

Pratiques et usages des outils numériques dans les communautés scientifiques en France

Mariannig Le Béchec, Aline Bouchard, Philippe Charrier, Claire Denecker, Gabriel Gallezot, Stéphanie Rennes

To cite this version:

Mariannig Le Béchec, Aline Bouchard, Philippe Charrier, Claire Denecker, Gabriel Gallezot, et al.. Pratiques et usages des outils numériques dans les communautés scientifiques en France. [Rapport de recherche] Comité pour la science ouverte. 2022, 112 p. ⟨hal-03545512⟩

HAL Id: hal-03545512

https://hal-lara.archives-ouvertes.fr/hal-03545512v1

Submitted on 27 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-ND 4.0 - Attribution - No Derivative Works - International License

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Liberté
Égalité
Fraternité

STATE OF OPEN SCIENCE PRACTICES IN FRANCE SOSP-FR



OUVRIR
LA SCIENCE !

Conception graphique et mise en page : Trëma Print EIRL

Except where otherwise noted, this work is licensed under
https://creativecommons.org/licenses/by-nd/4.0/deed.fr

STATE OF OPEN SCIENCE PRACTICES IN FRANCE SOSP-FR

Étude rédigée et conduite par

Mariannig LE BÉCHEC – Pilotage
(MCF HDR, Université Claude Bernard Lyon 1, UR ELICO)
https://orcid.org/0000-0003-3205-2238

Aline BOUCHARD
(Urfist de Paris)

Philippe CHARRIER
(IR HDR, Urfist de Lyon)
https://orcid.org/0000-0003-0285-5216

Claire DENECKER
(Urfist de Lyon)

Gabriel GALLEZOT
(MCF, Université Côte d'Azur, //TransitionS)
https://orcid.org/0000-0002-4443-3153

Stéphanie RENNES
(INRAE, Université de Strasbourg, UMR BETA)
https://orcid.org/0000-0003-1458-7773

Janvier 2022

RÉSUMÉ

L'enquête *State of Open Science Practices* in France (SOSP-FR) a été conduite entre juin 2020 et septembre 2020. Elle a pour but d'interroger les pratiques des outils numériques et autour des données de la recherche dans les communautés scientifiques françaises. Le questionnaire se compose de 38 questions réparties en 9 thématiques. Les questions portent sur des pratiques déjà établies et des pratiques ou usages émergents comme l'*open peer review* ou les articles de données dits *data paper*. Le nombre de répondants est de 1089, permettant d'interroger une répartition disciplinaire, genrée et statutaire assez représentative de l'état de l'emploi dans l'enseignement supérieur et de recherche en France.

Dans l'enquête, le focus sur le contexte de travail des répondants, qualifié de solitaire ou collectif, met en exergue des différences dans les pratiques, notamment d'archivage des données de recherche et dans les usages, particulièrement d'accès à l'information, aux infrastructures de recherche ou aux outils numériques institutionnels. Les réseaux sociaux des chercheurs semblent influencer les pratiques et les usages liés à la science ouverte en France. En distinguant les usages et les pratiques selon deux perspectives, l'une où la science ouverte est associée à une dimension humaine incluant une ouverture au plus grand nombre des résultats de recherche et l'autre où la dimension technique, incluant l'usage d'un environnement numérique libre et gratuit, est celle qui prévaut, les résultats aboutissent à des distinctions disciplinaires mais également statuaires, générationnelles et au niveau du contexte de travail. Les résultats sont équivalents quant à l'arrivée de nouveaux logiciels et langages de programmation, comme nous avons pu le constater avec R, Excel et Python. L'acculturation aux enjeux de la science ouverte passe par des collectifs, plus accessibles dans des environnements de recherche que dans le couple recherche-enseignement.

Selon les principes FAIR, les données de l'enquête sont accessibles via Zenodo au format CSV : https://doi.org/10.5281/zenodo.5827206

ABSTRACT

The State of Open Science Practices in France (SOSP-FR) survey was carried out between June 2020 and September 2020. It aims to question the practices of digital tools and about research data in French scientific communities. The questionnaire consists of 38 questions divided into 9 themes. The questions concern practices already fixed and those emerging uses such as open peer review or data papers. The number of respondents was 1089, making possible a fairly representative disciplinary, gender and status analyse of the state of employment in higher education and research in France.

In this survey, the focus on the respondents' work context, defined as lonely or collective, highlights differences in practices, particularly in the archiving of research data and in the uses, especially in access to information, of research infrastructures or institutional digital tools. Researchers' social networks seem to influence the practices and uses related to Open Science in France. By distinguishing uses and practices according to two approaches, one where Open Science is associated with a human dimension, including opening up research results to the wider range and the other where the technical dimension, including the use of a free digital environment, is what prevails, the results lead to disciplinary distinctions, but also to distinctions in terms of status, generation and work context. The outcomes are equivalent regarding the arrival of new software and programming languages, as we have been able to see with to R, Excel and Python. Acculturation to the challenges of Open Science is achieved through collectives, which are more accessible in research environments than in the research-teaching pair.

SOMMAIRE

Résumé 4
Abstract 5

1. L'ENQUÊTE STATE OF OPEN SCIENCE PRACTICES IN FRANCE – SOSP-FR 7

2. ÉTAT DE L'ART 9

La maîtrise de l'échantillonnage 11
Thématiques récurrentes 12
Thématiques rares 13
Les caractéristiques démographiques des répondants 15
Quelques hypothèses pour l'enquête SOSP-FR 16

3. MÉTHODES 18

4. RÉSULTATS 22

Votre profil et votre environnement de recherche 22
Usages et pratiques des outils numériques en science ouverte 31
Quels sont les logiciels et les langages de programmation utilisés ? 39
Focus sur 3 outils numériques et 3 logiques d'innovation 48
Les données de la recherche 55
Partage des données et des résultats de la recherche 69
Diffusion et valorisation des résultats et des données de la recherche 74

5. DIFFUSION DES RÉSULTATS DES RECHERCHES, DES DONNÉES ET PLACE DU CHERCHEUR DANS LA SOCIÉTÉ 86

Conclusion 99
Bibliographie 103
Index des figures 106
Index des tableaux 108
Annexes 110

1

L'ENQUÊTE STATE OF OPEN SCIENCE PRACTICES IN FRANCE – SOSP-FR

Une lettre de mission nous a été adressée le 16 janvier 2020 par Marin Dacos, de la Direction générale de la recherche et de l'innovation, ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation.

LA QUESTION POSÉE EST : Quelles sont les pratiques des outils numériques et autour des données de la recherche dans les communautés scientifiques françaises ?

Les grandes étapes de l'enquête sont les suivantes :

- État de l'art sur les enquêtes françaises et internationales ;
- Enquête sur les usages relatifs aux outils numériques et aux données de la recherche dans les communautés scientifiques françaises ;
- Passation d'un questionnaire auprès des populations ciblées que sont les membres des universités, des écoles et des EPST français (doctorants, chercheurs, enseignants-chercheurs) ;
- Communication des résultats et mise à disposition des matériaux de l'enquête.

Les livrables attendus sont :

- Étude rédigée ;
- *Executive summary* : traductions (potentiellement assurées avec d'autres moyens) ;
- Poster ;
- Données livrées au MESRI pour alimenter le baromètre de la science ouverte (BSO) ;

- Une restitution publique en France prévue le 22 juin 2022 et une restitution publique dans un événement international à l'étranger[1].

Le calendrier prévu est :

- Étude – 2020
- Résultats intermédiaires : réunion le 27 novembre 2020
- Résultats – 1er semestre 2021 : présentation au SPSO le 28 mai 2021, remise du rapport 22 juillet 2021.

Les membres de l'équipe sont :

- Mariannig LE BÉCHEC – Pilotage
  (MCF HDR, Université Claude Bernard Lyon 1, UR ELICO)
- **Aline BOUCHARD** (Urfist de Paris)
- Philippe CHARRIER (IR HDR, centre Max Weber, URFIST de Lyon)
- **Claire DENECKER** (Urfist de Lyon)
- Gabriel GALLEZOT (MCF, Université Côte d'Azur, //TransitionS)
- Stéphanie RENNES (INRAE, Université de Strasbourg, UMR BETA)

Pour citer cette étude : Le Béchec M., Bouchard A., Denecker C., Charrier P., Gallezot G., Rennes S., étude State of open science practices in France, 2021.

1. La crise sanitaire débutée en 2020 ayant conduit à des confinements qui ont perduré en 2021, il paraît difficile d'envisager une participation à un colloque à l'étranger.

2

ÉTAT DE L'ART

Les dix enquêtes consultées[1] procèdent de considérations diverses tout en s'inspirant explicitement les unes des autres. Elles visent à connaître les pratiques en matière d'accès ouvert et de communication des données, essentiellement vers les chercheurs. Il est sous-entendu que l'ouverture de données se déroule dans le cadre de l'activité de recherche. L'hypothèse d'échanges entre chercheurs par des voies informelles (en dehors du cadre de travail, réunion amicale, échanges de sociabilité) n'est donc pas considérée.

Le territoire enquêté

L'échelle considérée par les dix enquêtes est variable : un établissement universitaire (3), nationale (3) et internationale (4). L'échelle d'un établissement induit une limitation pour raisonner à une échelle plus large. L'échelle internationale pose la difficulté de comparer des pratiques alors que les contextes nationaux de recherches sont souvent très différents. L'échelle nationale permet donc un équilibre.

La population enquêtée

L'exploration locale et limitée à un établissement repose sur des questions de faisabilité et de finalité. La plus grande facilité à mobiliser des chercheurs de sa propre université provient d'un appui institutionnel lors de la passation, car les retombées de l'étude sont potentiellement directes et rendent compte d'une expression des besoins des répondants. Il est difficile de généraliser les résultats, dans un contexte où la recherche se joue dans un périmètre *a minima* national et de plus en plus au niveau international.

Malgré leur intérêt manifeste, les enquêtes étudiées comportent des écueils que nous avons tenté d'éviter dans l'enquête SOSP-FR.

1. Les 10 enquêtes sont listées dans le tableau 1.

RECOMMANDATIONS RETENUES POUR L'ENQUÊTE SOSP-FR SUITE À L'ÉTAT DE L'ART

- Maîtriser l'échantillon des répondants par rapport à une « population mère », afin de représenter la population de recherche du secteur public français
- Limiter les modes de passation qui influent sur les types de répondants et par conséquent sur les résultats
- Développer une autocritique sur la forme du questionnaire
- Respecter les droits afférents à l'usage des logotypes ou des images
- Utiliser une typologie accessible sur les données de recherche
- Poser des questions qui dépassent la dualité obstacles/facilités
- Interroger les valeurs propres à la recherche scientifique et leur présence ou absence de lien avec la science ouverte
- Distinguer questions sur les opinions et questions sur les pratiques *i.e.* les intentions qui ne sont ni des actes ni des réalisations probables
- Aborder le lien entre la pratique de diffusion des résultats et le support de publication adopté pour qualifier la bibliodiversité[2]
- Réutiliser les questions de l'enquête Couperin (2019) sur les *preprints*
- Interroger les nouvelles modalités d'évaluation par les pairs
- Situer les pratiques dans le déroulement de la carrière selon les travaux de l'enquête Harbinger (2018) et distinguer les statuts (permanent, non-permanent)
- Définir les responsabilités au sein de la recherche (direction de projet, direction d'équipe, responsabilités de formation) pour chaque répondant
- Interroger le contexte de travail, donc le réseau social du répondant comme variable de l'accès à l'information, ressources matérielles et dans l'acquisition des nouveaux outils
- Développer des questions démographiques : les réponses ouvertes permettent de recomposer les classes d'âge *a posteriori* et travailler sur des moyennes, écarts-types et médianes afin de repérer des basculements
- Considérer les questions « démographiques » comme de véritables variables qui individuellement ou en corrélation avec d'autres, permettent d'expliquer des comportements, des représentations et des postures différenciées
- Le genre déclaré permet de vérifier la représentativité des femmes
- Poser une question ouverte sur la discipline pour un recodage et atteindre une granularité que ne proposent pas d'autres enquêtes notamment au niveau des SHS

2. Appel de Jussieu, https://jussieucall.org/

Il convient donc de distinguer

- La faisabilité technique de l'étude en 2020, pendant une crise sanitaire;
- L'analyse des représentations (opinions) et les pratiques ou les usages (faits);
- Le contexte d'exercice de la recherche et le statut des répondants;
- Et les stratégies de carrière des répondants.

