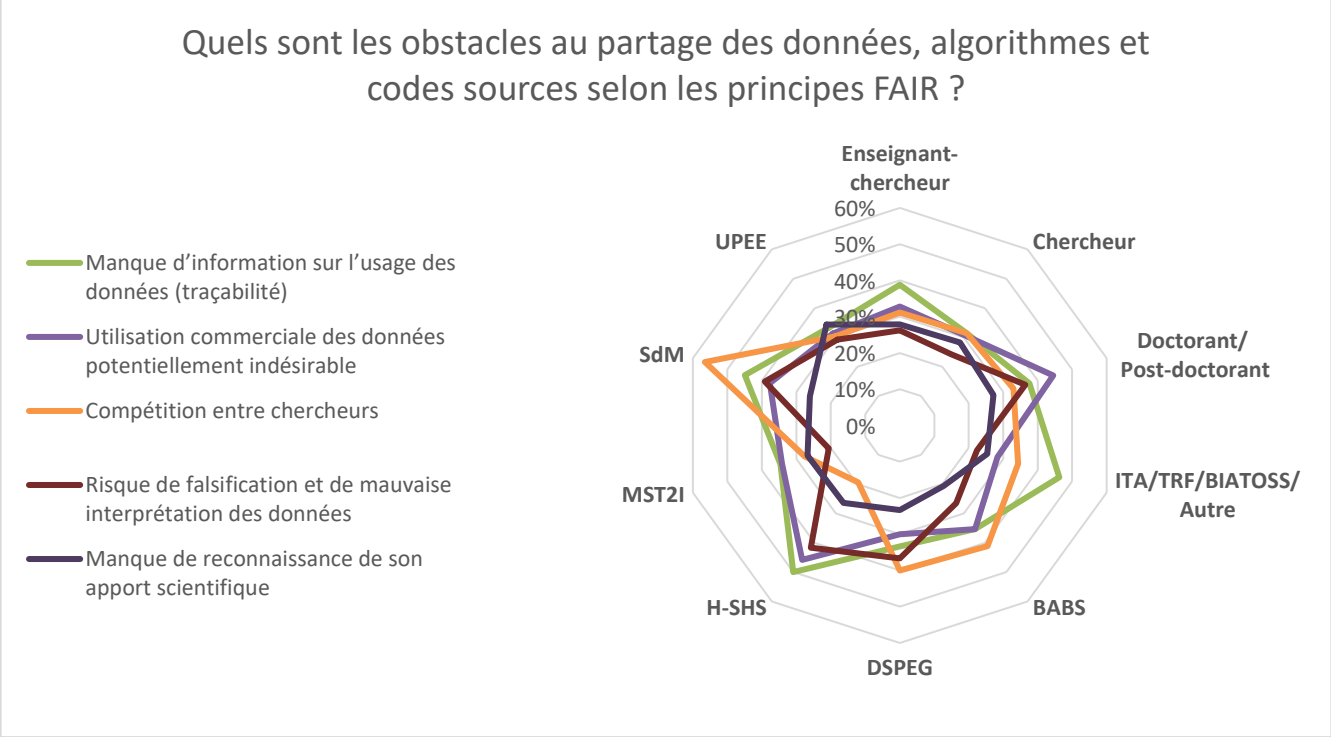

Figure 28 Obstacles au partage des données selon les principes FAIR (3)

La perspective d'ouverture des données, en particulier lorsqu'elle est précoce, est perçue comme un danger car elle exacerbe une compétition déjà sensible dans certaines disciplines.

La concurrence entre chercheurs est une préoccupation importante pour les répondants des pôles SdM (57 %), BABS (41 %) et DSPEG (40 %). L'ouverture des données, algorithmes et codes sources pose la question de la temporalité de l'exploitation des données et de leur partage.

> « […] le temps court des contrats n'est pas le temps (long) de la recherche. La recherche et l'exploitation des résultats n'est pas généralement achevé à la fin d'un contrat financé. Dans mon domaine, on peut exploiter les données produites à un moment jusqu'à plusieurs années après. Mettre ses données en accès libre sans restriction, c'est aussi se dépouiller de son travail et de sa propriété intellectuelle […] » (SdM)

> « - Période d'embargo le temps de valoriser les données qui peut être longue…» (UPEE)

Une autre inquiétude concerne le devenir des données avec deux freins proches : le manque d'information sur l'usage des données (traçabilité) et une utilisation commerciale des données potentiellement indésirable.

Ces freins touchent davantage les répondants des pôles H-SHS (50 %, 48 %) et SdM (45 %, 38 %). Les personnels ITA/TRF/BIATOSS soulignent plus l'écueil de la traçabilité (46 %) que les enseignants-chercheurs (39 %), les doctorants et post-doctorants (38 %) et les chercheurs (32 %).

> « …Important d'avoir l'assurance de traçabilité (qui fait quoi avec ces données) et donc de reconnaissance » (UPEE)

> « le problème est bien celui de l'absence de citation à terme de l'auteur même des données. Une plateforme sera citée, et les scientifiques qui ont acquis ces données n'existeront plus […] Des données doivent être situées dans des journaux de références et non pas diluées sur le web ». (H-SHS)

De façon plus large, la crainte de perdre la maîtrise de ses données s'exprime d'autant qu'elle s'accompagne d'un changement de culture. Le passage d'un modèle d'échange entre pairs à un modèle de dépôt et de partage des données est perçu comme contraignant, sans offrir de réelles garanties pour la réutilisation des données.

> « En effet, je pense qu'il est déjà assez simple de se procurer les données qui sont en lien à une publication… il suffit de demander à l'équipe qui a produit la publication. De là naît une véritable interaction entre le demandeur et les auteurs de l'article en question. En mettant en place une plateforme impersonnelle d'accès aux données provenant d'expériences diverses et variées, on perd ce lien et on perd la remise en contexte possible de l'acquisition de ces données. […] » (SdM)

À cette inquiétude s'ajoute celle du manque de reconnaissance de son apport scientifique. Les enquêtés du pôle UPEE (34 %) s'en préoccupent davantage que ceux des autres pôles (21 % à 27 %).

Les répondants semblent opposer le principe de réalité à la philosophie de l'ouverture en pointant les risques de la perte de contrôle sur leurs données :

- Le risque de falsification et de mauvaise interprétation des données est un obstacle constaté particulièrement pour les pôles H-SHS (42 %), SdM (39 %) et DSPEG (37 %).
- Le manque d'entrepôts de confiance ou reconnus par la communauté est un frein pour les répondants des pôles H-SHS et SdM (36 % et 35 %). Il en ressort un manque de visibilité ou de connaissance sur le fonctionnement de cet écosystème d'entrepôts.

Par ailleurs, cette enquête révèle une inquiétude sur les coûts de la gestion des données et du partage puisque 29 % des répondants (tous pôles confondus) considèrent cela comme un frein.

Ainsi le partage et la gestion des données induisent différents freins ou obstacles :

> « Donc c'est un problème humain là-derrière un vrai problème humain de compétence, d'envie, de temps, en même temps que matériel que système, mais matériellement le chercheur dans [domaine de recherche] les gens sont pas formés pour faire ça [18]» (SdM)

[18] Propos recueilli lors de l'enquête qualitative (juillet 2021)

6. Besoins et attentes

Cette dernière partie porte sur les besoins et les attentes des enquêtés en termes d'aide et d'accompagnement à la gestion des données.

6.1 Besoins

Parmi une liste de réponses prédéterminées (16), les participants devaient sélectionner celles correspondant à leurs besoins. S'ils répondaient Autre, ils devaient préciser leur attente dans un champ texte libre. 95 % des enquêtés ont déclaré avoir au moins un besoin[19].

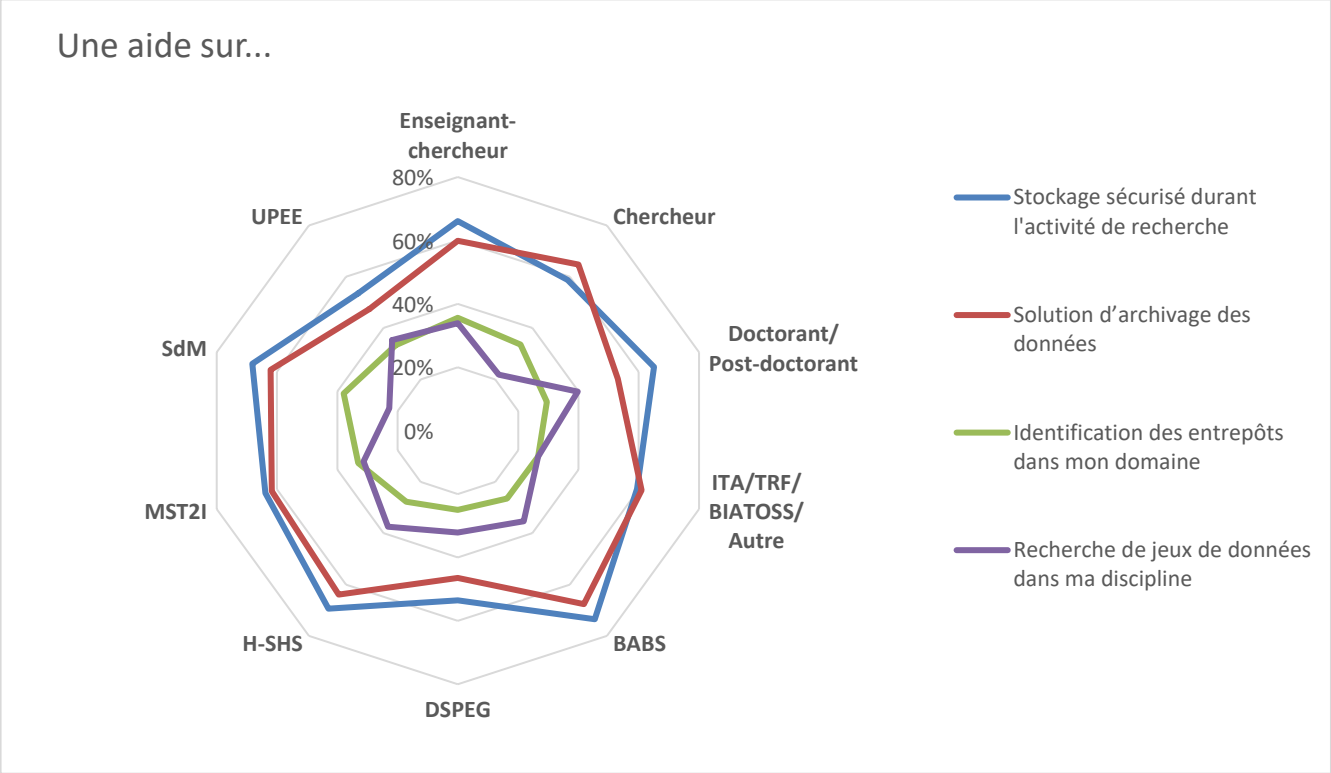

Figure 29 Eléments qui correspondent aux besoins selon le pôle et le profil

Les besoins prédominants pour les répondants sont de disposer d'un espace de stockage sécurisé durant l'activité de recherche (63%), mais aussi d'un espace d'archivage sécurisé et fiable pour entreposer des données après la fin du projet (59 %).

Cela témoigne en creux d'un manque de solutions de stockage ou d'archivage adaptées aux données des chercheurs ou d'un manque de connaissances sur les solutions existantes. Ce besoin est plus marqué dans les pôles BABS, H-SHS, MST2I et SdM.