1. LA MAÎTRISE DE L'ÉCHANTILLONNAGE

Quelle est la « population mère » ?

L'absence d'interrogation sur l'échantillon représentatif signifie que le poids de certains répondants dans les résultats est supérieur à d'autres, jouant ainsi sur les analyses produites. Par maîtrise de l'échantillon, il faut donc entendre le fait que les répondants ne sont pas représentatifs d'une population dite « population mère ».

La représentativité de l'échantillon ?

Les personnes répondant ne sont pas envisagées comme des représentants d'une population plus large, faisant de leurs positions et de leurs réponses, des postures qui pourraient être généralisables. Les répondants montrent un intérêt minimum pour répondre aux questions de l'enquête, par le temps qu'ils y ont accordé au moment de la réception de la demande. Ainsi, il est difficile d'identifier avec assurance des tendances ou plus encore des faits sociaux.

Par exemple, les répondants néerlandais et d'Afrique du Sud semblent surreprésentés dans l'enquête *101 innovations*. Cet exemple illustre la relative absence de considération par les enquêteurs envers la question de la représentativité.

Le mode de passation : diversification, taux de réponse et forme du questionnaire

Le mode de passation influe sur la volonté du répondant à aller au bout de l'enquête tout autant que celle d'y répondre de manière exhaustive. La formulation des questions – parfois longues et surtout complexes – amenuise progressivement l'envie de poursuivre.

Le taux de réponses est variable selon les enquêtes et selon les questions au sein d'une enquête. Par exemple, dans l'enquête *101 innovations* la question des outils spécifiques à chacun des pays a un taux de réponse de 10 % des répondants de l'enquête et un peu moins d'un répondant français sur deux. L'interprétation devient alors discutable.

Concernant la forme, au-delà de la formulation des questions, le recours à l'image, au logo des outils comme dans l'enquête *101 innovations*, permet de mobiliser la mémoire visuelle des répondants. Cependant, cet affichage réduit le nombre de possibilité de réponses et pourrait être uniquement possible pour les questions de faits, notamment pour enregistrer les pratiques des chercheurs.

2. THÉMATIQUES RÉCURRENTES

Les thématiques sont abordées selon deux axes : les pratiques et les représentations (plus précisément l'adhésion) des répondants à la science ouverte, dont l'*open access et l'open data,* par exemple, sont des composantes.

L'exploration ou la recherche de données et de publications

La question de l'usage de moteurs de recherche apparaît principalement dans l'enquête *101 innovations*. Cette pratique d'exploration est inhérente aux travaux de recherche mais semble peu questionnée dans les autres enquêtes.

Les pratiques en matière de données de la recherche

Chercher à définir une donnée de recherche

L'enquête de l'Université de Rennes 2 (2017) pose des questions visant à définir ce que représente une donnée de recherche pour le répondant. La démarche est louable, car la donnée n'a pas qu'une connotation technique, et il paraît important de poser une question visant à déterminer quelles sont les données qui ont une valeur scientifique. Mais cette démarche étant complexe à traiter par un questionnaire, elle est documentée également par le biais d'entretiens semi-directifs.

Définir une typologie des données de recherche

Les enquêtes proposent une ou plusieurs questions sur le type de données que les répondants utilisent ou produisent. Ces questions ne permettent pas la recherche d'une corrélation éventuelle avec les disciplines ou les sous-disciplines, car elles sont déjà définies à partir de pratiques disciplinaires.

Stockage et archivage des données

La question de l'information des chercheurs sur le stockage et l'archivage des données

Les résultats montrent généralement que les chercheurs sont loin d'être au fait des possibilités qui leur sont offertes, même si l'enquête de la *Research Data Alliance* (2019) souligne des avancées. À l'analyse, il apparaît clairement que pour une bonne part des chercheurs, la dimension technique n'est pas l'obstacle le plus important.

La question de l'accès aux infrastructures, logiciels, outils

Au-delà de l'information donnée aux chercheurs, l'hypothèse ne peut être évacuée que les pratiques scientifiques et culturelles font obstacle à l'appropriation. L'appropriation implique un travail d'acculturation probablement plus long dans le temps. Une des hypothèses absentes à travailler est le lien entre la culture professionnelle des chercheurs et la pratique d'archivage propre à des métiers liés à la documentation. Ainsi, se restreindre à des obstacles techniques paraît ne pas rendre compte de l'ensemble du problème, ni même de l'enjeu de la communication des données de recherche.

Diffusion des données de recherche

Les enquêtes analysées cherchent à quantifier les pratiques de diffusion des données et à évaluer le niveau de partage acceptable pour les chercheurs. Sur ce point les résultats montrent que ce qui tranche avec les déclarations d'intention à propos des dispositions à adopter pour cette pratique.

La diffusion des données est souvent corrélée avec la diffusion des résultats et des publications. Or, si les deux sont liés, les enjeux ne sont pas identiques. La diffusion des résultats *via* des articles, des ouvrages, des films par exemple est un enjeu important non seulement pour l'activité scientifique mais également pour la carrière du chercheur. La diffusion des données de recherche sans publication académique peut alimenter par exemple un objectif économique.

Les représentations du chercheur

Les questions sur les représentations portent sans détails sur l'adhésion et les justifications de la part des chercheurs. La focalisation sur les outils permet d'identifier leurs usages mais pas d'appréhender les représentations quant aux enjeux. De plus, des questions attendent des réponses qui font appel conjointement à des faits et des opinions.

La focalisation sur les obstacles, l'inadaptation des outils, évacue les questions sur « ce qui fonctionne bien » et empêche de comprendre pourquoi les chercheurs le considèrent comme positif. Or, connaître les dynamiques vertueuses est tout aussi important, d'autant qu'elles ne sont pas toujours aisées à identifier ni conformes à celles envisagées par les enquêteurs.

Par ailleurs, la connaissance des principes FAIR[3], acronyme récent, anglophone, portée par une vision techniciste des données, ne paraît pas être le meilleur indicateur pour tester l'adhésion à la science ouverte. Or sans avoir la connaissance de cet acronyme, un chercheur peut développer des pratiques de recherche ou adhérer à des valeurs afférentes aux principes FAIR.

3. THÉMATIQUES RARES

Les thématiques suivantes sont moins présentes parfois uniquement dans une seule enquête.

L'évaluation

Il s'agit de comprendre l'opinion des répondants sur le processus d'évaluation des articles par les pairs. Les critiques émergent particulièrement chez les répondants français dans les enquêtes internationales. La question de l'alternative à une évaluation par les pairs est posée.

3. FAIR : acronyme de *Findable, Accessible, Interoperable, Reusable*.

Quelques constats dans les pratiques :

- L'auto-archivage de *preprints* est une possibilité offerte de publier immédiatement sans relecture par les pairs, du moins dans un premier temps ;
- Les communautés scientifiques savent organiser en aval l'évaluation des *preprints* comme les revues traditionnelles, epi-revues, Peer Community In et plus généralement sur les forums, voire sur les plateformes de réseaux sociaux numériques ;
- Les *article processing charges* ou le Gold APC (modèle auteur-payeur) ;
- Les revues prédatrices avec une évaluation payante et sa version dévoyée, par des pseudo-éditeurs malhonnêtes ayant un intérêt personnel qui ne pratiquent pas une évaluation des textes et diffusent le texte dans sa version auteur (Cf. Grudniewicz et al, 2019, pour la définition consensuelle) ;
- L'OA ne veut pas donc dire « sans évaluation par les pairs », le modèle Diamond, les post-publications, et toutes les initiatives d'Open Peer Reviewing montrent le contraire (Tennant et al, 2017).

Pour nombre de chercheurs, l'évaluation par les pairs est gage de qualité et fait partie du mode de fonctionnement de la recherche. L'enquête Couperin (2019) a questionné la position du répondant notamment sur sa participation concrète à l'évaluation par les pairs et montre que tous les chercheurs ne participent pas à ce type d'évaluation et différemment selon les disciplines. L'enquête *101 innovations* montre, quant à elle, que la publication en *open acccess* n'est pas un critère pris en compte dans l'évaluation de la production des chercheurs.

La pratique des *preprints* ou prépublication

Le *preprint* permet de diffuser une version du projet de publication, de prendre date et de valoriser des résultats et des analyses dans des espaces scientifiques concurrentiels. Cette pratique reste à décrire et à documenter, notamment dans le contexte de crise sanitaire dans laquelle notre enquête se développe et selon notamment les modalités mises en place par l'enquête Couperin (2019).

Le besoin d'outils

L'enquête *RDM Survey Delft* (2019) a montré une demande d'outils techniques. La nécessité de formation et de fournir de nouveaux outils se retrouve dans les enquêtes. Notre objectif est d'abord de faire un point sur les pratiques pour envisager ensuite les ressources. Mieux comprendre l'adoption des outils permettra d'identifier les besoins en formation.

Politiques menées en faveur de la science ouverte : focus sur l'*Open access* et de l'*open data*

La difficulté est de qualifier l'acculturation des chercheurs s'ils ne sont pas en charge de projets de recherche pour lesquels les financeurs ont développé des politiques. La prise de responsabilité peut être discriminante au niveau de l'acculturation. De même, des répondants sans connaissance des politiques publiques et des termes spécifiques peuvent développer leurs propres pratiques de Science Ouverte.

4. LES CARACTÉRISTIQUES DÉMOGRAPHIQUES DES RÉPONDANTS

Les caractéristiques démographiques sont aussi des caractéristiques sociologiques. Il faut donc être très attentifs à les recueillir d'une part et à faire en sorte qu'elles soient opérantes et précises d'autre part.

Absence de caractéristiques essentielles

D'un point de vue sociologique, les variables démographiques mobilisées dans les enquêtes analysées sont trop limitées. Ce sujet constitue un écueil important pour les enquêtes internationales, dont les résultats combinent des données issues de plusieurs systèmes de recherche nationaux. Les fonctions mêlent parfois les professeurs et les maîtres de conférences alors qu'il s'agit d'une distinction importante en France, notamment en termes de responsabilités et de moyens d'action.

Les enquêtes proposent des classes d'âge prédéfinies qui ne permettent pas de fortes nuances, car elles se limitent généralement à trois regroupements qui ne sont plus modifiables ensuite.

Des caractéristiques démographiques pour décrire la population et créer des variables

Les caractéristiques démographiques sont majoritairement utilisées à des fins descriptives pour caractériser les répondants. Ces caractéristiques peuvent devenir de véritables ou hypothétiques variables par des tris croisés et permettre de réaliser des analyses plus dynamiques et précises.

La discipline

Toutes les enquêtes posent la question de la discipline d'appartenance, avec plus ou moins de précisions possibles dans les réponses. Les catégories anglo-américaines ne recoupent pas complètement les disciplines françaises. Par exemple, la notion *humanities* n'a pas véritablement de correspondance en France.