Les répondants ont parfois précisé leurs attentes en termes de stockage [20]:

> « Notre laboratoire a un fort besoin de serveurs qui sont accessibles pour des projets communs entre différents membres d'équipes du laboratoire. […] Il est très important de

[19] Ces besoins ont été classés par ordre décroissant selon le pourcentage de réponses obtenu pour chaque pôle (cf. annexe, Tableau 6 Classement des besoins par pôle de recherche en fonction du % de réponses obtenues (question « Sélectionnez les éléments qui correspondent à vos besoins »).

[20] Les citations ci-dessous sont extraites des commentaires libres de l'enquête quantitative.

pouvoir dans l'avenir avoir ces volumes (*sic*) collaboratifs avec des pratiques de sauvegarde. » (SdM)

« Le fait qu'il s'agisse de données de santé est une spécificité à considérer car elle implique des méthodes de stockage et de partage spécifiques » (BABS)

Ainsi, une aide sur le stockage durant l'activité de recherche doit tenir compte des spécificités disciplinaires, de la sensibilité des données et des aspects collaboratifs.

La question de pérennité de la solution d'archivage a été soulevée par un des enquêtés (BABS). L'aide fournie doit prendre en compte l'évolutivité des outils et supports garantissant ainsi l'accès pérenne aux données.

Une aide pour rechercher des jeux de données existants est un besoin secondaire exprimé par 32 % des répondants. Cette aide est plus sollicitée dans le pôle H-SHS (38 %) et par les doctorants et post-doctorants (40 %).

« Le soutien dont ont besoin les utilisateurs c'est comment retrouver, comment faire une requête, pour pouvoir retrouver des données qui nous intéressent. Donc avoir des interfaces personne-machine, des IHM qui sont très ergonomiques et simples d'utilisation pour des non-experts comme sait très bien le faire google, Google est utilisé parce qu'on tape son texte, on appuie sur entrée et ça marche quoi. [21]» (MST2I)

Pour rechercher des jeux de données existants, la communauté scientifique a besoin de savoir où chercher mais également de disposer d'un outil de recherche avec des fonctionnalités de plus en plus riches.

[21] Propos recueilli lors de l'enquête qualitative (juillet 2021)

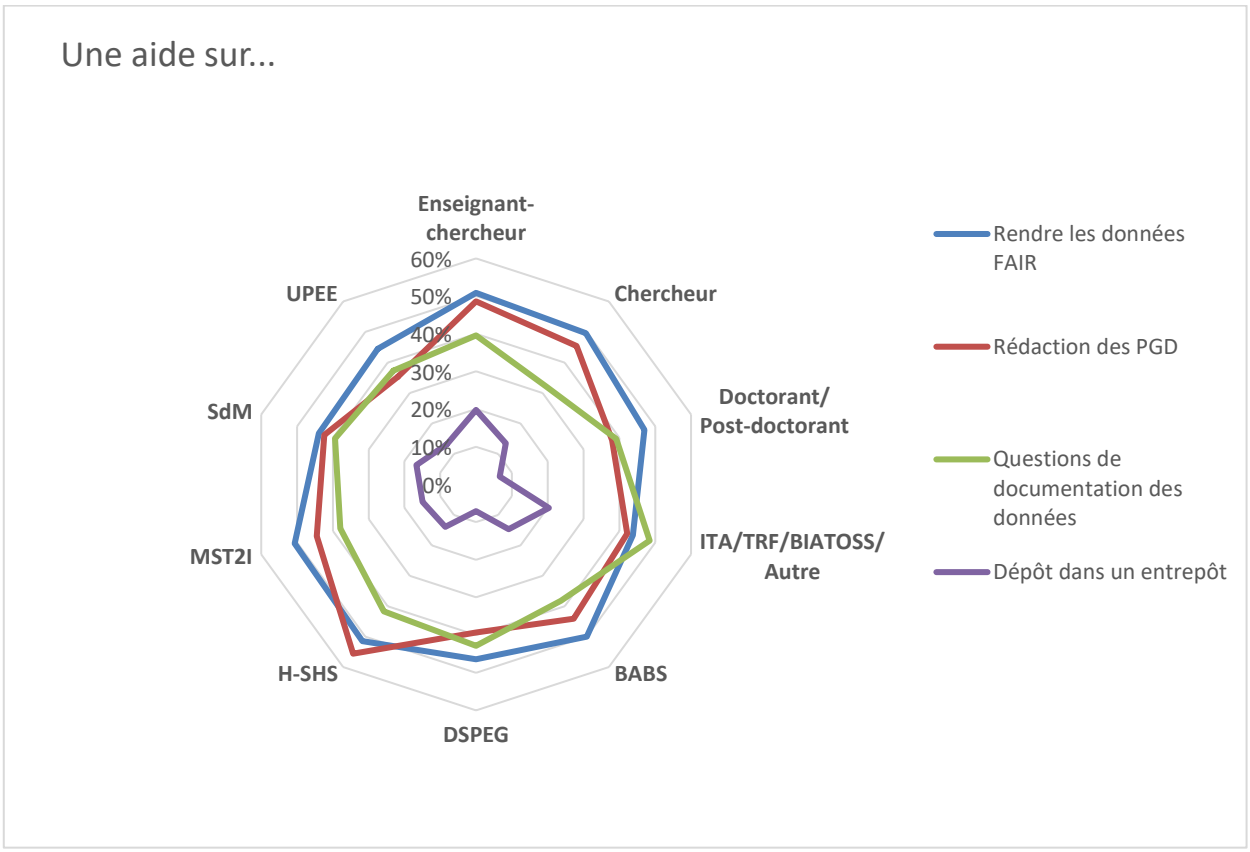

Figure 30 Eléments qui correspondent aux besoins selon le pôle et le profil (2)

45 % des répondants indiquent avoir besoin d'une aide à la rédaction du plan de gestion de données. Parmi eux, la moitié n'en a jamais rédigé.

Ces résultats sont à mettre en lien avec une proportion encore assez faible de chercheurs concernés par des projets financés avec une obligation de PGD. Un appui sur ce sujet est sollicité par 56 % des enquêtés du pôle H-SHS. Concernant le statut, l'aide à la rédaction d'un PGD est davantage demandée par les enseignants-chercheurs que par les doctorants ou les ingénieurs.

49 % des enquêtés souhaitent recevoir une aide pour rendre les données FAIR. Cette aide est plus sollicitée par les enseignants-chercheurs, chercheurs, doctorants et post-doctorants que par les personnels ITA/TRF/BIATOSS. Ce besoin démontre la sensibilité des chercheurs à la réutilisation et à l'accessibilité de leurs données.

D'autre part, une aide sur la documentation des données (description des jeux de données, métadonnées, nommage des fichiers) intéresse 39 % des enquêtés. Parmi les personnels ITA/TRF/BIATOSS, qui souvent génèrent les données brutes et réalisent les manipulations sur les différents appareils à leur disposition, 48 % disent vouloir un appui sur ce sujet.

> « L'étiquetage des données, ça je pense que ça peut être fait par des gens qui vont être dédiés à ça, des gens qui savent manier les outils des *métadatas*, tous les *repository*, télécharger les documents dans le *repository* et les estampiller de *métadatas* [22]» (SdM)

Tous pôles confondus, 14 % des répondants ont déclaré avoir besoin d'une aide pour déposer des données dans un entrepôt. 20 % des personnels ITA/TRF/BIATOSS et des enseignants-chercheurs déclarent en avoir besoin. Il s'agit de la thématique sur laquelle une aide est la moins sollicitée. Cette aide intéresse en majorité ceux qui

[22] Propos recueilli lors de l'enquête qualitative (juillet 2021)

diffusent déjà des données et/ou algorithmes et codes sources (57 %). Seulement 15 % des répondants diffusent via un entrepôt de données. Il peut s'agir d'un besoin latent.

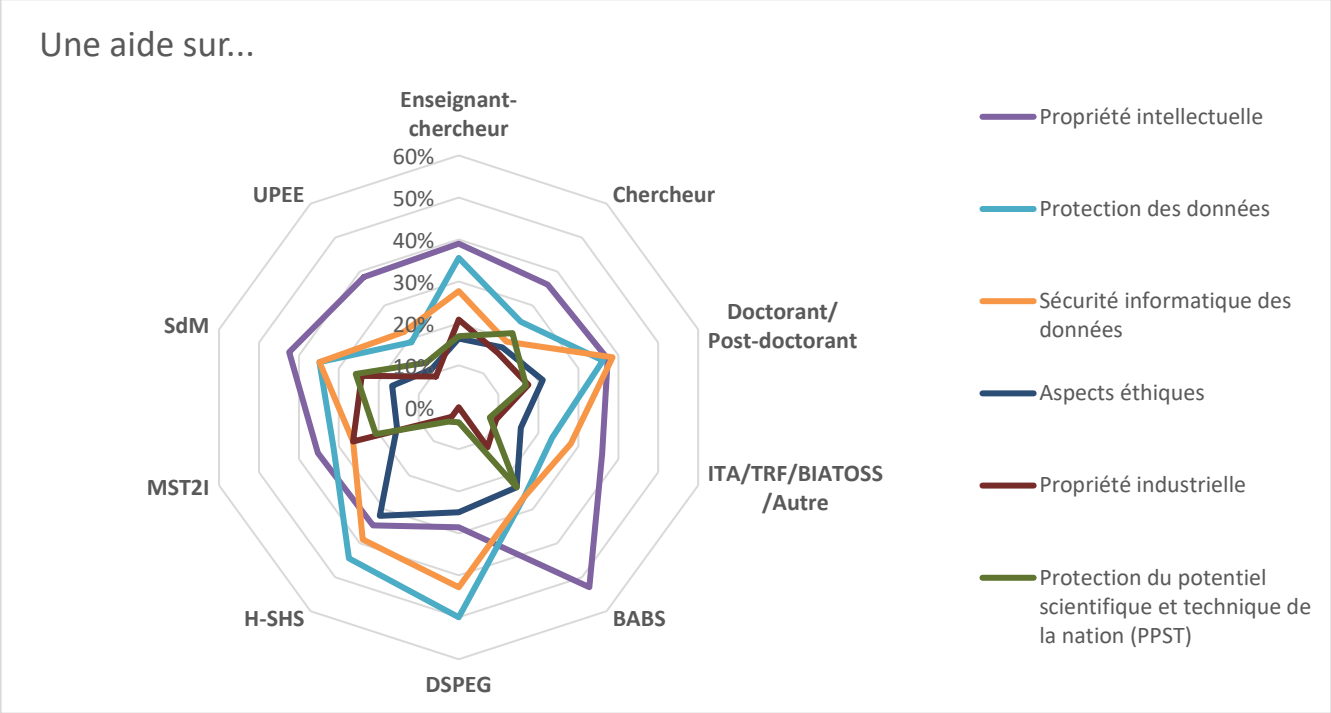

Figure 31 Eléments qui correspondent aux besoins selon le pôle et le profil (3)

Contrairement aux thématiques précédentes sur lesquelles tous les pôles ont déclaré avoir des besoins relativement équivalents, une aide sur les aspects juridiques et éthiques apparaît davantage conditionnée par l'appartenance à un pôle :

- Pôle Biologie, Agronomie, Biotechnologie, Santé : la majorité des répondants attend des conseils sur les aspects de propriété intellectuelle (53 %). De plus, 24 à 27 % des enquêtés ont besoin d'une aide sur les questions de sécurité informatique des données, de protection des données, les aspects éthiques et le PPST.

- Pôle Droit, Science Politique, Economie, Gestion : la majorité des répondants déclare avoir des besoins sur les aspects de protection des données (50 %). Viennent ensuite les aspects de sécurité informatique des données (43 %), de propriété intellectuelle (29 %) et éthiques (25 %).

- Pôle Humanités, Sciences Humaines et Sociétés : les répondants ont besoin d'aide sur les aspects de protection des données (44 %), de sécurité informatique des données (39 %), les questions de propriété intellectuelle (35 %) et éthiques (32 %).