L'activité de publication

Caractériser les chercheurs publiant régulièrement de ceux qui le font moins se heurte à de multiples problèmes, comme le fait que les réponses soient déclaratives, mais également que le niveau de publications et la nature de celles-ci ne sont pas équivalentes selon les disciplines.

Le rattachement institutionnel

La plupart des enquêtes demandent le rattachement institutionnel du répondant: université, ou l'organisme de recherche d'appartenance, ou le laboratoire. La réponse ne suffit pas à qualifier le contexte de travail du répondant. Or les conditions d'exercice sont variables et, par exemple, les réseaux de travail du répondant ne sont pas mentionnés alors qu'il s'agit probablement d'un élément qui oriente la capacité à échanger des données de recherche. Ce constat amène à développer cette hypothèse du contexte de travail dans l'enquête SOSP-FR.

5. QUELQUES HYPOTHÈSES POUR L'ENQUÊTE SOSP-FR

Le lien entre le statut/ la carrière du chercheur et l'engagement dans la science ouverte (SO)

L'étude *Early career researchers* (2018) développe une analyse qualitative et se concentre sur les jeunes chercheurs. En substance, l'étude montre une tension assez forte entre une adhésion aux principes de la science ouverte chez les jeunes chercheurs et une pratique qui n'est pas complètement en cohérence, notamment en termes de diffusion des publications et qui se lient à des facteurs institutionnels. La recherche permet de considérer l'hypothèse selon laquelle les pratiques sont en lien avec le statut (permanent, non-permanent) des chercheurs, que ce soit dans le sens d'un plus grand engagement ou bien une certaine résistance, notamment chez les plus jeunes, pour accéder à un poste stable selon les critères d'évaluation en cours dans les universités.

Les conditions de l'activité de recherche

L'environnement du chercheur répondant est absent des enquêtes étudiées. Il serait sans doute intéressant de connaître cet environnement ou le contexte de recherche, puisqu'ils sont susceptibles d'influencer sur les pratiques ou même la diffusion des idées liées à la science ouverte. L'initiation par les pairs doit être qualifiée quant à son influence potentielle sur les pratiques tout comme l'isolement ou la solitude.

L'absence de la recherche privée

Lors de l'acceptation de la mission pour le développement de l'enquête SOSP-FR, le souhait émis était de développer l'analyse des pratiques aux chercheurs du secteur privé, ce qui constitue une nouveauté dans le paysage des enquêtes réalisées jusqu'à présent sur le sujet.

Titre de l'enquête	Porteur de l'enquête	Pays	Portée	Date
Réseaux sociaux de la recherche et *Open Access*. Perception des chercheurs	Consortium Couperin	France	Nationale	2014
Researchers and their data	Université de Vienne	Autriche	Nationale	2015
Les données de la recherche en SHS	Université Lille 3 (GERiiCO)	France	Locale	2015
101 Innovations	Utrecht University Library	Pays-Bas	Internationale	2016
Données de recherche en SHS. Pratiques, représentations et attentes des chercheurs.	URFIST de Rennes 2	France	Locale	2017
Early Career Researchers: the harbingers of change	Publishing Research Consortium	International	Internationale	2018
Les pratiques de publications et d'accès ouvert des chercheurs français	Consortium Couperin	France	Nationale	2019
Open data surveys: how comparable are they and their policy development applications	Research Data Alliance (RDA)	International	Internationale	2019
Données de la recherche : Quelles pratiques ? Quels besoins ?	Cellule Open Access – Aix-Marseille Université	France	Locale	2019
RDM Survey Delft	Delft University of Technology	Pays-Bas	Locale	2019

TABLEAU 1 : Les dix enquêtes consultées pour l'état de l'art

3 MÉTHODES

Notre positionnement

L'objectif est d'obtenir un échantillon de répondants le plus proche possible, sur certains critères, de la « population mère ». Elle correspond à l'ensemble des chercheurs et enseignant-chercheurs exerçant en France soit environ 238 000 personnes en France, 72 000 pour le secteur public (en 2016) et 166 000 pour le secteur privé (en 2015) (MESRI, 2020).

Élaboration du questionnaire

Comprendre les pratiques

Notre difficulté est de définir les outils mobilisés par les communautés scientifiques. Dans l'enquête *101 innovations* les 496 répondants français de cette enquête ne mentionnent pas des outils qui couvrent les questions des données de la recherche. Outre les 5 premiers résultats, les usages mentionnés restent très peu représentatifs de la population avec une mention de multiples outils. La plateforme HAL est utilisée fortement (de 50 % à près d'un tiers) par des chercheurs statutaires (enseignant-chercheur, chercheur, ingénieur de recherche et d'études). Dès que l'on a des répondants moins installés, mais aussi plus jeunes, la part tombe nettement en dessous de 50 %. Quels impacts de l'âge et du statut ?

La question de la date de la première publication sera donc posée dans le questionnaire.

Afin de mieux cerner les usages, 12 chercheurs présents dans notre cercle de connaissances ont été contactés par mail en mars 2020 avec une liste de questions.

Questions posées pour la pré-enquête

Nous souhaiterions, en amont d'une plus vaste enquête par questionnaire, connaître brièvement **quels sont vos outils numériques et vos pratiques** en termes de :

- Recherche d'informations ou d'exploration pour réaliser une recherche :

  .....

- Production des recherches :

  .....

- Réalisation de vos analyses :
- Rédaction de vos publications :
- Choix des formats de publication :
- D'enregistrements de vos données de recherches :
- Diffusions de vos résultats de recherches (partage de données) :
- Valorisation de vos recherches :

Les réponses obtenues mettent en exergue l'hétérogénéité des outils mobilisés, confortant l'option de ne pas définir une liste d'outils (logiciels, langage de programmation) *a priori*.

Constitution du questionnaire

Suite à un travail de classification des outils tirés de l'enquête *101 innovations,* de traitement des réponses des profils de chercheurs, le questionnaire a été rédigé par Philippe Charrier en avril 2020 puis discuté au sein de l'équipe projet le même mois.

Le 27 avril 2020, il a été soumis à discussion lors d'une réunion avec les deux pilotes du **Collège Données de la recherche du Comité pour la Science ouverte (CoSO) :** Véronique Stoll, Directrice de la bibliothèque de l'Observatoire de Paris et Pierre-Yves Arnould, Responsable des systèmes d'information, OSU OTELo (Observatoire Terre Environnement de Lorraine) – CNRS. Suite à ce premier échange, il a été convenu de diffuser ce questionnaire auprès des 14 membres du collège Données de la recherche du Comité pour la Science Ouverte, (chercheurs et professionnels de l'information) durant deux semaines au mois de mai 2020. Nous souhaitons les remercier ici pour leurs nombreux retours, leur éclairage, leurs remarques qui nous ont permis de modifier nos questions ou de confirmer certains de nos choix.

Le questionnaire a été soumis en juin 2020 aux membres commanditaires du MESRI. Nous remercions également ici Marin Dacos pour ses conseils.

Passation

Le questionnaire a été ouvert du 17 juin au 30 septembre 2020 via la plateforme SphynxOnline de la TGIR Human-Num.

Les modes de diffusion ont été multiples via des réseaux de chercheurs (réseau des URFIST, relais des universités ou organismes de recherche et sollicitation d'entreprises privées), via des comptes sur des plateformes comme Twitter (réseau des URFIST), LinkedIn (membres du MESRI) ou les sites web institutionnels (CoSO).

- Les annonces de cette enquête par des comptes institutionnels ont toujours été marquées par une augmentation du nombre de répondants. **Il semble que la voie institutionnelle soit à privilégier pour la passation** ;
- **La déclinaison en local** doit être prise en compte pour la passation de ce type d'enquêtes. Nous avons eu des demandes émanant d'établissements afin de décliner l'enquête en local. La mise en œuvre initialement prévue de se déplacer dans les établissements n'ayant pas été possible en raison de la crise

sanitaire et cette demande en cours de passation obligeant également à revoir notre demande de consentement en début de questionnaire, nous n'avons pas pu répondre favorablement. Afin de faire valider la passation dans des établissements et obtenir l'appui de leur institution pour la passation auprès de leurs chercheurs, nous avons mis à disposition le questionnaire sur ZENODO. Nous avons pu constater que cette mise en ligne a profité à d'autres enquêtes.

Questionnaire disponible via ZENODO

En français :
https://zenodo.org/record/3935958

En anglais :
https://zenodo.org/record/3935971

La durée estimée pour répondre au questionnaire et les contraintes de réponses

Le questionnaire se compose de 9 parties (votre environnement de recherche ; découvrir ; produire des données scientifiques ; écrire ; publier, diffuser, communiquer ; évaluer ; l'environnement de la recherche et de la science ; votre profil) soit 38 questions, dont deux questions ouvertes d'opinion. Le temps moyen de réponse est de 22 mn et 30 s. La médiane est à 16 mn 30 s.

Ne souhaitant pas conserver de données temporaires, type adresse IP, nous n'avons pas laissé la possibilité d'enregistrer les réponses. Ce choix peut entraîner une perte de réponses étant donné la durée d'engagement demandée aux répondants.

Les répondants doivent également répondre à toutes les questions, la dernière exceptée, pour valider le questionnaire.

Nous avons donc pu perdre également des répondants par ce choix.

L'outil utilisé

Nous avons mobilisé les outils de la TGIR Huma-Num et utilisé dans un premier temps SphinxOnline. Suite à un problème technique au lancement de l'enquête ne permettant pas de récupérer les résultats sur le web, nous avons donc récupéré les résultats de l'enquête et procédé à l'analyse sur la version desktop Sphinx IQ2.

La deuxième phase qualitative de l'enquête

Entre septembre 2020 et novembre 2020, 30 entretiens ont été conduits avec des chercheurs ayant laissé leurs coordonnées pour que nous les contactions. Cette question était laissée libre, mais nous avons obtenu plus de réponses qu'attendues. Nous avons sélectionné 40 contacts en variant les disciplines et les statuts. Aucune des réponses au questionnaire n'a été corrélée avec ces entretiens semi-directifs. Cette deuxième phase fait l'objet d'un travail de recherche complémentaire et n'est pas restitué dans cette étude rédigée.