- Pôle Mathématiques, Sciences et Technologies de l'Information et de l'Ingénierie : les répondants ont besoin d'aide sur les aspects de propriété intellectuelle, de protection des données ainsi que sur leur sécurité informatique. Viennent ensuite sur les questions de propriété industrielle (26 %) et le PPST (21 %).

- Pôle Sciences de la Matière : 42 % des répondants déclarent avoir besoin d'une aide sur les questions de propriété intellectuelle, 35 % sur la protection des données et leur sécurité informatique tandis que 24 à 26 % veulent de l'aide sur la PPST et la propriété industrielle.

- Pôle Univers, Planète, Espace, Environnement : 38 % des répondants ont besoin d'une aide sur les questions de propriété intellectuelle. Viennent ensuite la protection des données et leur sécurité informatique (22 % à 19 %) ainsi que les questions sur la PPST, aspects éthiques et propriété industrielle (9 à 13 %).

De façon globale, les aspects éthiques, les questions de propriété industrielle ou sur la protection du PPST sont moins souvent cités par les répondants (moins de 20 %). En effet, ce sont des notions qui ne concernent pas tous les domaines de recherche.

6.2 Forme de l'aide et niveau de priorité

Après avoir sélectionné les sujets sur lesquels ils avaient besoin d'aide, les participants de l'enquête devaient indiquer sous quelle forme cette aide pouvait leur être apportée et avec quel niveau de priorité.

Le questionnaire proposait cinq formes d'aide :

- des informations générales
- des ateliers ou séminaires techniques
- un accompagnement individualisé
- du personnel dédié à la gestion des données de la recherche
- une automatisation des tâches

Pour chacune, les enquêtés devaient indiquer un degré de priorité (en priorité, important, éventuellement, pas nécessaire).

Les réponses ont été regroupées en deux ensembles : d'une part une catégorie « nécessaire » qui comprend les réponses « en priorité », « important », « éventuellement » et d'autre part une catégorie « pas nécessaire ». On constate que toutes les formes d'aide sont jugées utile par une large proportion de répondants **:**

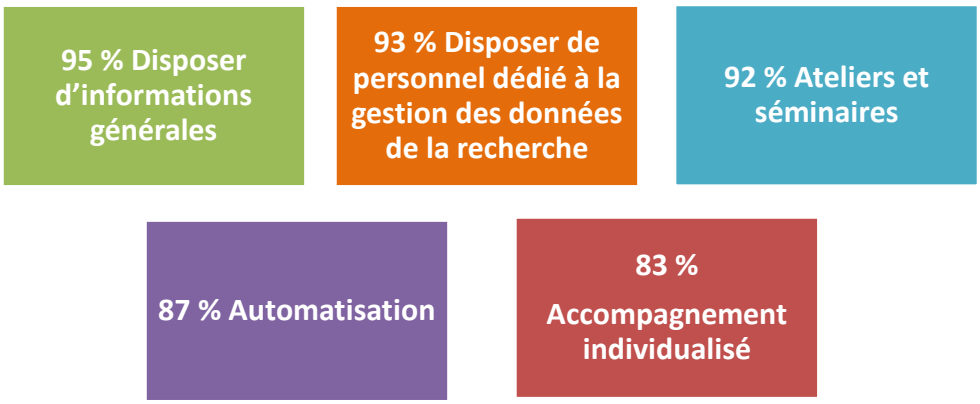

Figure 32 Aides identifiées comme nécessaires

Cependant, pour analyser les résultats, il convient de se concentrer sur les niveaux « en priorité » et « important » afin de cerner les principales attentes des répondants. Ces derniers ont déclaré prioritaire ou importante les formes d'aide prédéterminées suivantes :

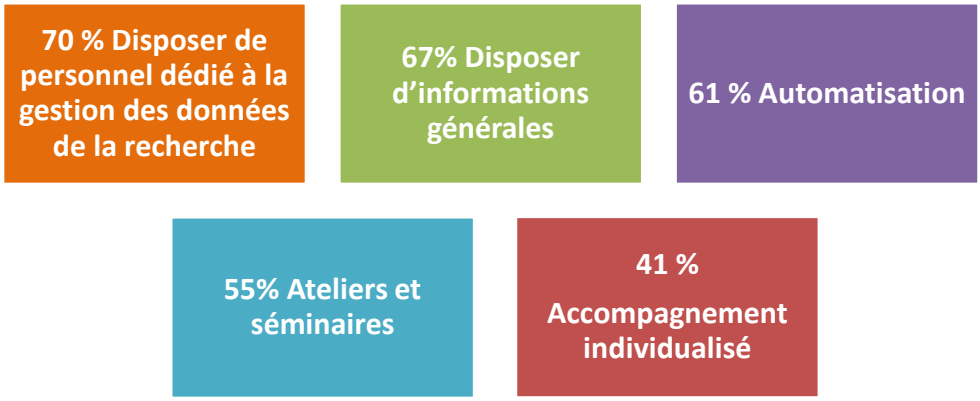

Figure 33 Aides identifiées comme prioritaires ou importantes

Pour quatre formes d'aide sur cinq, plus de la majorité des répondants ont déclaré un niveau de priorité « en priorité » ou « important ».

Du personnel dédié à la gestion des données

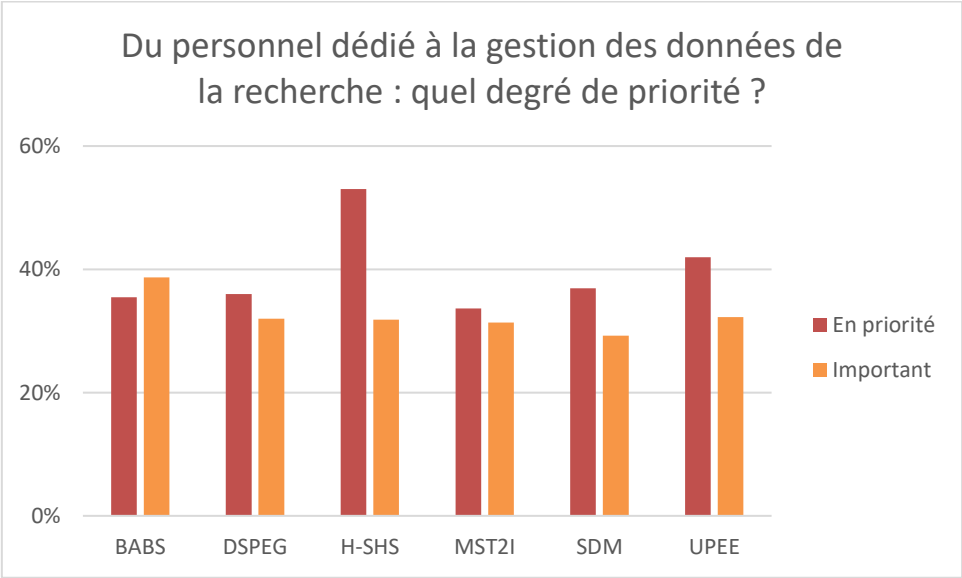

Figure 34 Degré de priorité pour du « Personnel dédié à la gestion des données de la recherche »

La grande majorité des répondants (70 %) considère que « disposer de personnel dédié à la gestion des données de la recherche » est prioritaire ou important. Parmi les formes d'aide proposées, c'est celle qui a recueilli le plus de réponses « Prioritaire » dans chaque pôle. En H-SHS, 53 % des enquêtés la considèrent comme prioritaire tandis que 32 % lui attribuent le niveau « Important ».

Les enseignants-chercheurs (45 %) et les chercheurs (44 %) déclarent davantage ce besoin comme prioritaire que les doctorants, post-doctorants (30 %) et les personnels ITA/TRF/BIATOSS (33%).

Au regard des propos recueillis dans le cadre de l'enquête qualitative et des commentaires laissés par les répondants de l'enquête quantitative, il semble y avoir deux raisons pour lesquelles les répondants souhaitent avoir du personnel dédié[23].

> « Ce travail de gestion des données […] est extrêmement chronophage. […] DMP Opidor est très redondant dans ses questions. On demande toujours plus de process aux chercheurs ce qui empiète largement le temps qu'ils consacrent à leur recherche effective. Les chercheurs ont besoin de déléguer toutes ces tâches.[24] » (H-SHS)

> « J'aimerais être accompagnée, […] je fais de la recherche en droit et j'aimerais pouvoir me consacrer à ma recherche en droit. J'aimerais qu'on me dise, ben toi tu t'occupes d'analyser les résultats […] moi je m'occupe de trouver la bonne plateforme, de mettre tous les documents sur la plateforme, de vous faciliter les accès […] » (DSPEG)

La gestion des données de la recherche, algorithmes et codes sources demande un travail chronophage qui s'ajoute à une masse de travail déjà conséquente, que ce soit pour préparer les données à la diffusion (nommage, format), choisir un entrepôt pour les diffuser ou encore rédiger un plan de gestion de données.

> « L'idée d'avoir du personnel compétent et dédié à cette question pourrait grandement faciliter la mise en place d'une gestion adaptée des données de recherche. [25]» (MST2I)

> « […] La gestion de la donnée je pense que c'est un métier différent […] et c'est pas le nôtre.» (SdM)

Par ailleurs, disposer de personnel pour la mise en œuvre de la gestion des données est perçu comme un facilitateur de la mise en place d'une bonne gestion des données. Pour les chercheurs qui se présentent comme non familier de ces pratiques c'est aussi la garantie d'avoir une expertise apportée par un tiers.

> « […] je suis ingénieure plateforme, c'est à dire que je suis au service de tout le monde, et là en plus […] j'ai un poste qui est mutualisé […] Donc ça permet d'avoir une vision des besoins » (BABS)

Au sein des structures de recherche, ces personnels de soutien connaissent les besoins des équipes de recherche. Lors des entretiens qualitatifs, une chercheuse du pôle BABS a souligné le caractère indispensable de l'aide reçue par le personnel d'appui.

Néanmoins, il ne suffit pas d'avoir du personnel dédié à la gestion des données mais également d'identifier les personnes ressources :

> « Savoir au sein d'un écosystème universitaire, il y a aussi tel bureau qui s'occupe de ça, telle personne qui sait faire ça […] » (UPEE)

> « Rien que identifier qui est le DPO rien que ça c'est pas évident » (DSPEG)

> « C'est vraiment des questions d'éthique qui peuvent se poser ou des fois on sait pas trop à qui s'adresser » (DSPEG)

[23] Sauf mention contraire, les citations ci-dessous sont issues des propos recueillis dans le cadre de l'enquête qualitative (juillet 2021)

[24] Commentaire recueilli au cours de l'enquête quantitative

[25] Commentaire recueilli au cours de l'enquête quantitative

Des informations générales

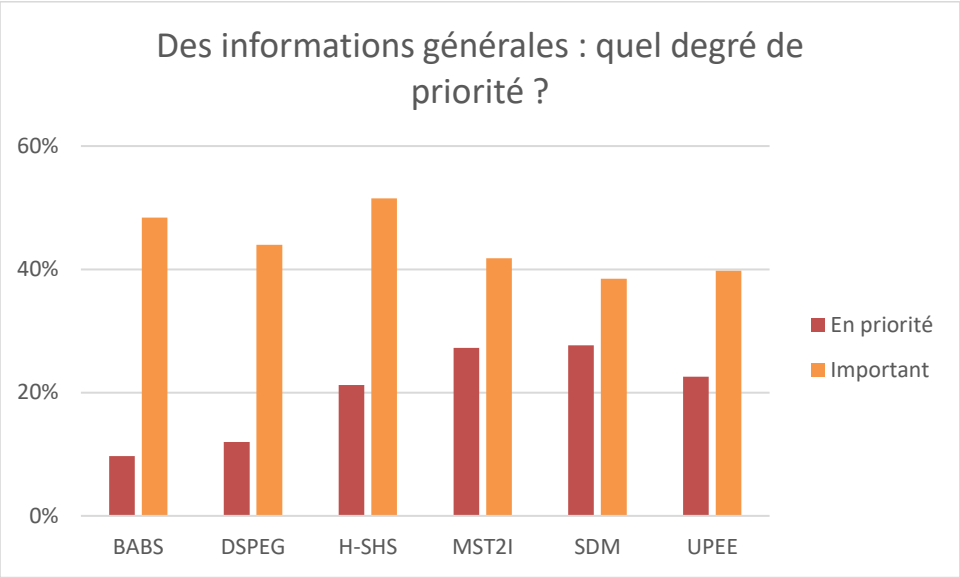

Figure 35 Degré de priorité pour "Des informations générales"

La seconde forme d'aide considérée prioritaire ou importante par une majorité de répondants est de disposer d'informations générales (67 %). Si 43 % des répondants l'ont déclarée « important », seuls 24 % en font une priorité.