STATE OF OPEN SCIENCE PRACTICES IN FRANCE 2020

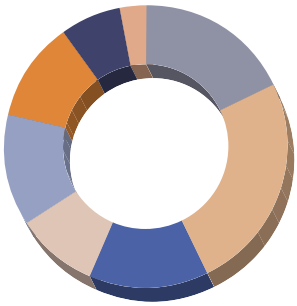
Aline Bouchard
Philippe Charrier
Claire Dénecker
Gabriel Gallezot
Mariannig Le Béchec
Stéphanie Rennes

RÉPARTITION PAR GENRE

FEMME	HOMME	ND
44,1 %	51,9 %	4 %

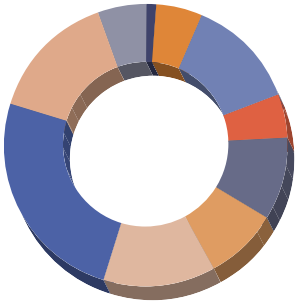
Part féminine conforme répartition MESRI (2020)

FONCTIONS



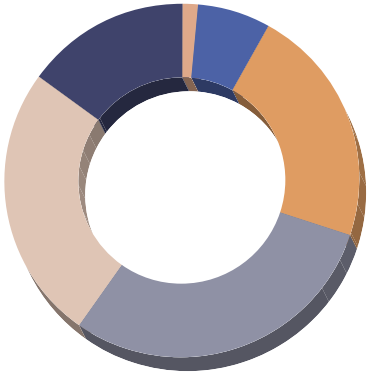
- MCF et assimilés 25,1 %
- Doctorats & CIFRE 17,8 %
- PU et assimilés 13,8 %
- Chargé de recherche 12,9 %
- Ingénieur de recherche 11,5 %
- Directeur de recherche 9,3 %
- Ingénieur d'études 6,9 %
- Chercheurs secteur privé 2,8 %

DISCIPLINES



- Lettres & arts 5,3 %
- Sciences sociales 14,8 %
- Sciences humaines 25,1 %
- sciences du vivant 12,7 %
- Sciences de l'ingénieur 8,4 %
- NR 1,2 %
- Chimie, matériaux 5,4 %
- Mathématiques et informatique 12,8 %
- Médecine 5,0 %
- Physique 9,5 %

ANNÉES DE NAISSANCE



- 1990 et plus 14,97 %
- 1960-1969 21,95 %
- 1980-1989 25,16 %
- 1950-1959 6,61 %
- 1970-1979 29,75 %
- Avant 1950 1,56 %

PASSATION QUESTIONNAIRE

17/06 au 31/09 2020

NOMBRE DE RÉPONSES

1 089

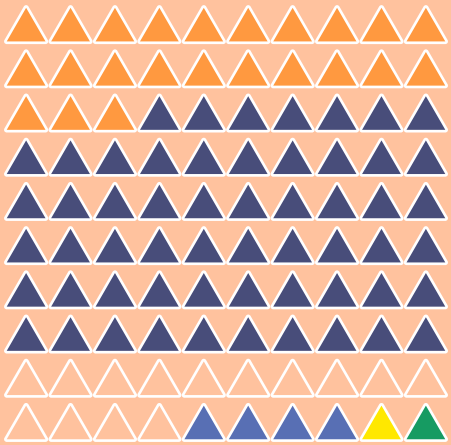
ENTRETIEN SEMI-DIRECTIFS

septembre-novembre 2020

NOMBRE D'ENTRETIENS

29

CONTEXTE DE TRAVAIL



- Seul 23,22 %
- 2 et 5 pers. 57,45 %
- 6 et 10 pers. 13,88 %
- 11 et 20 pers. 3,89 %
- 21 et 50 pers. 1,02 %
- + de 50 pers. 0,56 %

SAUVEGARDE DES DONNÉES (À LONG TERME)

52,9 % OUI

RÉUTILISATION DE DONNÉS

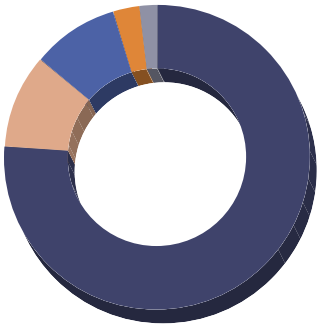
3,5 % TOUJOURS

OUTILS NUMÉRIQUES RÉPERTORIÉS

492

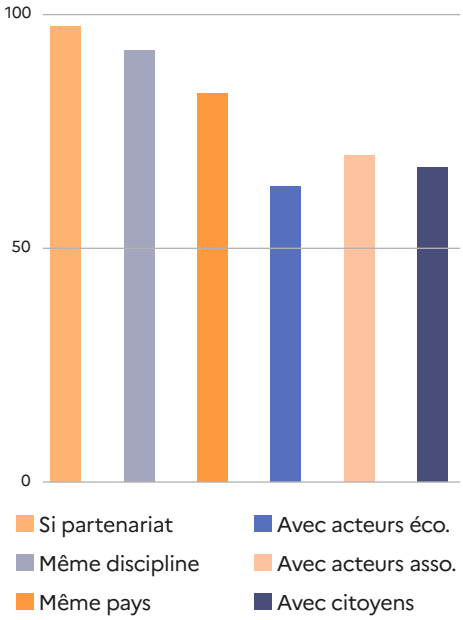
LES PRATIQUES VARIENT-ELLES SELON LE STATUT ?

STATUT PROFESSIONNEL DES RÉPONDANTS



- Permanent 76,3 %
- Non-permament 9,9 %
- Contrat doctoral 9,1 %
- Post-doc 2,8 %
- Indépendants 1,9 %

% OUI AU PARTAGE DE DONNÉES



OUVRIR LA SCIENCE !

RÉSULTATS

4

1. VOTRE PROFIL ET VOTRE ENVIRONNEMENT DE RECHERCHE

Une diversité des disciplines

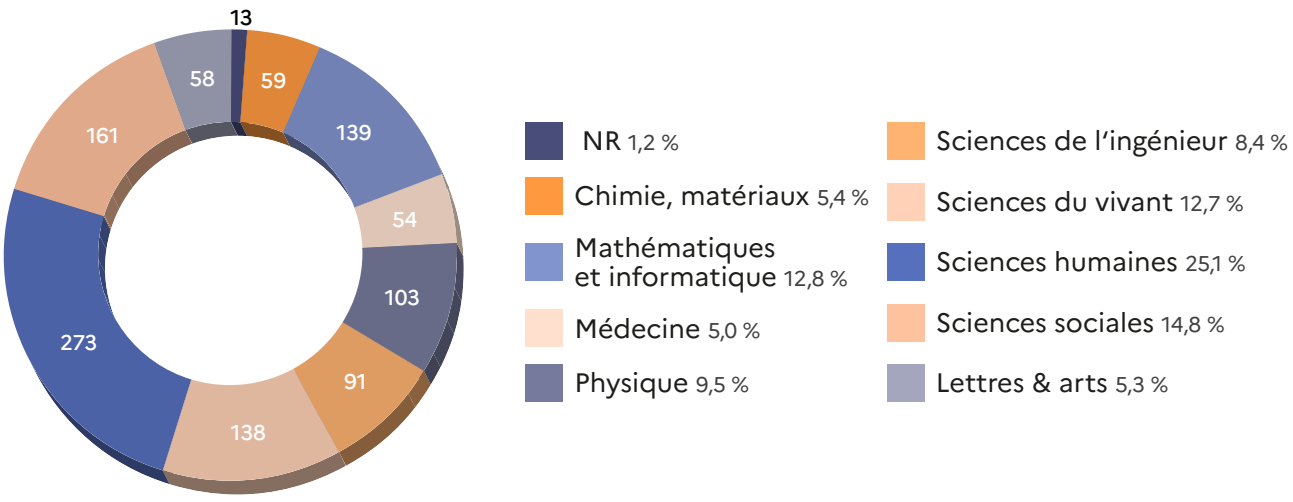


Figure 1: Effectifs des répondants par disciplines (9 catégories)

Près de 250 disciplines différentes ont été déclarées par les répondants et ont été regroupées en 9 catégories utilisées tout au long de ce rapport.

La population répondante (dénommée ci-après « POP ») n'est pas complètement représentative de l'ensemble des chercheurs qui exercent en France. Pour réaliser une comparaison, nous avons eu recours aux données portant sur les effectifs des chercheurs en France fournies par les enquêtes du ministère de la Recherche

et de l'Enseignement Supérieur, précisément, son enquête annuelle sur l'État de l'emploi scientifique en France (2020)[1].

Au regard des écarts, nous n'avons pas procédé à un redressement statistique.

D'une part la discipline n'est qu'une variable sociodémographique parmi d'autres et la redresser aurait signifié que celle-ci prenait une importance *a priori* décisive, d'autre part les écarts statistiques se sont révélés trop importants pour que cette opération n'entraîne pas, elle-même, des biais.

La diversité disciplinaire des répondants est un atout significatif pour nos analyses.

Une comparaison entre notre POP et la nomenclature développée par le MESRI[2] montre quelques différences entre la population répondante et les effectifs théoriques attendus :

Disciplines	% Enquête MESRI	Eff. POP_SOSP	% POP_SOSP	% POP_SOSP (hors doctorants et post-doctorants)
Mathématiques	9,7	139	12,8	13,6
Sciences physiques	8,8	54	5,0	5,2
Chimie	6,7	47	4,3	4,1
Sciences de l'ingénieur 1	9,9	57	5,2	5,3
Sciences de l'ingénieur 2	8,5	46	4,2	4,1
Sciences de la terre/ Environnement	5,5	86	7,9	8,4
Sciences agricoles	0,4	12	1,1	1,3
Sciences biologiques	21,3	87	8,0	8,3
Sciences médicales	6,5	54	5,0	4,5
Sciences sociales	11,1	215	19,7	18,5
Sciences humaines	10,1	221	20,3	20,1
Sûreté, sécurité	0,9	0	0	0
STAPS/Pluridisciplinaire	0,6	58	5,3	5,2
Non réponse		13	1,2	1,4
Total	**100 %**	**1089**	**100 %**	**100 %**

TABLEAU 2 : Comparaison entre effectifs et part des chercheurs entre l'enquête du MESRI et SOSP-FR

1. La tâche est assez complexe, car bien souvent, il y a une distinction entre les chercheurs exerçant dans les grands organismes de recherche (CNRS, INSERM, INRAE, CEA…) et les enseignants-chercheurs exerçant dans les universités ou des grandes écoles. https://www.enseignementsup-recherche.gouv.fr/cid154848/www.enseignementsup-recherche.gouv.fr/cid154848/www.enseignementsup-recherche.gouv.fr/cid154848/l-etat-de-l-emploi-scientifique-en-france-edition-2020.html

2. Cf. Annexes : Tableau 42 ; Tableau 43 ; Tableau 44.

Des écarts notables sont observables :

- Les SHS sont nettement surreprésentées sans doute au détriment des sciences biologiques qui auraient dû être représentées par deux fois plus de répondants ;
- Les sciences de l'ingénieur sont sous-représentées.

La répartition déséquilibrée s'explique en partie par la présence notable des doctorants dans notre enquête et qui ne sont pas comptabilisés dans l'enquête ministérielle (MESRI, 2020). Toutefois, les doctorants et post-doctorants qui ont répondu viennent de disciplines également très diversifiées. En effet, les pourcentages ne varient guère et la présence des SHS et lettres est toujours très importante.

La fonction exercée dans la recherche

Dans la figure 2, seuls 31 chercheurs travaillant dans le secteur privé ont répondu à l'enquête. Cette proportion (2,9 %) ne permet pas une analyse différenciée.

Outre la question de l'accès aux chercheurs du privé, la difficulté à recueillir des données d'enquête issues du secteur privé tient également au caractère confidentiel des activités de recherche réalisées dans un contexte hautement concurrentiel et susceptibles d'être couvertes par le secret des affaires. Ces restrictions affectent directement le statut et les obligations professionnelles des chercheurs concernés. C'est un obstacle que SOSP-FR a rencontré, malgré une diffusion dans des réseaux de la R&D privée (celui de l'ANRT par exemple).

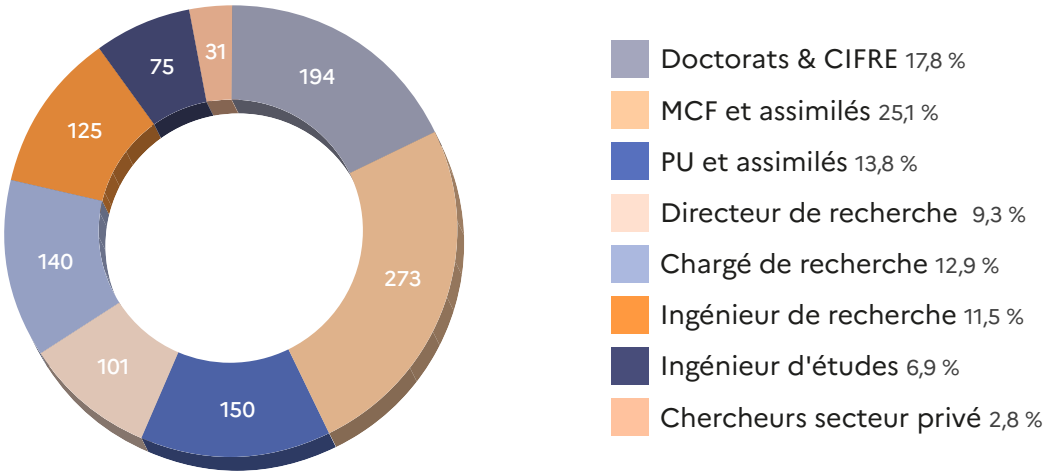


Figure 2 : Fonction déclarée dans la recherche des 1089 répondants

Nous pouvons distinguer 4 sous-populations :

- Les universitaires, maîtres de conférences et professeurs des universités ou assimilés : 423 répondants soit 38,9 % de la POP et 40 % des chercheurs du secteur public ;
- Les chercheurs des grands établissements publics à caractère scientifique et technologique (EPST) ou des établissements publics à caractère industriel et commercial (EPIC) : 241 répondants soit 22,2 % de la POP et 22,8 % des chercheurs du secteur public ;
- Les personnels de soutien à la recherche (ingénieurs de recherche et d'études) travaillant soit dans les universités soit dans les EPST ou EPIC : 200 répondants et 18,4 % de la POP et 18,9 % des chercheurs du secteur public ;

- Les doctorants, au nombre de 194 répondants soit 17,8 % de la POP et 18,3 % des chercheurs du secteur public.

Cette distinction va se retrouver dans l'analyse de l'usage des outils numériques.

Statut professionnel permanent ou non-permanent

HYPOTHÈSE : il existe une différence dans les pratiques selon le statut

Les répondants travaillent dans le secteur public avec des statuts différents (Tableau 3). La fonction dans la recherche ne converge pas nécessairement avec le statut (Cf. Encadré n° 2). Pour les besoins de l'analyse, nous utiliserons régulièrement un regroupement visant à distinguer les répondants qui possèdent un statut permanent de ceux qui n'en bénéficient pas.

STATUT PERMANENT = FONCTIONNAIRE OU ASSIMILÉ + CDI + PROFESSEUR ÉMÉRITE + RETRAITÉ + TRAVAILLEUR INDÉPENDANT

STATUT NON-PERMANENT = CDD + POST-DOC + DOCTORANT NON FINANCÉ + CHÔMAGE + VACATAIRE

Trois types sont distingués

- Type 1 : statut fonctionnaire pour près des deux tiers (65,3 %).
- Type 2 : statut CDI (10 %) : les personnes travaillent dans le secteur privé, ou sont chercheurs dans les grands organismes de recherche.
- Type 3 : statut CDD (8,7 %) : les contractuels, les attachés temporaires d'enseignements et de recherche, les chercheurs, les ingénieurs de recherche et d'étude.

Le statut non-permanent concerne les plus jeunes générations. Les 31-35 ans bénéficient d'un statut stable pour 60 % d'entre eux (contre 78 % pour la POP) et la part se réduit à 10 % pour les 26-30 ans. Le basculement vers la stabilité se fait à partir de 35 ans, même si chez les 36-40 ans, encore 15 % de répondants déclarent un statut non-permanent.

La place des doctorants dans l'enquête SOSP-FR

Avec près de 200 répondants, les doctorants sont bien représentés dans notre échantillon. En 2018, le MESRI comptabilise 176 840 chercheurs et personnels de soutien à la recherche, tout type d'établissement confondu, pour 71 200 doctorants inscrits en 2018 (toutes disciplines et années confondues). Nous considérons que les doctorants participent à la recherche, qu'ils sont également un vecteur important des usages des outils techniques destinés à la recherche. Ainsi les doctorants sont intégrés aux effectifs totaux des chercheurs actifs et nous obtenons donc un total de 247 214 chercheurs dans le secteur public. Dans ces conditions, les doctorants représentent approximativement 29 % des personnes ayant une activité de recherche, nettement plus que la part qu'ils représentent dans notre échantillon (17,8 %). Dès lors, les doctorants qui ont répondu à notre enquête sont loin d'être surreprésentés. Il est même possible d'en déduire qu'ils ne sont pas assez nombreux à avoir répondu au questionnaire.

Si la fonction de doctorant est homogène, le statut ne l'est pas. Les doctorants peuvent bénéficier d'un financement lors de leur parcours de thèse, comme ne pas être financés. Nous ventilons la sous-population

des doctorants en fonction de leur statut : statut spécifique si contrat doctoral d'une durée limitée à 3 ans ou un équivalent et statut non-permanent si le financement n'est pas stable durant leur parcours doctoral. Ces derniers seront classés parmi les chercheurs au statut non-permanent, où sont comptabilisés les vacataires, les chômeurs et le personnel en CDD.

	Nombre	% Obs.
Fonctionnaire ou assimilé	711	65,30 %
En CDI	109	10,00 %
Contrat doctoral	99	9,10 %
En CDD	95	8,70 %
Post-doc	30	2,80 %
Travailleur indépendant	20	1,80 %
Doctorant non financé	7	0,60 %
Chômage	6	0,60 %
Retraité	5	0,50 %
Professeur/Chercheur émérite	3	0,30 %
Autre	2	0,20 %
Vacataire	2	0,20 %
Chercheur associé non rémunéré	0	0,00 %
Total	**1089**	**100,00 %**

TABLEAU 3 : Statut professionnel des répondants

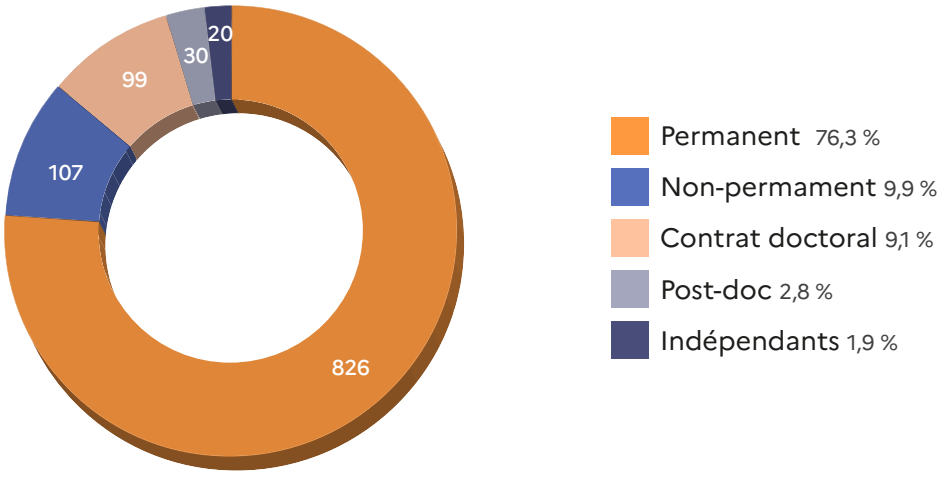


Figure 3 : Répartition statutaire agrégée

La répartition par genre et par âge

Genre déclaré	Pourcentage
Homme	51,9 %
Femme	44,1 %
Je ne souhaite pas répondre	4 %

TABLEAU 4 : Genre déclaré par le répondant

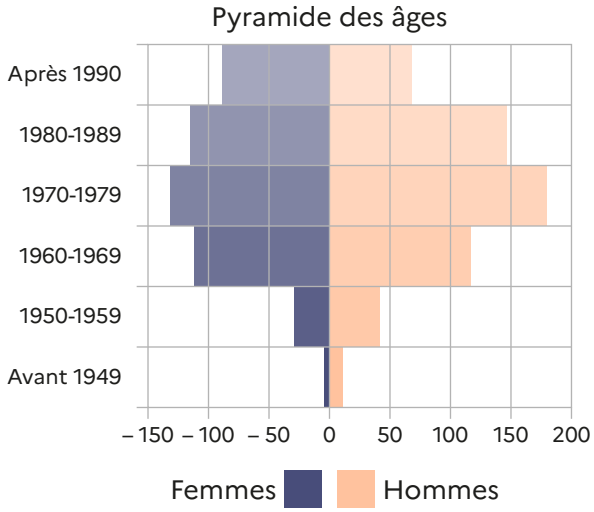


Figure 4 : Année de naissance déclarée répartie par genre

La présence féminine est conforme à la part des femmes dans le secteur public de la recherche qui est de 43,8 % en 2017[3]. Reste que 4 % des répondants n'ont pas souhaité répondre à la question, ce qui ne permet pas de déterminer si la part féminine pourrait être légèrement sous-représentée ou même sur-représentée, car rien n'indique que ces répondants se répartissent harmonieusement entre homme et femme. Notre échantillon ne pâtit pas en représentativité sur la question du genre. Parmi les caractéristiques sociodémographiques la question de l'âge du répondant peut avoir un impact sur les réponses et les opinions. Nous avons préféré raisonner indirectement en demandant au répondant de renseigner son année de naissance, ce qui permet *a posteriori* de construire différentes classes d'âge, ce que n'autorise pas une question fermée proposant au répondant de se situer dans une classe d'âge prédéfinie.

Les classes d'âge et le basculement entre statut non-permanent et permanent

Parmi nos 1 089 répondants, il y a 285 répondants (26,2 % de l'échantillon) de « jeunes chercheurs », c'est-à-dire de moins de 35 ans (générations > 1985). Rappelons que l'âge moyen de recrutement dans l'enseignement supérieur (MCF) est de 34,5 en 2018, de 32,6 ans pour les chargés de recherche 2e classe des 5 principaux EPST et 33 ans pour les ingénieurs et cadres non confirmés des 8 EPIC et ISBL[4].

Les répondants nés à partir de 1990 (30 ans et moins) sont à près de 40 % dans des statuts non-permanents et seulement 1/10 exercent avec un statut permanent, la majeure partie d'entre eux bénéficiant d'un contrat doctoral (près de 50 %). La classe d'âge supérieure marque un basculement conséquent, près de 58 % ayant accédé à un statut permanent. La stabilité statutaire est une norme peu contestée (80,8 %) pour les plus de 35 ans.

RECOMMANDATION : maintenir la distinction statut permanent, non-permanent et contrat doctoral

3. L'état de l'Emploi scientifique en France – Rapport 2017. MESRI.
4. EPST : Établissement Public Scientifique et Technique ; ISBL : Institution Sans But Lucratif. L'état de l'Emploi scientifique en France – Rapport 2020. MESRI, pp. 18-19.

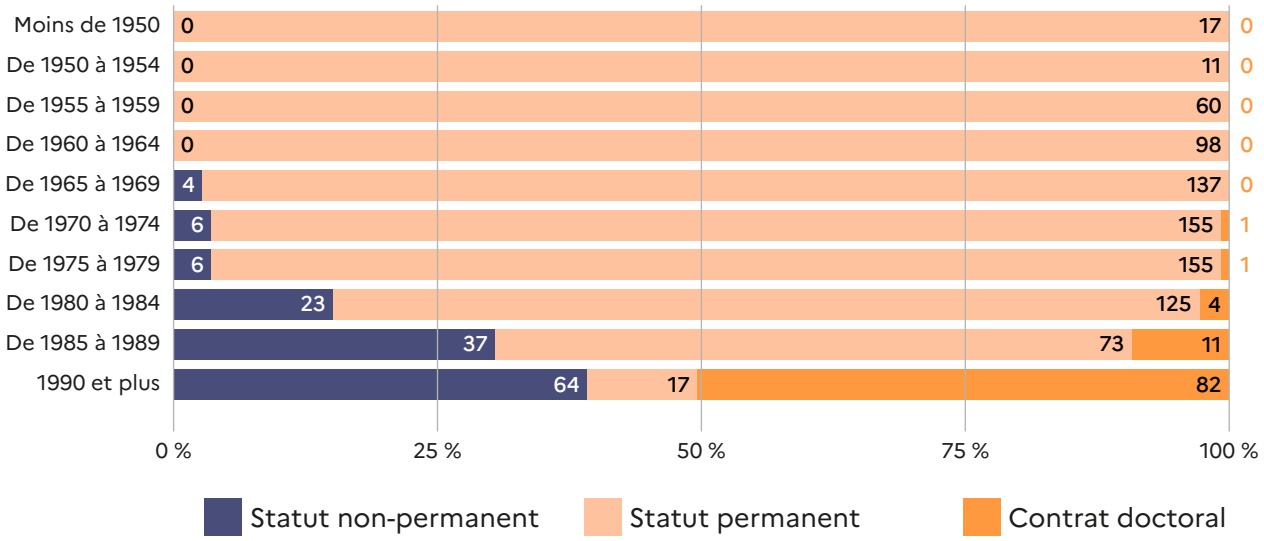


Figure 5: Statut professionnel agrégé en fonction de l'année de naissance

Le contexte de travail

HYPOTHÈSE: le contexte de travail influe sur l'accès à l'information et aux infrastructures de recherche.