Ces informations générales sur la gestion des données peuvent être diffusées sous forme de guide qui présente un ensemble de bonnes pratiques pour bien gérer ses données. Il s'agit aussi de disposer de supports concis pour faciliter l'accès à l'information [26].

> « Il nous faudrait le code de la route de la gestion des données [27]» (MST2I)
>
> « Commencer par des brochures / pages concises (memo des éléments importants) […] » (UPEE)

Tout niveau de priorité confondu, 95 % des répondants déclarent que disposer d'informations générales est nécessaire. Néanmoins, une aide sous cette forme apparaît complémentaire au fait de disposer de personnel dédié ou des ateliers et séminaires techniques.

> « Besoin de référents sur les questions générales et techniques de Fairisation, et autres questions en relation avec la science ouverte […] » (UPEE)

Le personnel dédié à la gestion des données peut être un moyen pour disposer d'informations générales sur la gestion des données et plus généralement sur la science ouverte.

[26] Sauf mention contraire, les citations ci-dessous sont issues des commentaires recueillis dans le cadre de l'enquête quantitative

[27] Propos recueillis lors de l'enquête qualitative (juillet 2021)

Des ateliers et séminaires techniques

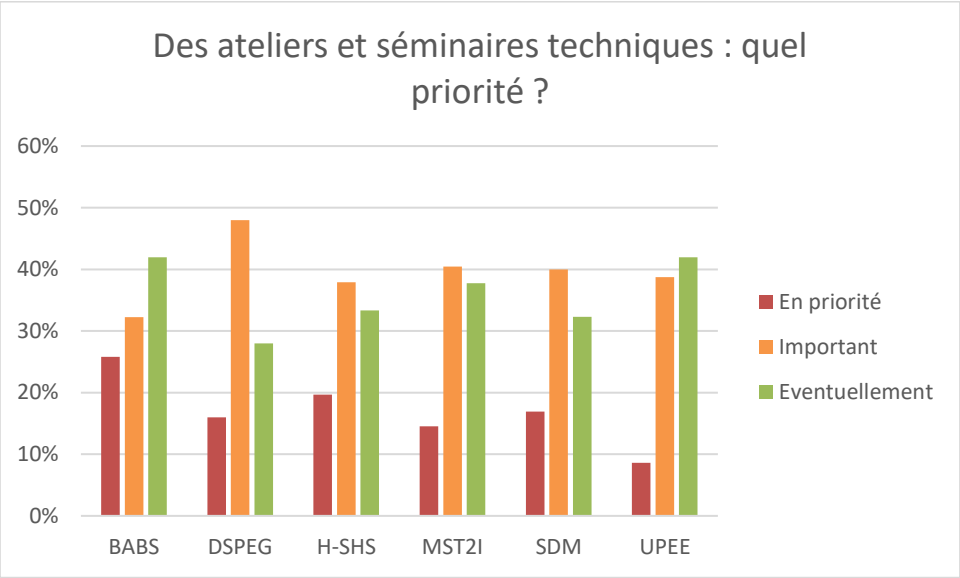

Figure 36 Degré de priorité pour "Des ateliers et séminaires techniques"

Les ateliers et séminaires techniques apparaissent comme une forme d'aide secondaire. Les résultats sont assez partagés, oscillant entre 40 % qui le considère comme important et 37 % qui le considère comme une aide éventuelle.

Les personnels ITA/TRF/BIATOSS accordent plus d'importance (53 %) et de priorité (21 %) à ce type d'aide que les enseignants-chercheurs (39 % ; 14 %), chercheurs (35 % ; 12 %) et doctorants et post-doctorants (37 % ; 16 %).

> « Des formations avec des personnes qui pourront indiquer les solutions idoines pour chacun.e. [28]» (H-SHS)
>
> « j'ai pu assister à des webinaires ou autres sur ce types (*sic*) de sujets mais […] sans mise en pratique, on oublie rapidement les informations collectées. Un document ou outil d'aide permettant d'être guidé sur les différentes étapes de la procédure serait vraiment d'un grand secours. » (MST2I)

Les ateliers et formations doivent être adaptés aux problématiques rencontrées par les participants pour proposer des solutions (outils, ressources, etc.) accessibles qui permettent une mise en pratique de la gestion des données. Un guide pratique peut être utile après un atelier ou séminaire technique afin de disposer d'un outil aide-mémoire auquel se reporter.

> « Il faudrait qu'il y ait des formations […] dans les écoles d'ingénieurs et les universités […] une bonne éducation, formation à la gestion des données comme on en a eu avec le matériel concret, les déchets [29]» (MST2I)
>
> « Une courte formation (demi-journée) sur les bonnes pratiques m'aurait été utile pendant mon doctorat. » (MST2I)

[28] L'ensemble des citations sont extraites de l'enquête quantitative, sauf mention contraire.
[29] Propos recueillis lors de l'enquête qualitative (juillet 2021)

Ces séminaires techniques doivent être accompagnés par des formations dispensées dans le cadre des études universitaires afin de sensibiliser les étudiants à la gestion des données de la recherche.

> « Mon principal problème est que je fais une étude longitudinale en sciences du langage / sciences de l'éducation. J'ai donc énormément de données, variées, donc je suis dans le qualitatif et le quantitatif. Je me rends compte que j'ai besoin d'une aide informatique. Toutes les formations proposées en doctorat sont sur des données homogènes et quantitatives, et les logiciels correspondants. J'ai besoin de formations sur ce qui existe dans les logiciels de traitement de données hétérogènes, quantitatives et qualitatives. […] » (H-SHS)

> « Une formation pour me rassurer sur mes pratiques actuelles et éventuellement les faire progresser... » (UPEE)

En termes de formation, les attentes portent sur des sujets très divers (par exemple sur les bonnes pratiques de gestion ainsi que les solutions disponibles pour les mettre en œuvre, identifier et utiliser des outils de traitement de données, etc.) avec des attentes variées de la sensibilisation à la maîtrise d'outils. La formation est également considérée comme un moyen de confronter voire améliorer sa pratique de gestion de données.

L'automatisation des tâches

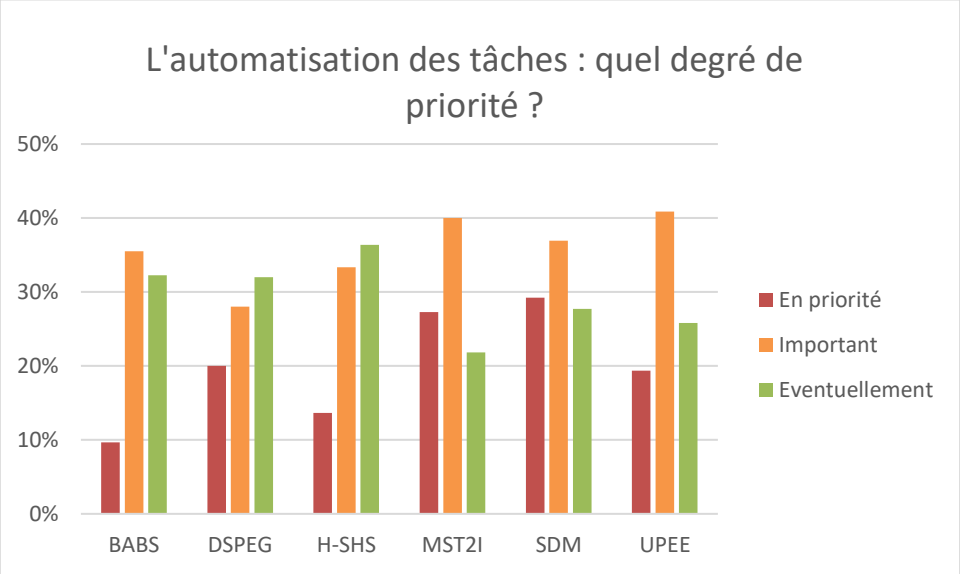

Figure 37 Degré de priorité pour une "automatisation des tâches"

Une aide sous la forme d'une automatisation des tâches intéresse également une majorité de répondants (61 %). Néanmoins, on constate qu'elle est plus sollicitée par les pôles constitués par des disciplines Sciences et techniques. Elle est en effet considérée comme importante voire prioritaire par la majorité des enquêtés des pôles MST2I (67 %), SdM (66 %) et UPEE (60 %). Dans les pôles BABS, DSPEG et H-SHS, les avis sur le degré de priorité de cette aide sont plus divisés. En effet, la proportion de répondants qui ont répondu « important » (36 %, 28 % et 33 %) est proche de celle des répondants qui ont répondu « éventuellement » (32 %, 32 % et 36 %).

Si cette aide, tout degré de priorité confondu, intéresse une grande majorité de répondants, on peut se demander ce que représente l'automatisation des tâches pour ces derniers.

« […] on clique sur le bouton, tac ça nous dit que c'est de la donnée personnelle, capteur, ceci et on nous dit quelles sont les recommandations et dans les dépôts de données ça nous envoie sur les entrepôts avec ce qu'il y a à remplir.[30] » (MST2I)

L'automatisation peut être perçue comme un outil qui guide la gestion des données en fonction du type de données utilisées et indique les bonnes pratiques à suivre. Par conséquent, l'automatisation des tâches apparaît comme un facilitateur pour la gestion des données.

La notion d'automatisation des tâches recouvre certainement différentes réalités pour ceux qui l'ont déclaré nécessaire. Néanmoins, dans la mesure où la gestion des données est perçue comme une activité chronophage par une partie des répondants[31], il est probable que cette forme d'aide soit envisagée comme un gain de temps par ceux qui l'ont déclaré nécessaire.

Un accompagnement individualisé

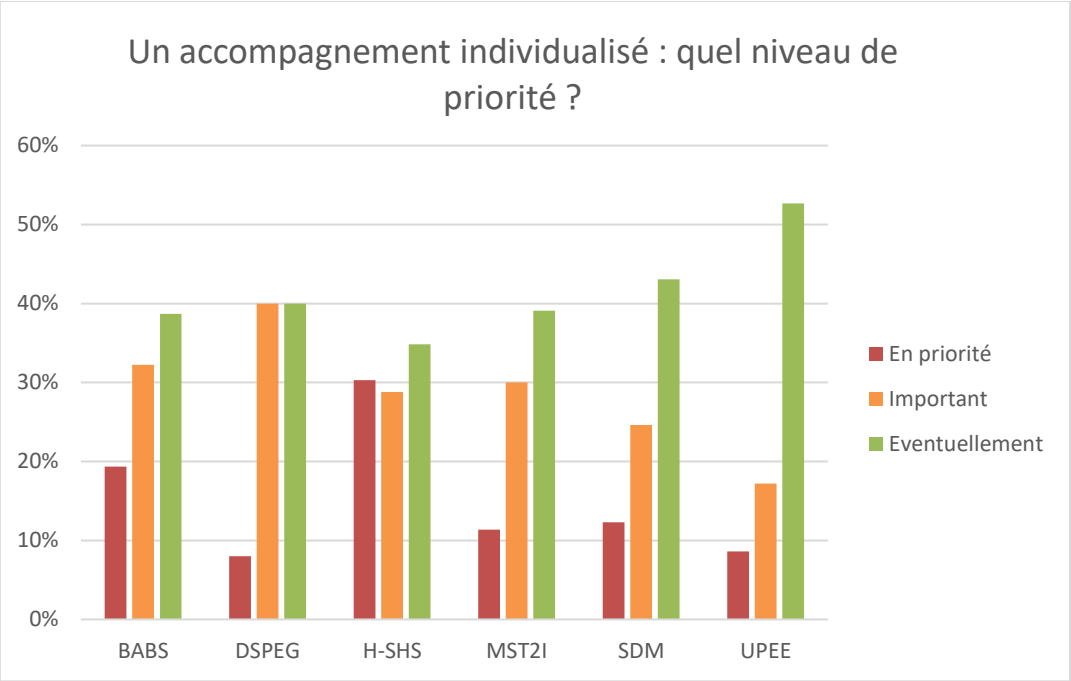

Figure 38 Degré de priorité pour un "accompagnement individualisé"

83 % des enquêtés ont déclaré qu'un accompagnement individualisé est nécessaire. Néanmoins, le degré de priorité exprimé varie selon les pôles.

Cette forme d'aide est davantage considérée importante ou prioritaire par les répondants des pôles H-SHS (59 %), BABS (52 %) et DSPEG (48 %). Néanmoins, pour les répondants du pôle DSPEG, ce n'est pas une priorité.

Par ailleurs, les avis sur le degré de priorité sont divisés au sein des pôles BABS, DSPEG et MST2I. Dans ces trois pôles, le pourcentage de répondants qui considèrent l'accompagnement individualisé comme une aide prioritaire ou importante et le pourcentage de ceux qui la déclarent comme éventuelle sont proches. Ainsi les attentes des répondants au sein d'un même pôle peuvent être hétérogènes.

[30] Extrait d'un entretien de l'enquête qualitative
[31] Cf. Du personnel dédié à la gestion des données

La proportion de réponses « éventuellement » est relativement équivalente dans chaque pôle (35 % à 43 %) hormis pour le pôle UPEE. En effet, 53 % des répondants affiliés à celui-ci ont déclaré un degré de priorité « éventuellement ».

Au regard des besoins exprimés, l'accompagnement individualisé est considéré par 42 % des enquêtés comme une forme d'aide éventuelle à laquelle ils peuvent recourir potentiellement, certainement en fonction des problématiques ou des besoins qu'ils rencontrent.

Le pôle H-SHS semble avoir des attentes sur plusieurs formes d'aide avec un plus haut niveau de priorité que les autres pôles. Pour trois formes d'aide sur cinq, le pourcentage de répondants du pôle H-SHS qui déclarent un degré de priorité « important » ou « en priorité » est plus élevé que dans les autres pôles. Il s'agit des aides sous forme de personnel dédié à la gestion des données (85 %), d'informations générales (73 %) et d'accompagnement individualisé (59 %).

Conclusion

L'objectif de cette enquête était d'appréhender les pratiques et les besoins des personnels scientifiques de l'Université de Toulouse en ce qui concerne la gestion des données de recherche, algorithmes et codes sources. Les réponses recueillies révèlent plusieurs éléments intéressants qui ouvrent de nouvelles perspectives et permettent d'ores et déjà de proposer des solutions concrètes.

À l'Université de Toulouse, on constate un large panel de données. Au sein d'un même pôle, la nature des données utilisées est multiple, même si à l'échelle de l'UT ce sont les données de simulation/modélisation qui sont les plus couramment utilisées (56 %), suivies des données d'observation (52 %) et d'expérimentation (45 %). Ce premier résultat est à mettre en regard avec une enquête européenne de 2021, dans laquelle les données expérimentales (64 %) et les données d'observation (58 %) sont prépondérantes[32]. La réutilisation des données est une pratique courante qui nécessite presque toujours un travail de curation. La majorité des répondants a l'habitude de documenter ses données : le fichier « Lisez-moi » est la forme la plus répandue dans tous les pôles. À travers ces pratiques et la fairisation des données, ce sont les enjeux de la qualité des données et de l'éthique qui apparaissent en filigrane.

La rédaction de plan de gestion de données est une pratique encore peu répandue mais analogue à celle observée dans d'autres enquêtes[33]. Parmi les enquêtés, 43 % des doctorants et post-doctorants indiquent ne pas savoir de quoi il s'agit. Une aide sur la rédaction de PGD est d'ailleurs demandée dans tous les pôles et pourrait être davantage sollicitée à l'avenir en raison du contexte national et des exigences des financeurs. Conscients des besoins existants, des services de proximité sont déjà actifs au sein des établissements de l'Université de Toulouse et coopèrent aujourd'hui pour mieux répondre aux attentes. Une sensibilisation doit aussi être menée auprès de l'Ecole des Docteurs et reconduite chaque année. Le plan de gestion de données est une porte d'entrée à un accompagnement plus large à la gestion raisonnée des données.

Les solutions de stockage utilisées sont multiples. Le stockage des données de recherche nécessite des espaces avec une capacité importante. L'enquête révèle que l'archivage résulte d'une pratique individuelle et dénote la méconnaissance ou l'absence de politiques d'archivage au niveau des structures.
Les besoins exprimés révèlent de fortes attentes autour d'outils techniques (espace de stockage sécurisé pendant la recherche, espace d'archivage des données). Une politique de stockage et d'archivage ainsi qu'une offre d'accompagnement par les structures pourraient répondre aux attentes de la communauté scientifique. Le DROCC, data center régional d'Occitanie Ouest, pourra amener des solutions répondant aux spécificités disciplinaires et aux aspects collaboratifs.

Les répondants sont favorables au partage des données de recherche et/ou algorithmes et codes sources selon les principes de la science ouverte. Toutefois, le délai lié à leur publication est un élément à prendre en compte. L'accès aux données d'autres chercheurs et une plus grande transparence de la recherche motivent cette démarche de partage. La majorité des répondants a déjà diffusé des données, algorithmes ou codes sources après un projet de recherche. Cette pratique reste inégalement répandue selon les domaines disciplinaires. Pour

[32] Commission européenne, Direction générale pour la Recherche et l'Innovation (2022). *European Research Data Landscape : final report*, Office des publications de l'Union européenne. https://data.europa.eu/doi/10.2777/3648

[33] En 2022, l'université Paris Saclay note que 23 % des répondants de son enquête ont déjà rédigé un PGD (Brenel M., Mercier C., Suhan S., Kassas A., Ménard C., et al. (2022). *Rapport d'analyse – Enquête : Les données de la recherche à l'université Paris Saclay, panorama et perspectives*. Université Paris-Saclay, p. 22), le CNAM 19 % (Bertram M.-L., (2022). *Gestion et ouverture des données de la recherche : pratiques, représentations, besoins*. Conservation National des Arts et Métiers, p. 30, https://cnam.hal.science/hal-03824487/) tandis qu'ils sont 14 % des enquêtés à l'Université Grenoble Alpes (Cellule Data Grenoble. (2022). *Rapport sur l'enquête sur les usages et les besoins pour la gestion des données de la recherche sur le site de l'Université Grenoble Alpes*, p. 50.)

appuyer cette dynamique, des préconisations au niveau de l'établissement ou des unités de recherche pourraient aider la communauté scientifique dans cette dissémination. Les structures pourront aussi s'appuyer sur l'écosystème Recherche Data Gouv, et notamment sur le futur Atelier de la donnée Occitanie Ouest, pour accompagner la diffusion des données qui est amenée à se développer avec la mise en œuvre du 2nd Plan national pour la Science ouverte. Cette évolution va nécessiter un accompagnement accru à la curation et au dépôt des données dans des entrepôts ainsi qu'une sensibilisation aux principes FAIR et à leur application.

L'enquête permettait aux participants d'indiquer la forme de l'aide souhaitée. Des priorités se dégagent à la fois à l'échelle du site et par pôle.
Disposer de personnel dédié pour la gestion des données est une priorité dans tous les pôles. Cette gestion implique des compétences sur différents aspects (techniques, juridiques, informatiques, documentaires etc.) qu'il est nécessaire de réunir et rendre visibles.

Disposer d'informations générales sur la gestion des données est nécessaire pour l'ensemble des répondants. Un guide de bonnes pratiques, en lien avec une politique d'établissement sur la gestion des données permettrait de répondre à ce besoin.

Ces deux derniers points trouvent un écho dans la mise en œuvre des ateliers de la donnée proposée par l'écosystème Recherche Data Gouv.

Cette enquête a permis de relever des spécificités par pôle de recherche (cf. Tableau 6 Classement des besoins par pôle de recherche en fonction du % de réponses obtenues (question « Sélectionnez les éléments qui correspondent à vos besoins ») qu'il serait nécessaire de prendre en compte dans l'accompagnement afin de mieux répondre à chaque communauté.
Les données manipulées par les pôles H-SHS, DSPEG et BABS nécessitent plus souvent un traitement des données personnelles en lien avec la réglementation (RGPD) ainsi qu'une anonymisation avant leur partage. Ces pôles sont également davantage concernés par les questions éthiques. Cela implique une charge de travail supplémentaire qui pourrait être allégée grâce à des services d'accompagnement spécialisés dans ce domaine. Une aide sur ce sujet pourrait faciliter le partage des données dans ces pôles.
Dans les pôles UPEE, SdM, MST2I et BABS, on observe une pratique de diffusion des données plus avancée. Une offre de services comprenant une orientation vers des entrepôts disciplinaires et un accompagnement au dépôt des données dans le respect des principes FAIR permettrait d'enrichir les pratiques et de répondre aux besoins exprimés.

Le pourcentage relativement élevé de réponses « Je ne sais pas » à plusieurs questions (PGD, RGPD, licences, espace nécessaire au stockage) interpelle sur l'importance de proposer une sensibilisation générale à la science ouverte et à la gestion des données de la recherche à l'ensemble de la communauté scientifique et son renouvellement à tous les nouveaux entrants.

Cet accompagnement autour des données, algorithmes et codes sources ne peut se faire sans les infrastructures, les services associés, les ressources humaines et compétences suffisantes. Ces questions sont portées à la fois dans les établissements et à l'échelle de l'université de Toulouse par le Comité de la science ouverte (CéSO) et le projet d'atelier de la donnée Occitanie Ouest.