QUESTION 2: Habituellement, pour une opération de recherche, avec combien de personnes travaillez-vous (chercheurs, techniciens, ingénieurs compris)?

Il s'agit de distinguer deux types d'activité de recherche par le contexte de travail pour les répondants: solitaire ou collective.

L'indicateur permet de caractériser dans une perspective de management des connaissances, les conditions de réalisation de l'activité de recherche du répondant et les ressources sociales et techniques disponibles.

Le contexte de travail des répondants est majoritairement collectif (76,2 %). Le travail collectif se réalise surtout dans des petites équipes (de 2 à 5 personnes), le cas des opérations de recherche mobilisant plus de 20 personnes étant relativement rare (1,6 %). Près d'un quart des répondants travaillent seul.

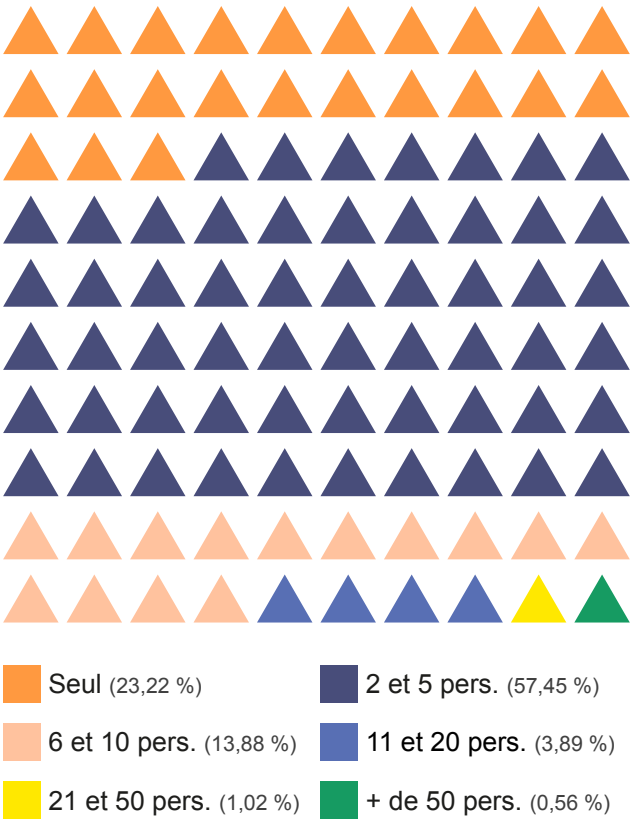


Figure 6: Contexte de travail par le nombre de personnes participant à une opération de recherche

Ce contexte de travail est dépendant de trois éléments (l'âge, la discipline et la fonction dans la recherche) :

Âge	▸ Il est significatif qu'un peu plus de 30 % des jeunes chercheurs (≤ 35 ans) travaillent dans un contexte solitaire. ▸ Le travail solitaire se réduit progressivement dans les autres classes d'âge pour augmenter à nouveau pour les chercheurs les plus âgés (> 60 ans). ▸ La présence de chercheurs en fin de carrière, voire au statut émérite, peut laisser envisager un éloignement des espaces collectifs de travail comme les laboratoires.
Discipline	▸ Pour la médecine, la chimie, la physique, mathématiques et informatiques et les sciences du vivant, la part des chercheurs qui travaillent seuls est réduite (entre 3,4 et 9,3 %). ▸ Au sein des sciences de l'ingénieur, le travail solitaire est assez faible sans être rare (17,6 %). ▸ Dans l'ensemble des lettres, arts et SHS, la solitude semble être un mode d'activité en tant que tel, plus encore pour les sciences humaines (60,3 %) que les sciences sociales (environ 40 %).
Fonction dans la recherche	▸ Les doctorants sont les plus isolés (près de 46 %), suivis des universitaires (plus chez les MCF, 28,1 % que les professeurs, 24,2 %). ▸ Les chercheurs des organismes de recherche (19,3 %) et les directeurs de recherche (6,9%) travaillent peu seuls. Ils connaissent des contextes de travail plus collectifs, avec des collectifs plus larges pour les seconds. ▸ Les personnels d'encadrement sont ceux qui travaillent le plus collectivement, ce qui est inclus dans leur fonction. ▸ Il existe une fracture entre universitaires et chercheurs d'organismes de recherche. Son identification permet de mieux prendre en compte les modalités de circulation de l'information autour des enjeux et des pratiques liées à la science ouverte.

TABLEAU 5 : Tableau de comparaison du contexte de travail selon trois variables (âge, discipline, fonction)

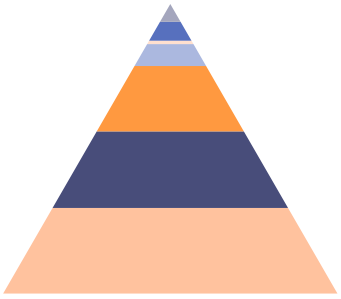
RECOMMANDATION : Il sera intéressant d'observer si l'accès à l'information et les usages des infrastructures de recherche progressent, notamment pour les personnels des universités dans le développement de politiques publiques en faveur de la science ouverte

STATE OF OPEN SCIENCE PRACTICES IN FRANCE 2020

Aline Bouchard
Philippe Charrier
Claire Dénecker
Gabriel Gallezot
Mariannig Le Béchec
Stéphanie Rennes

USAGES ET PRATIQUES

Sources d'information sur les outils numériques



- Par des collègues, des amis 29,55 %
- En cherchant par moi-même 26,25 %
- Par mon institution, mon entreprise 22,94 %
- Au cours d'une formation 7,7 %
- Par la veille sur ces questions 7,63 %
- Via les réseaux sociaux numériques 5,77 %
- Par démarchage commercial 0,16 %

Focus sur 3 outils et 3 logiques d'innovation (Alter, 2015)

- **R, la voie de l'institutionnalisation**

244 (usagers) / 1 089 (répondants)

- **Excel, le logiciel institutionnalisé**

369 / 1 089

- **Python, une pratique émergente**

170 / 1 089

PARTAGE ET DIFFUSION

DES DONNÉES AVEC LES CHERCHEURS

- entre chercheurs avec qui j'ai un partenariat (97,8 % oui)
- entre chercheurs de la même discipline ou du même domaine (92,20 % oui)
- entre chercheurs du même pays ou du même continent (83,10 % oui)

DES DONNÉES AVEC LA SOCIÉTÉ CIVILES

- avec des membres du domaine économique et médiatique (63,6 % oui)
- avec des membres du secteur associatif et / ou caritatif (70,2 % oui)
- avec tout citoyen, sans restriction professionnelle ni géographique (67,4 % oui)

DIFFUSION ET VALORISATION

arXiv

20,20 % d'usage fréquent sur les plateformes « *not for profit* »

des variations disciplinaires dans une pratique limitée

zenodo

Usage fréquent

Supplementary materials 11,3 % > data papers 2,8 %

Entrepôts disciplinaires 5 % > entrepôts généralistes 3,4 %

des SHS en retrait face à des pratiques dépendantes des modèles d'édition des revues

31,9 % d'usage déclaré de GitHub, Git, Software Heritage, ...)

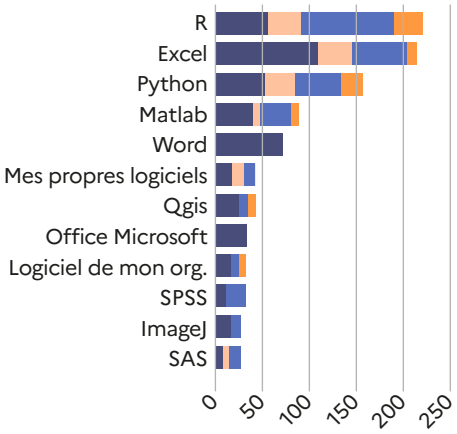
quelle reconnaissance de ces pratiques émergentes mais mobilisées

OUVRIR LA SCIENCE !

PRODUCTION DE DONNÉES DE RECHERCHE

492 LOGICIELS ET LANGAGES

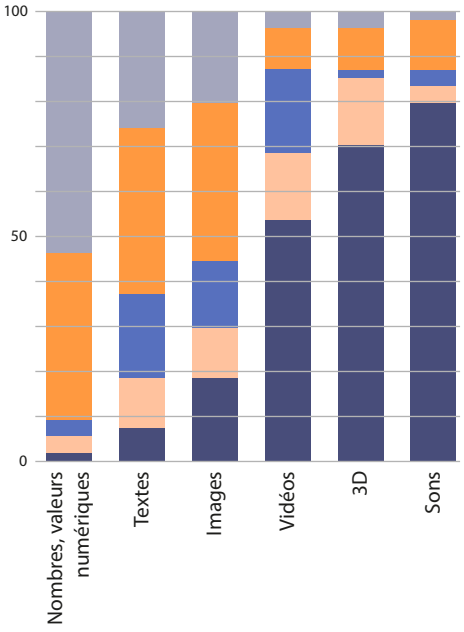
Le contexte de travail marque un clivage dans les usages. L'appartenance à l'université ou aux grands organismes de recherche explique les usages différenciés des outils de production des données de recherche.

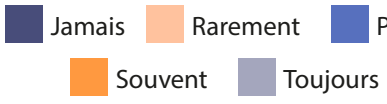




LES DONNÉES DE RECHERCHE PRODUITES SELON LA DISCIPLINE

- Sons : sciences sociales
- Données 3D : disciplines littéraires et artistiques
- Données images : l'usage fréquent concerne entre 15 et 25 % des répondants d'une discipline et ne dépend pas d'une discipline.





2. USAGES ET PRATIQUES DES OUTILS NUMÉRIQUES EN SCIENCE OUVERTE

Les systèmes d'exploitation

HYPOTHÈSE : Linux est un système d'exploitation plus propice à recevoir des logiciels libres, nous soumettons cette population à l'hypothèse que ces répondants sont plus sensibles aux enjeux de la science ouverte.

Le système d'exploitation peut signaler des postures particulières envers les outils numériques. Lorsque l'on s'éloigne de l'usage normatif du système d'exploitation Windows, et que nous recensons dans cette étude, la catégorie de chercheur qui utilise résolument les systèmes Linux ou d'autres Unix pour son association avec les logiciels dits libres à l'image des « libristes » qu'a distingués Célya Gruson-Daniel (2018 : 261) « pour désigner les personnes qui s'attachent à l'éthique et à la culture "libre" ».