Annexes

Répartition des personnels en fonction de leur statut par pôle de recherche

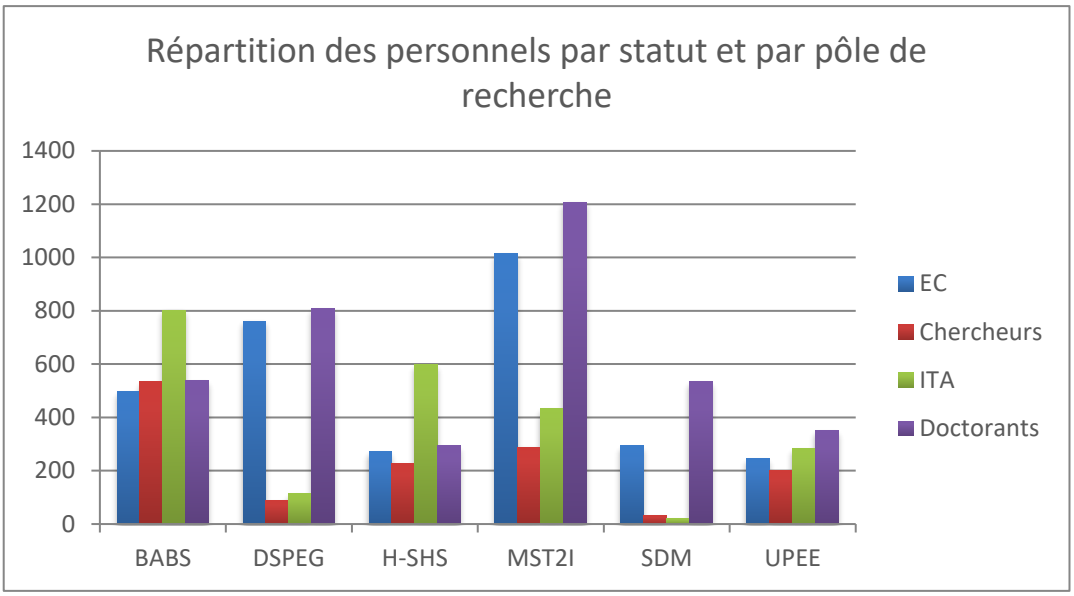

Figure 39 Répartition des personnels en fonction de leur statut par pôle de recherche

Source : données chiffrées Université de Toulouse (2019)

Répartition des répondants (genre, âge, statut) par pôle de recherche

Tableau 4 Répartition, genre, âge, statut par pôle de recherche (%, n=547)

		À quel pôle de recherche appartenez-vous ?					
	Total	BABS	DSPEG	H-SHS	MST2I	SdM	UPEE
Nombre de répondants	*547*	*34*	*30*	*72*	*243*	*69*	*99*
%	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Homme	57,8	50,0	36,7	33,3	68,7	56,5	58,6
Femme	32,7	35,3	56,7	56,9	23,0	34,8	29,3
Autre/ Je ne souhaite pas répondre	9,5	14,7	6,7	9,7	8,2	8,7	12,1
20-29	23,2	20,6	23,3	8,3	28,8	21,7	22,2
30-39	21,4	17,6	36,7	34,7	21,0	14,5	14,1
40-49	22,7	20,6	23,3	26,4	20,2	21,7	27,3
50+	25,4	20,6	13,3	26,4	24,7	34,8	25,3
Je ne souhaite pas répondre	7,3	20,6	3,3	4,2	5,3	7,2	11,1
Enseignant-chercheur	33,5	23,5	33,3	26,4	41,2	33,3	23,2
Chercheur	22,7	23,5	3,3	13,9	19,3	31,9	36,4
Doctorant/Post-doctorant	31,6	29,4	53,3	44,4	31,7	23,2	22,2
ITA/TRF/BIATOSS/ Autre	12,2	23,5	10,0	15,3	7,8	11,6	18,2

Répartition des répondants par établissement

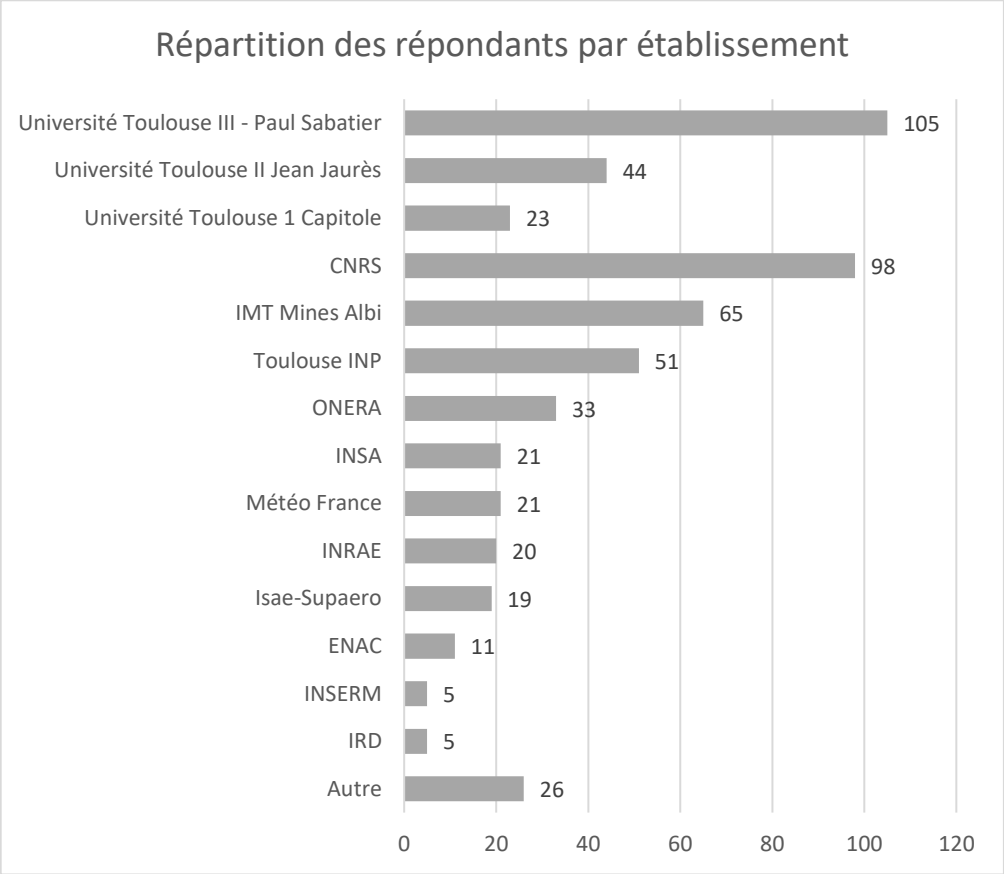

Figure 40 Répartition des répondants par établissement (nombre de réponses, n=547)

Nombre de répondants par discipline et appartenance aux pôles de recherche

Tableau 5 Nombre de répondants par discipline et appartenance aux pôles de recherche

Disciplines Science et Technologie (n=415, 76 %)		
Sciences informatiques et informatique : systèmes informatiques et d'information, sciences informatiques, calcul scientifique, systèmes intelligents	106	83,0 % en pôle MST2I
Ingénierie des produits et des procédés : conception de produits, conception et contrôle des procédés, méthodes de construction, génie civil, systèmes énergétiques, ingénierie des matériaux	86	87,2 % en pôle MST2I
Sciences du Système Terre : géographie physique, géologie, géophysique, sciences de l'atmosphère, océanographie, climatologie, cryologie, écologie, changements environnementaux globaux, cycles biogéochimiques, gestion des ressources naturelles	86	80,2 % en pôle UPEE
Mathématiques : tous les domaines des mathématiques, pures et appliquées, plus les fondements mathématiques des sciences informatiques, la physique mathématique et les statistiques	50	92,0 % en pôle MST2I
Physique de la matière condensée : structure, propriétés électroniques, fluides, nanosciences, physique biologique	47	89,4 % en pôle MST2I et SdM
Ingénierie des systèmes et de la communication : ingénierie électrique, électronique, de la communication, optique et des systèmes	45	91,1 % en pôle MST2I
Chimie de synthèse et matériaux : synthèse des matériaux, relations structure-propriétés, matériaux fonctionnels et avancés, architecture moléculaire, chimie organique	41	75,6 % en pôle SdM

Sciences de l'Univers : astro-physique/chimie/biologie; système solaire; astronomie stellaire, galactique et extragalactique, systèmes planétaires, cosmologie, science de l'espace, instrumentation	28	89,3 %	en pôle	UPEE
Chimie physique et analytique : chimie analytique, théorie chimique, chimie physique/physico-chimie	26	61,5 %	en pôle	SdM
Constituants fondamentaux de la matière : physique des particules, nucléaire, des plasmas, atomique, moléculaire, des gaz et optique	16	62,5 %	en pôle	SdM
Disciplines SHS (n=104, 19 %)				
Le monde social, diversité, population : sociologie, psychologie sociale, anthropologie sociale, démographie, éducation, communication	51	78,4 %	en pôle	H-SHS
Institutions, valeurs, environnement et espace : science politique, droit, science de la durabilité, géographie, étude et aménagement du territoire	39	94,9 %	en pôle	DSPEG et H-SHS
L'esprit humain et sa complexité : sciences cognitives, psychologie, linguistique, philosophie de l'esprit	27	85,2 %	en pôle	H-SHS
Individus, marchés et organisations : économie, finance et management	16	68,8 %	en pôle	DSPEG
Cultures et production culturelle : littérature, philologie, études culturelles, étude des arts, philosophie	7	85,7 %	en pôle	H-SHS
L'étude du passé humain : archéologie et histoire	5	100,0 %	en pôle	H-SHS
Discipline Vie et Santé (n=28, 6 %)				
Neurosciences et troubles neurologiques	7	85,7 %	en pôle	BABS
Biologie moléculaire, biochimie, biologie structurale et biophysique moléculaire	6	83,3 %	en pôle	BABS
Génétique moléculaire, génétique quantitative, épidémiologie génétique, épigénétique, génomique, métagénomique, transcriptomique, protéomique, métabolomique, glycomique	6	100,0 %	en pôle	BABS
Démécologie, synécologie, écologie des écosystèmes, biologie de l'évolution, écologie comportementale, écologie microbienne	6	66,7 %	en pôle	BABS
Sciences de la vie appliquées, biotechnologie, et ingénierie moléculaire et des biosystèmes	6	83,3 %	en pôle	BABS
Technologies médicales appliquées, diagnostics, thérapies et santé publique	3	100,0 %	en pôle	BABS
Physiologie, physiopathologie et endocrinologie	2	100,0 %	en pôle	BABS
Biologie cellulaire et du développement	1	100,0 %	en pôle	BABS

Questionnaire

Enquête sur la gestion des données de la recherche, algorithmes et codes sources

Les données de la recherche deviennent un enjeu majeur pour la gestion et la diffusion des connaissances scientifiques. Sous forme d'enregistrements factuels (chiffres, textes, images et sons), elles sont utilisées comme sources principales pour la recherche scientifique, et sont généralement reconnues par la communauté scientifique comme nécessaires pour valider des résultats de recherche.

Le Comité de réflexion pour le partage et la valorisation des données de la recherche et coordination de la Science Ouverte de l'université fédérale Toulouse Midi-Pyrénées (CéSO UFTMiP) lance **du 30 mai 2022 au 15 août une enquête en ligne afin de mieux connaître les pratiques et besoins en matière de gestion des données de la recherche**. Il s'agit d'un état des lieux, à destination des enseignants-chercheurs, chercheurs, doctorants et autres personnels scientifiques des établissements et organismes de recherche de l'UFTMIP.

En répondant à ce questionnaire, **vous participez à définir la future offre de services d'accompagnement au plus près de vos besoins.**

Répondre à cette enquête vous prendra environ 10 mn. Le questionnaire comporte 5 thèmes : vos données de recherche, algorithmes et codes sources, vos pratiques de gestion de données, les freins et les leviers par rapport à la science ouverte, vos besoins et attentes. La participation est anonyme. Consultez les informations sur la gestion des données personnelles de l'enquête ci-dessous.

La synthèse des résultats de l'enquête sera consultable sur la page web du CéSO à l'automne 2022.

I. Vos données de recherche, algorithmes et codes sources

Intéressons-nous dans un premier temps à la nature des données que vous utilisez dans vos travaux de recherche.

Q1. En règle générale, quelles données utilisez-vous dans vos recherches ?