Les systèmes d'exploitation utilisés par les répondants placent Windows comme le plus fréquent (63,9 %). Il est suivi à la fois par l'environnement Macintosh (31,2 %) et Linux (29,6 %). Les autres systèmes (autres Unix ou autres OS, comme Android…) sont rares et ne concernent que 6,5 % des répondants. Toutefois, une part importante des répondants use de plusieurs systèmes, ce qui relativise le poids du système Windows.

Ce raisonnement permet de pondérer la notoriété des systèmes. En distinguant les différentes configurations possibles en termes d'usage de système d'exploitation, nous constatons que l'usage exclusif pour les trois principaux OS est plus marqué pour Windows (68 % des utilisateurs de Windows le font en exclusivité) que pour Macintosh et surtout Linux, ce dernier étant majoritairement un OS que l'on utilise en complément d'un autre (45 % d'exclusivité). Parmi les combinaisons possibles, c'est celle entre Windows et Linux qui apparaît la plus fréquente, avec 7,5 % des répondants qui fonctionnent avec ces deux OS parallèlement.

Système d'exploitation (OS)	Effectif	% (/POP)	Rapport exclusivité/Usage total
Windows exclusivement	476	43,7 %	0,68
MacOS exclusivement	186	17,1 %	0,55
Linux exclusivement	146	13,4 %	0,45
Autres OS exclusivement	3	0,3 %	0,05
Autres Unix exclusivement	0	0,0 %	0
Association Windows/Mac	67	6,2 %	
Association Windows/Linux	82	7,5 %	
Association Linux/Mac	38	3,5 %	
Total	**1089**	**88,2 %**	

TABLEAU 6 : Répartition des différentes configurations en termes d'OS et leur poids parmi la population répondante

RECOMMANDATION : La science ouverte dans le cadre notamment des enjeux de la reproductibilité des résultats induit une diffusion des résultats et une exploitation des outils numériques d'analyse, il est donc intéressant de suivre la part d'usage des OS.

L'usage du téléphone à des fins de recherche

RÉSULTATS : L'usage du téléphone portable ou mobile à des fins de recherche est pratiqué par une minorité des chercheurs, soit 15,8 % pour « souvent » et « toujours » et 21,6 % pour « parfois ». Cet usage n'est pas ancré.

Les outils de gestion du travail

HYPOTHÈSE : La pandémie Covid-19 a modifié les usages des outils de gestion de travail.

QUESTION 8 : AVANT LE CONFINEMENT, quels types d'outils de gestion de travail individuel et collectif utilisiez-vous ?

Nous avons interrogé les répondants sur leurs usages en matière d'outils de gestion du travail scientifique, en distinguant la période antérieure au premier confinement[5] et celle qui la suit **(QUESTION 9)**.

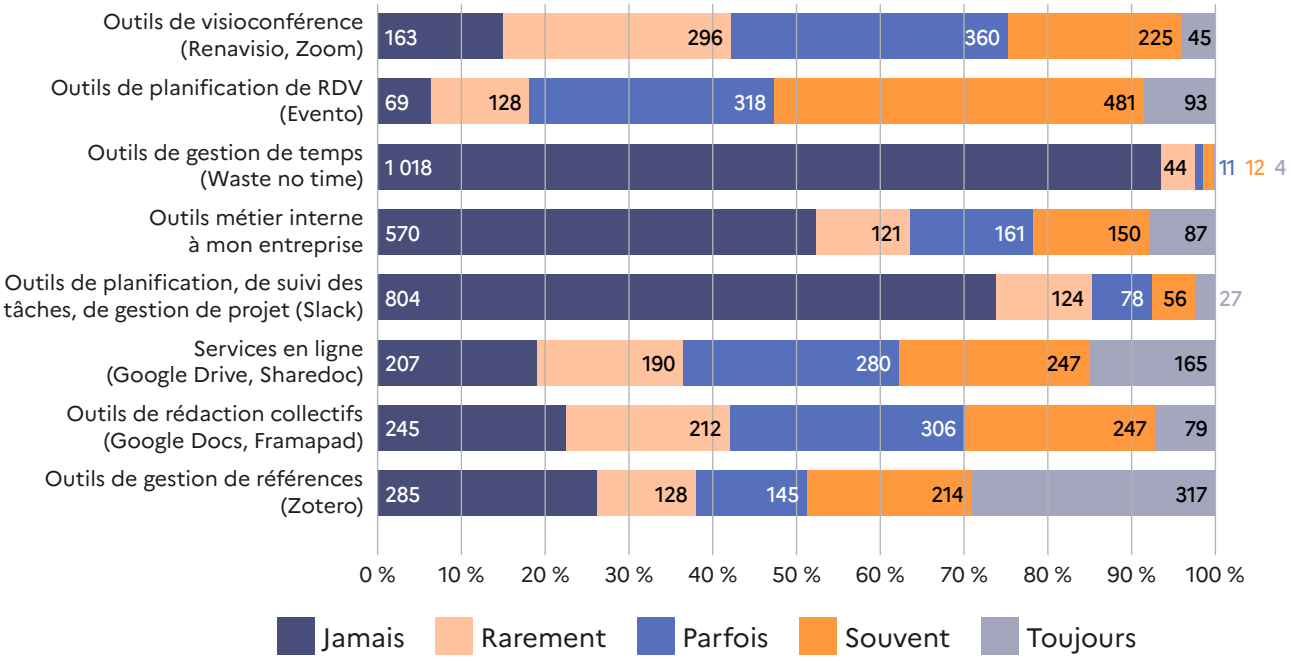


Figure 7 : Usage des outils de gestion du travail avant le confinement

Les pratiques des chercheurs en cette période de crise ont convergé vers deux attentes : échanger et écrire collectivement, sans changements notables dans les pratiques de travail.

5. Du 17 mars au 11 mai 2020.

Avant la situation de confinement, les outils de planification de rendez-vous sont utilisés fréquemment par plus d'un chercheur sur deux. Les outils de gestion de références semblent être entrés dans les pratiques scientifiques, ainsi que les services en ligne, et les outils de rédaction collectifs rassemblent un tiers de répondants pratiquants. À l'inverse, les usages des outils de visioconférence, de planification et de gestion de projet, de gestion de temps demeurent peu répandus.

La situation de confinement du premier semestre 2020 n'a pas fait considérablement évoluer les usages à deux exceptions près. D'une part, la visioconférence a véritablement progressé passant de 25 % d'usage fréquent (« souvent » et « toujours ») à 79,2 %.

D'autre part, de manière moins marquée, les outils de rédaction collectif sont plus sollicités par les répondants, puisqu'ils ne concernaient fréquemment que de 30 % d'entre eux alors qu'à la suite du confinement cette part est passée à 36,4 %.

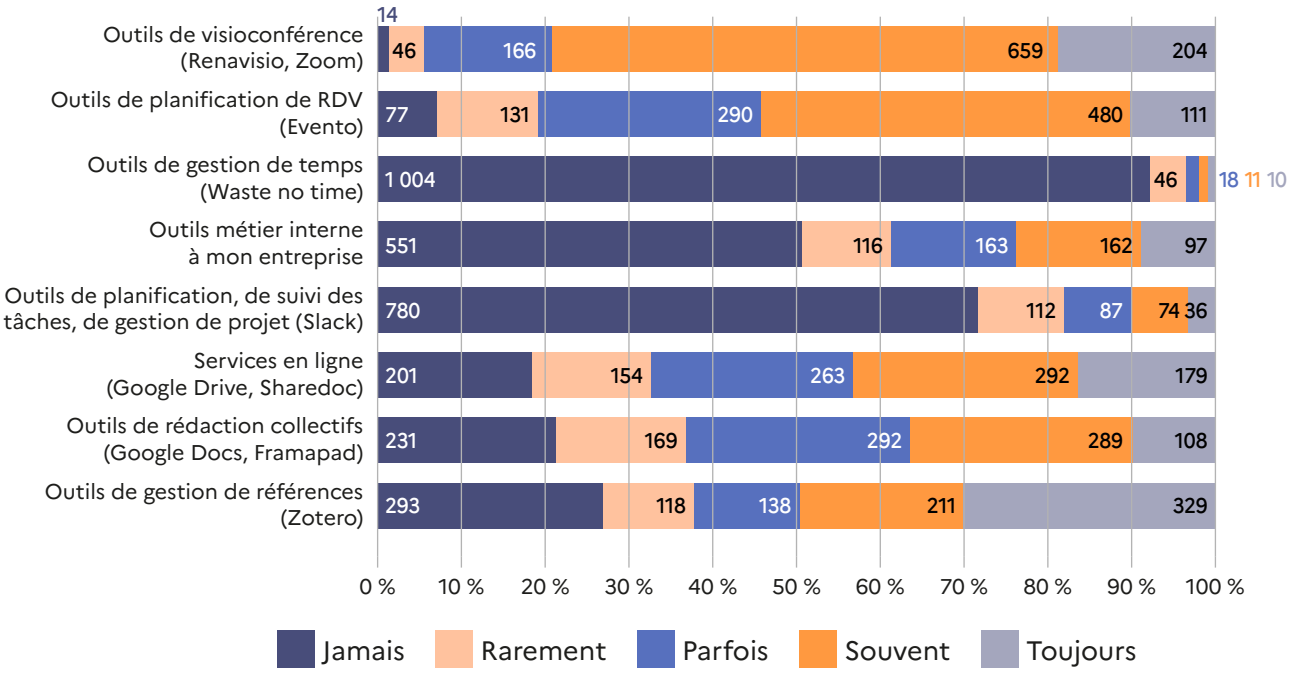


Figure 8: Usage des outils de gestion du travail après le confinement

RECOMMANDATION: L'augmentation des usages des outils de visioconférence et de rédaction collective doit être vérifiée ultérieurement pour savoir s'il s'agit là d'un phénomène conjoncturel ou de l'amorce de nouvelles pratiques numériques.

Découvrir – les canaux d'accès à l'information scientifique

QUESTION 10: Comment accédez-vous aux informations utiles à vos recherches ?

Les outils de collecte de l'information nécessaires sont multiples. Dans la figure 9, les canaux d'accès aux informations peuvent être regroupées :

- 4 modes d'accès semblent privilégiés : les accès institutionnels (59,6 % d'usage fréquent) ; les moteurs de recherche

scientifiques (62,6 % d'usage fréquent) ; les moteurs de recherche généralistes (62 % d'usage fréquent) et dans une moindre mesure, les archives ouvertes (46,9 % d'usage fréquent) ;

- Les bibliothèques « clandestines », les bibliothèques numériques et les réseaux sociaux académiques recueillent entre 20 et 27 % d'usage fréquent. Ces modes d'accès viennent sans doute en appui lorsque le chercheur connaît des difficultés pour obtenir l'information souhaitée (Boudry et al., 2019) ;
- Le « *pay per view* », les demandes sur le web, les extensions sur un navigateur et enfin l'usage de plateformes dites « entrepôt de données » ne sont pas utilisés par près de 8 chercheurs sur 10 ;

La sollicitation directe de l'auteur d'une publication n'a pas une fréquence très importante (7,7 % d'usage fréquent) mais à peine un quart des répondants déclarent ne jamais le faire, plus des ⅔ d'entre eux (68,3 %) le faisant avec parcimonie. Cette pratique vient pallier l'absence de solution des autres modes d'accès.