Veuillez choisir toutes les réponses qui conviennent :

1. Observation (remontées de terrain : données de capteurs, images satellites, enquêtes, entretiens...)
2. Expérimentation (par exemple les données obtenues en laboratoire, séquençage ADN, chromatographie, eye-tracking, données comportementales...)
3. Simulations et modélisations (générées par des modèles informatiques…)
4. Données dérivées (issues du traitement et de la compilation de données brutes : bases de données, text/data mining…)
5. Données de référence (préalablement organisées, gérées, voire publiées : banques de données de séquences ADN, de structures chimiques, portails de données spatiales, données INSEE...)
6. Autre, précisez :
7. Je n'utilise pas de données **[renvoi à Q23]**

Q2. Dans quel(s) format(s sont ces données ?

Veuillez choisir toutes les réponses qui conviennent :

1. Données textuelles (issues de corpus de textes, réseaux sociaux, archives, cahier de laboratoire, séquences biologiques : ADN, ARN, protéine, …)
2. Données audiovisuelles (enregistrements audio, vidéo, films...)
3. Données d'images fixes (photographies, dessins… d'objets, de paysages, d'architectures, spectroscopies, imagerie…)
4. Données numériques/chiffrées (séries statistiques, logs de serveurs, traces...)
5. Données géographiques
6. Données 3D
7. Codes sources
8. Algorithmes
9. Autre, précisez :

Réutilisation des données

Q3. Avez-vous déjà réutilisé des données de recherche et/ou algorithmes et codes sources ?

Veuillez choisir toutes les réponses qui conviennent :

1. Oui, transmises par un partenaire ou un pair
2. Oui, disponibles en ligne
3. Oui, mes propres données
4. Non, je n'ai pas réutilisé de données

➢ **Si oui**

Q4. La réutilisation a-t-elle majoritairement imposé un travail de nettoyage, sélection, correction… ?

Veuillez choisir toutes les réponses qui conviennent :

1. Oui, que j'ai réalisé
2. Oui, que j'ai sous-traité
3. Non

➢ **Si non**

Q4ter. Pourquoi n'avez-vous pas réutilisé de données, algorithmes ou codes sources?

Veuillez choisir toutes les réponses qui conviennent :

1. Difficiles à trouver
2. Nécessitent un travail de curation (donnée) ou de portabilité (codes)
3. Soulèvent des questions de qualité, de fiabilité, de précision
4. Freins juridiques
5. Je n'en ai pas vu la nécessité
6. Autre

Q5. Utilisez-vous, lorsque c'est possible, des formats ouverts (.txt, .csv, .png, .html, .mkv, .ods, .odt, …) quand vous générez des données de recherche ?

« Les formats ouverts correspondent à des fichiers **encodés de façon transparente**, leur recette de fabrication fait partie du domaine public. Ils sont **interopérables**, c'est à dire qu'ils peuvent être créés, lus et modifiés par tous les logiciels destinés à traiter le type du fichier (image, texte, audio, etc.). » (Doranum)

Veuillez sélectionner une seule des propositions suivantes :

1. Oui
2. Non
3. Je ne sais pas

Q6. Avez-vous des précisions à ajouter sur la nature, les formats de vos données de recherche, algorithmes et codes sources ?

Veuillez écrire votre réponse ici :

II. Pratiques de gestion de données

Dans cette partie nous voudrions connaitre vos habitudes et pratiques concernant la gestion de données de recherche.

Planification et description

Q7. Avez-vous déjà rédigé un plan de gestion de données (PGD)/ Data Management Plan (DMP) ou participé à sa rédaction ?

Veuillez sélectionner une seule des propositions suivantes :

1. Oui
2. Non
3. Je ne sais pas ce qu'est un plan de gestion de données

Q8. Avez-vous l'habitude de décrire/documenter vos données de recherche et/ou algorithmes et codes sources ?

Veuillez sélectionner une seule des propositions suivantes :

1. Oui
2. Non
3. Je ne sais pas

➢ **Si oui**

Q9. Sous quelle forme décrivez-vous/ documentez-vous vos données de recherche et/ou algorithmes et codes sources ?

Veuillez choisir toutes les réponses qui conviennent :

1. Fichier Lisez-moi / Read me
2. Standards de métadonnées
3. Autre, précisez :

Q10. Vos données sont-elles concernées par le Règlement Général sur la Protection des Données (RGPD) ?

Veuillez sélectionner une seule des propositions suivantes :

1. Souvent
2. Parfois
3. Rarement
4. Jamais
5. Je ne sais pas

Stockage, sauvegarde et partage

Les questions suivantes concernent les données, algorithmes et codes sources en cours d'utilisation dans le cadre d'une activité de recherche (projet, doctorat, expérience, …).

Q11. À combien estimez-vous l'espace nécessaire au stockage de vos données de recherche et/ou algorithmes et codes sources ?

Veuillez sélectionner une seule des propositions suivantes :

1. Moins de 2 Go
2. Moins de 10 Go
3. Moins de 50 Go
4. Entre 50 et 500 Go
5. Entre 500 Go et 1 To (1 Tera-octet = 1000 Go)
6. Entre 1 To et 1 Po (1 Peta-octet = 1000 To)
7. Plus d'1 Po
8. Je ne sais pas

Q12. Où stockez-vous vos données et/ou algorithmes et codes sources actuellement ?

Veuillez choisir toutes les réponses qui conviennent :

1. sur mon ordinateur professionnel
2. sur mon ordinateur personnel
3. sur un support externe (disque dur externe, clé USB, CD, DVD,…)
4. sur un serveur géré au sein de ma structure (équipe, laboratoire, département, institut, …)
5. sur un cloud institutionnel
6. sur un cloud privé
7. Autre

Q13. Lors de votre activité de recherche, faites-vous (vous ou votre équipe) régulièrement des copies de sauvegarde ?
Veuillez sélectionner une seule des propositions suivantes :

1. Oui, automatiquement
2. Oui, manuellement
3. Non
4. Je ne sais pas

Q14. En dehors de vous, qui est autorisé à accéder à vos données et/ou algorithmes et codes sources ?
Veuillez choisir toutes les réponses qui conviennent :

1. Personne
2. Mon équipe, mon groupe de travail, les membres du projet
3. Mon unité de recherche
4. Tout le monde
5. Autre, précisez :

Q15. Avez-vous déjà été sollicité pour une demande d'accès à vos données et/ou algorithmes et codes sources ?
Veuillez sélectionner une seule des propositions suivantes :

1. Oui
2. Non

Archivage et diffusion de vos données

Dans cette partie, nous nous intéressons à ce que deviennent vos données, algorithmes et codes sources une fois votre travail de recherche terminé.

Q16. Existe-t-il une pratique d'archivage de données de recherche et/ou algorithmes et codes sources (critères, modalités, nommage des fichiers...) après le travail de recherche … ?
Choisissez la réponse appropriée pour chaque élément :

	Oui	Non	Je ne sais pas

dans votre établissement	☐	☐	☐
dans votre unité de recherche	☐	☐	☐
à votre niveau individuel	☐	☐	☐

Q17. Vous-même, avez-vous déjà diffusé vos données et/ou algorithmes et codes sources ?
Choisissez la réponse appropriée pour chaque élément :

	Oui	Non	Non concerné.e
Données	☐	☐	☐
Algorithmes et codes sources	☐	☐	☐

Si oui :

Q18. Attribuez-vous une licence (licence ouverte/ EtatLab, Creative Commons, CDbL, …) à ces …
(a) … données (si Q17 Données=oui)
(b) … algorithmes et codes sources (si Q17 Algorithmes et codes sources =oui)
(c) … données, algorithmes et codes sources (si Q17 Données ET Algorithmes et codes sources =oui)
précisant les droits et obligations par rapport à la réutilisation ?
Veuillez sélectionner une seule des propositions suivantes :

1. Oui
2. Non
3. Je ne sais pas

Q19. Par quel(s) moyen(s) diffusez-vous vos …
(a) … données ? (si Q17 Données=oui)
(b) … algorithmes et codes sources ? (si Q17 Algorithmes et codes sources =oui)
(c) … données, algorithmes et codes sources ? (si Q17 Données ET Algorithmes et codes sources =oui)

Veuillez choisir toutes les réponses qui conviennent :
1. Un entrepôt de données (plateforme spécialisée dans l'archivage et la diffusion des données de recherche)
2. Une plateforme collaborative de type Git
3. Une archive ouverte (de type HAL), associée à une publication
4. Associé à une publication dans une revue scientifique
5. Un serveur de votre (vos) établissement(s)
6. Le site ou le blog de votre unité de recherche/le site dédié au projet de recherche
7. Votre blog ou votre site personnel
8. Un réseau social de chercheurs (ResearchGate, Academia…)
9. Autre, précisez :

Q20. Avez-vous des précisions à ajouter sur vos pratiques de gestion de données de recherche, algorithmes et codes sources (planification, documentation, stockage, sauvegarde, archivage et diffusion) ?
Veuillez écrire votre réponse ici :

III. Vers une science plus ouverte

Extrait du Plan National de la Science Ouverte :

« La science ouverte est la diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et – autant que possible – aux données, aux codes sources et aux méthodes de la recherche […] en conformité avec les principes FAIR (faciles à trouver, accessibles, interopérables, réutilisables)… »

Q21. De manière générale, une fois votre travail de recherche terminé, êtes-vous d'accord pour partager vos données de recherche et/ou algorithmes et codes sources selon les principes de la science ouverte ?
Veuillez sélectionner une seule des propositions suivantes :

1. Oui, sans délai
2. Oui, après un délai
3. Oui, après avoir publié
4. Plutôt non
5. Non, pas du tout
6. Je ne sais pas

Q22. Parmi les raisons suivantes, une fois votre travail de recherche terminé, qu'est-ce qui vous incite/inciterait à partager vos données de recherche et/ou algorithmes et codes sources ?
Veuillez choisir toutes les réponses qui conviennent :

1. Une obligation faite par le financeur de votre recherche (dans le cadre des projets européens, ANR...)
2. Une politique de votre établissement ou de votre tutelle
3. Une meilleure visibilité de vos travaux de recherche
4. Une prise en compte des pratiques de gestion de données de la recherche dans l'évaluation que l'on fera de votre travail de recherche
5. Favoriser la reproductibilité des résultats en assurant l'intégrité des données
6. Partager et avoir accès aux données des autres dans un esprit de réciprocité
7. L'occasion de susciter de nouveaux contacts, de nouvelles coopérations scientifiques
8. Une plus grande transparence de la recherche pour la science et la société
9. Rien
10. Autre, précisez :

Q23. Parmi les points suivants, d'après vous, lesquels sont des obstacles au partage des données selon les principes FAIR (Facilement trouvable, Accessible, Interopérable et Réutilisable) ?
Veuillez choisir toutes les réponses qui conviennent :

1. La charge de travail supplémentaire
2. Les coûts de la gestion des données et du partage

3. Le manque de compétences nécessaires pour gérer et partager les données
4. Le manque d'entrepôts de confiance ou reconnus par la communauté
5. Le manque d'information sur l'usage des données (traçabilité)
6. Le risque de falsification et de mauvaise interprétation des données
7. Une utilisation commerciale des données potentiellement indésirable
8. La compétition entre chercheurs
9. La protection des données et les exigences de confidentialité
10. Les restrictions légales (copyright, brevet)
11. Le manque de reconnaissance de son apport scientifique
12. Autre
13. Je ne sais pas

Q24. Précisez ici si vous voyez d'autres obstacles au partage des données selon les principes FAIR (Facilement trouvable, Accessible, Interopérable et Réutilisable) qui ne sont pas mentionnés ci-dessus :
Veuillez écrire votre réponse ici :

IV. Besoins et attentes

Cette dernière partie porte sur vos besoins et attentes afin de définir des services adaptés.

Q25. Sélectionnez les éléments qui correspondent à vos besoins
Veuillez choisir toutes les réponses qui conviennent :

1. Un espace de stockage sécurisé durant l'activité de recherche
2. Une solution d'archivage des données au-delà d'une activité de recherche
3. Une aide à la rédaction des plans de gestion de données
4. Une aide pour rendre les données FAIR (Facilement trouvable, Accessible, Interopérable et Réutilisable)
5. Une aide sur les questions de documentation (description des jeux de données, métadonnées, nommage des fichiers,...)
6. Une aide sur la sécurité informatique des données
7. Une aide sur la protection des données
8. Une aide sur les aspects éthiques
9. Une aide sur les questions liées à la propriété intellectuelle
10. Une aide sur les questions liées à la propriété industrielle
11. Une aide sur la protection du potentiel scientifique et technique de la nation (PPST)
12. L'identification des entrepôts qui correspondent à mon domaine
13. La recherche de jeux de données disponibles dans ma discipline
14. Le dépôt dans un entrepôt
15. Je n'ai pas de besoin
16. Autre, précisez :

Q26. Sous quelle forme cette aide pourrait-elle vous être apportée ?
Choisissez la réponse appropriée pour chaque élément :

Q27. Avez-vous des précisions à ajouter sur vos besoins et attentes par rapport à votre gestion de données de recherche, algorithmes et codes sources ?
Veuillez écrire votre réponse ici :

	En priorité	Important	Éventuellement	Pas nécessaire
Des informations générales	☐	☐	☐	☐
Des ateliers ou séminaires techniques	☐	☐	☐	☐
Un accompagnement individualisé	☐	☐	☐	☐
Du personnel dédié à la gestion des données de la recherche	☐	☐	☐	☐
L'automatisation des tâches	☐	☐	☐	☐

V. Informations personnelles

Q28. Cochez la case qui vous correspond :

Veuillez sélectionner une seule des propositions suivantes :

Enseignant-chercheur	☐
Chercheur	☐
Post-doctorant	☐
Doctorant	☐
ITA/ITRF/BIATOSS	☐
Autre	☐

Q29. Quel est votre établissement ou organisme de rattachement principal ?
Veuillez sélectionner une seule des propositions suivantes :

CHU	☐	INRAE	☐	Toulouse INP	☐
CNES	☐	INSA	☐	Toulouse INP ENIT	☐
CNRS	☐	INSERM	☐	Toulouse INP ENM	☐
CREPS	☐	INU Champollion	☐	Toulouse INP ENSAT	☐
ENAC	☐	IRD	☐	Toulouse INP ENSEEIHT	☐
ENSA	☐	Isae-Supaero	☐	Toulouse INP ENSIACET	☐
EnsaT	☐	IsdaT	☐	Toulouse INP PURPAN	☐
ENSFEA	☐	Météo France	☐	Université Toulouse 1 Capitole	☐
ENVT	☐	ONERA	☐	Université Toulouse II Jean Jaurès	☐

ICAM	☐	Sciences Po Toulouse	☐	Université Toulouse III - Paul Sabatier	☐
IMT Mines Albi	☐	TBS	☐	Autre	☐

Q30. Quel est votre champ disciplinaire principal ? D'après la nomenclature ERC

Veuillez sélectionner une seule des propositions suivantes :

1. Sciences Humaines et Sociales
2. Sciences et Technologies
3. Vie et Santé

Q31. Précisez votre/ vos discipline(s) …

(a) … en Sciences Humaines et Sociales (si Q30=1)

Veuillez choisir toutes les réponses qui conviennent :

1. Individus, marchés et organisations : économie, finance et management
2. Institutions, valeurs, environnement et espace : science politique, droit, science de la durabilité, géographie, étude et aménagement du territoire
3. Le monde social, diversité, population : sociologie, psychologie sociale, anthropologie sociale, démographie, éducation, communication
4. L'esprit humain et sa complexité : sciences cognitives, psychologie, linguistique, philosophie de l'esprit
5. Cultures et production culturelle : littérature, philologie, études culturelles, étude des arts, philosophie
6. L'étude du passé humain : archéologie et histoire

(b) … en Sciences et Technologies (si Q30=2)

Veuillez choisir toutes les réponses qui conviennent :

1. Mathématiques : tous les domaines des mathématiques, pures et appliquées, plus les fondements mathématiques des sciences informatiques, la physique mathématique et les statistiques
2. Constituants fondamentaux de la matière : physique des particules, nucléaire, des plasmas, atomique, moléculaire, des gaz et optique
3. Physique de la matière condensée : structure, propriétés électroniques, fluides, nanosciences, physique biologique
4. Chimie physique et analytique : chimie analytique, théorie chimique, chimie physique/physico-chimie
5. Chimie de synthèse et matériaux : synthèse des matériaux, relations structure-propriétés, matériaux fonctionnels et avancés, architecture moléculaire, chimie organique
6. Sciences informatiques et informatique : systèmes informatiques et d'information, sciences informatiques, calcul scientifique, systèmes intelligents
7. Ingénierie des systèmes et de la communication : ingénierie électrique, électronique, de la communication, optique et des systèmes
8. Ingénierie des produits et des procédés : conception de produits, conception et contrôle des procédés, méthodes de construction, génie civil, systèmes énergétiques, ingénierie des matériaux
9. Sciences de l'Univers : astro-physique/chimie/biologie; système solaire; astronomie stellaire, galactique et extragalactique, systèmes planétaires, cosmologie, science de l'espace, instrumentation
10. Sciences du Système Terre : géographie physique, géologie, géophysique, sciences de l'atmosphère, océanographie, climatologie, cryologie, écologie, changements environnementaux globaux, cycles biogéochimiques, gestion des ressources naturelles

(c) … en Vie et Santé (si Q30=3)

Veuillez choisir toutes les réponses qui conviennent :

1. Biologie moléculaire, biochimie, biologie structurale et biophysique moléculaire : synthèse, modification, mécanismes et interaction moléculaires, biochimie, biologie structurale, biophysique moléculaire transduction de signaux
2. Génétique, "omiques", bioinformatique et biologie des systèmes : génétique moléculaire, génétique quantitative, épidémiologie génétique, épigénétique, génomique, métagénomique, transcriptomique, protéomique, métabolomique, glycomique, bioinformatique, biologie computationnelle, biostatistique, biologie des systèmes
3. Biologie cellulaire et du développement : biologie cellulaire, physiologie cellulaire, transduction du signal, organogenèse, génétique du développement, formation de motifs et biologie des cellules souches, chez les végétaux et les animaux ou les microorganismes
4. Physiologie, physiopathologie et endocrinologie : physiologie des organes, physiopathologie, endocrinologie, métabolisme, vieillissement, cancérogenèse, maladies cardio-vasculaires, syndromes métaboliques
5. Neurosciences et troubles neurologiques : signalisation et fonction neuronale, neuroscience des systèmes, bases neurales des processus cognitifs et comportementaux, troubles neurologiques et psychiatriques
6. Immunité et infection : le système immunitaire et les troubles associés, biologie des agents infectieux et de l'infection, bases biologiques de la prévention et du traitement des maladies infectieuses
7. Technologies médicales appliquées, diagnostics, thérapies et santé publique : développement d'outils de diagnostic, surveillance et traitement des maladies, pharmacologie, médecine clinique, médecine régénérative, épidémiologie et santé publique
8. Ecologie, évolution et biologie environnementale : démécologie, synécologie, écologie des écosystèmes, biologie de l'évolution, écologie comportementale, écologie microbienne
9. Sciences de la vie appliquées, biotechnologie, et ingénierie moléculaire et des biosystèmes : sciences appliquées des végétaux et des animaux, sylviculture, sciences des aliments, biotechnologie appliquée, biotechnologie environnementale et marine, bioingénierie appliquée, biomasse et biocarburants, dangers biologiques

Q32. A quel pôle de recherche appartenez-vous ?

Veuillez sélectionner une seule des propositions suivantes :

1. BABS : Biologie, Agronomie, Biotechnologie, Santé
2. DSPEG : Droit, Science politique, Economie, Gestion
3. H-SHS : Humanités, Sciences humaines et sociétés
4. MST2I : Mathématiques, Sciences et Technologies de l'Information et de l'Ingénierie
5. SdM : Sciences de la Matière
6. UPEE : Univers, Planète, Espace, Environnement

Q33. Précisez, si vous le souhaitez, votre unité de recherche :

Veuillez choisir toutes les réponses qui conviennent :

[Sélection dans la liste des unités de recherche]

Q34. Vous êtes :

Veuillez sélectionner une seule des propositions suivantes :

Un homme	☐
Une femme	☐
Autre	☐
Je ne souhaite pas répondre	☐

Q35. Quel est votre âge ?

Veuillez sélectionner une seule des propositions suivantes :

1. 20-29 ans
2. 30-39 ans
3. 40-49 ans
4. 50-59 ans
5. 60 ans ou plus
6. Je ne souhaite pas répondre

Q36. Laissez un commentaire/une remarque

Veuillez écrire votre réponse ici :

Nous vous remercions d'avoir répondu à ce questionnaire et vous souhaitons une bonne continuation. La synthèse des résultats de l'enquête sera consultable sur la page web du CéSO à l'automne 2022.

Classement des besoins par pôle de recherche

Tableau 6 Classement des besoins par pôle de recherche en fonction du % de réponses obtenues (question « Sélectionnez les éléments qui correspondent à vos besoins »)

BABS	DSPEG	H-SHS	MST2I	SdM	UPEE
Espace de stockage sécurisé	Espace de stockage sécurisé	Espace de stockage sécurisé	Espace de stockage sécurisé	Espace de stockage sécurisé	Espace de stockage sécurisé
Solution d'archivage	Protection des données	Solution d'archivage	Solution d'archivage	Solution d'archivage	Solution d'archivage
Questions liées à la propriété intellectuelle	Solution d'archivage	Rédaction des plans de gestion de données	Rendre les données FAIR	Rendre les données FAIR	Rendre les données FAIR
Rendre les données FAIR	Rendre les données FAIR	Rendre les données FAIR	Rédaction des plans de gestion de données	Rédaction des plans de gestion de données	Questions liées à la propriété intellectuelle
Rédaction des plans de gestion de données	Sécurité informatique des données	Protection des données	Questions de documentation	Questions liées à la propriété intellectuelle	Questions de documentation
Questions de documentation	Questions de documentation	Questions de documentation	Questions liées à la propriété intellectuelle	Questions de documentation	Rédaction des plans de

Recherche de jeux de données	Rédaction des plans de gestion de données	Sécurité informatique des données	Identification des entrepôts	Identification des entrepôts	gestion de données Recherche de jeux de données
Identification des entrepôts Protection des données Sécurité informatique des données	Recherche de jeux de données	Recherche de jeux de données	Recherche de jeux de données Protection des données	Protection des données	Identification des entrepôts
	Questions liées à la propriété intellectuelle	Questions liées à la propriété intellectuelle		Sécurité informatique des données	Sécurité informatique des données
	Identification des entrepôts Aspects éthiques	Aspects éthiques	Sécurité informatique des données Questions liées à la propriété industrielle	Protection du PPST	Protection des données
Aspects éthiques Protection du PPST		Identification des entrepôts		Questions liées à la propriété industrielle	Le dépôt dans un entrepôt Protection du PPST
	Le dépôt dans un entrepôt	Le dépôt dans un entrepôt	Protection du PPST	Recherche de jeux de données	
Le dépôt dans un entrepôt	Protection du PPST	Protection du PPST	Aspects éthiques	Aspects éthiques Le dépôt dans un entrepôt	Aspects éthiques
Questions liées à la propriété industrielle		Questions liées à la propriété industrielle	Le dépôt dans un entrepôt		Questions liées à la propriété industrielle

Les propositions de réponses ayant obtenues le même pourcentage de réponses sont regroupées dans la même cellule

Table des illustrations