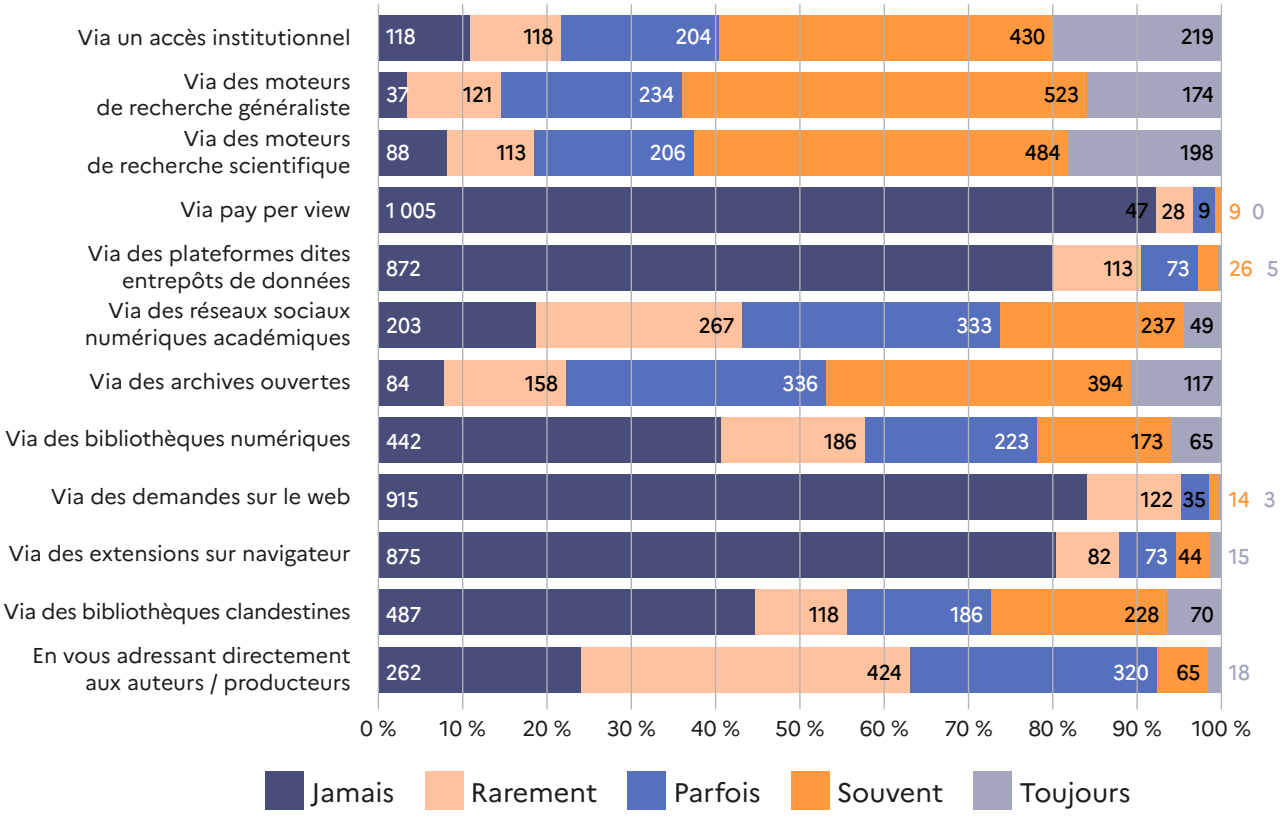


Figure 9 : Canaux d'accès à l'information scientifique

Dans les résultats, les bibliothèques clandestines et les réseaux sociaux académiques sont les résultats les plus originaux. Les 128 répondants se caractérisent par le fait de travailler nettement plus souvent seul (46,9 % contre 23,8 %), par une plus forte représentation des femmes (53,9 % contre 44,1 %), avec une plus forte proportion de doctorants (28,7 % contre 17,8 % pour la POP) et de professeurs d'université (20,3 contre 13,7 %). La propension à travailler seul est une variable explicative des pratiques alternatives numériques, dans le sens où dans la population générale, les femmes

qui travaillent isolément représentent 31,3 % des chercheurs alors que leurs homologues masculins pèsent seulement pour 17 % des chercheurs. Ces pratiques informationnelles que nous avons nommées stratégies alternatives sont développées par des chercheurs plus isolés des moyens et des voies institutionnelles de découverte de l'information.

RECOMMANDATION : Ces pratiques paraissent conformes aux autres enquêtes. Ces questions pourraient ne pas être reconduites.

Les réseaux sociaux comme sources de connaissance des outils numériques utilisés

HYPOTHÈSE : Le contexte de travail influe sur l'accès à l'information en matière d'outils numériques.

Par quel canal les chercheurs ont-ils découvert les outils numériques qu'ils utilisent au moment de l'enquête ? L'objectif est d'évaluer la part de l'information obtenue par les pairs ou le cercle amical, et la part émanant des ressources institutionnelles ou issue des temps de formation.

Les résultats sont en faveur du cercle amical et celui des collègues. Les usages des outils numériques des chercheurs se diffusent avant tout de bouche-à-oreille, les autres canaux de communication étant secondaires.

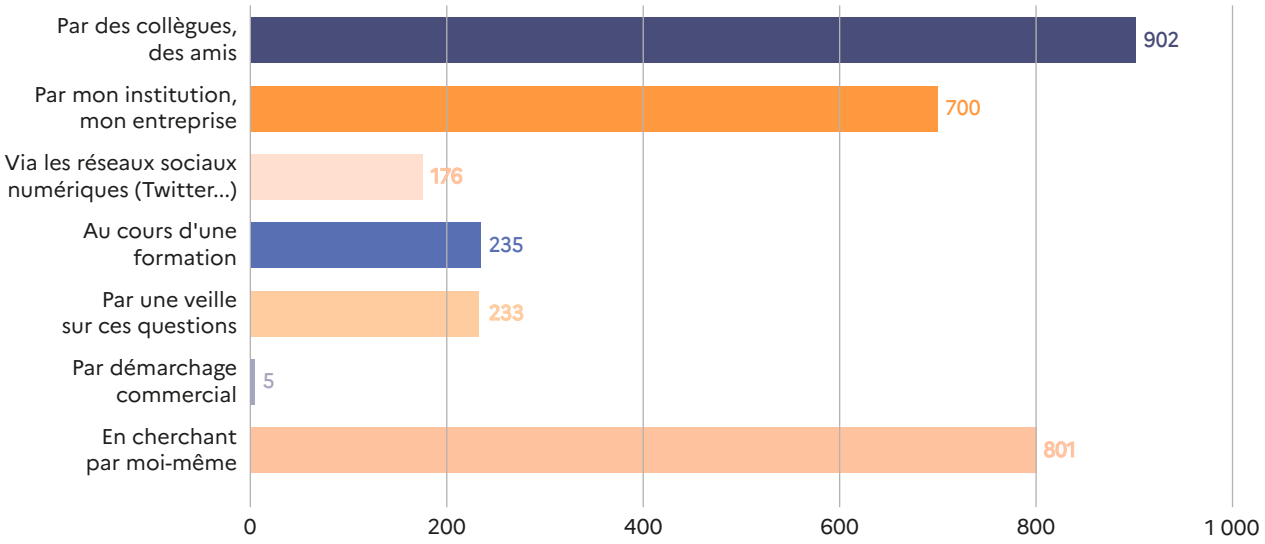


Figure 10 : Les canaux de découverte des outils numériques

Une diversité des canaux d'information

Sachant que les répondants pouvaient cocher plusieurs réponses, en moyenne ils ont signalé près de 3 modalités (2,8 précisément), ce qui marque une diversité dans les canaux d'information des outils numériques. Les collègues, les amis ou par une recherche personnelle sont le moyen de communication le plus plébiscité (82,8 %). La découverte *via* les ressources institutionnelles est le fait de 64,3 % des répondants. Les répondants privilégient avant tout comme source d'information leur réseau social de proximité au-delà de leurs capacités d'accès à l'information (moi-même, les collègues, mon laboratoire ou mon université). Les autres modalités semblent relever de solutions secondaires : les formations

(21,6%), la veille (21,4%), les réseaux sociaux numériques (16,2%). L'accès par un démarchage commercial est quant à lui rare (0,5%).

Le réseau social influe sur les réponses. Si le chercheur bénéficie de réseaux de collégialité et amicaux solides, d'un soutien institutionnel, il est probable que le chercheur sera plus au fait des outils numériques qui lui conviennent, puisqu'il bénéficiera de plusieurs canaux d'information.

Le contexte de travail et le poids de la proximité

À de nombreuses reprises, l'impact de l'indicateur de contexte du travail, construit autour de la question du nombre de personnes avec lequel le répondant travaille dans son activité de recherche est souligné. Il faut préciser sans doute que cet indicateur permet indirectement de connaître, même si c'est de manière limitée, les conditions de réalisation de l'activité de recherche du répondant mais également les ressources sociales disponibles. Car travailler seul ou collectivement implique des contextes relationnels variables. La question de l'usage des outils numériques est liée à celle de l'information disponible et des capacités d'accès à cette information. Le partage de « bonnes pratiques », des codes sources peut servir à alimenter cette accessibilité de l'information. Pour ces raisons, nous avons choisi d'analyser plus profondément les résultats bruts au regard de cet indicateur du contexte de travail.

	Travail seul	Entre 2 et 5 pers.	Entre 6 et 10 pers.	Entre 11 et 20 pers.	+ de 20 personnes
	% Obs.	% Obs.	% Obs.	% Obs.	% Obs.
Par des collègues, des amis	76,80%	85,20%	82,70%	83,30%	88,20%
Par mon institution, mon entreprise	57,50%	64,70%	70,70%	66,70%	88,20%
Via les réseaux sociaux numériques (Twitter…)	17,80%	15,80%	15,30%	16,70%	11,80%
Au cours d'une formation	25,10%	20,50%	20,00%	21,40%	23,50%
Par une veille sur ces questions	22,00%	20,10%	26,00%	14,30%	35,30%
Par démarchage commercial	0,40%	0,30%	0,70%	2,40%	0,00%
En cherchant par moi-même	82,20%	71,70%	68,70%	59,50%	88,20%

TABLEAU 7 : Répartition des sources de connaissance des outils selon le contexte de travail

Les relations ne sont pas significatives. Mais le premier canal d'information sur les outils techniques est le cercle amical ou des collègues. Si la plupart des répondants affirment qu'ils doivent la connaissance des outils numériques à un collègue ou un ami (82,8%), ce chiffre n'est plus que 76,8% pour ceux qui travaillent seuls. Ces derniers ont également moins recours à leur institution (57,5%). Ce résultat est indispensable à prendre en compte dans un objectif de diffusion des pratiques et des outils liés à la science ouverte.

RECOMMANDATION : cibler la communication de ces outils en fonction du contexte de travail (individuel ou collectif) peut induire une plus grande capacité à connaître des outils numériques et à les relayer auprès des pairs.

Produire des données de recherche

La mobilisation des outils numériques pour la recherche

Cette partie s'intéresse aux logiciels afin de distinguer les usages des outils payants, libres ou partiellement gratuits et les fonctions qu'ils remplissent. Les fonctions étudiées sont : production, analyse, nettoyage, visualisation, que nous détaillons dans la partie suivante.

Préférence en matière de choix de logiciels

Au sujet des logiciels de production de données, nous avons demandé quels types de logiciels étaient utilisés. L'objectif est de savoir vers où se situe leur préférence, sachant que nous demandons de hiérarchiser au mieux 5 réponses.

Logiciels	Nombre	% Obs.	Imp.
Libres et gratuits	757	69,50 %	2,97
Payants	679	62,40 %	2,83
Partiellement ou complètement gratuits mais propriétaires	292	26,80 %	1,05
Créés à cette fin (seul ou avec des collègues)	230	21,10 %	0,79
Issus d'un projet de recherche	183	16,80 %	0,52
Internes à mon organisation	173	15,90 %	0,56
Payants « crackés »	102	9,40 %	0,29
Version d'essai ou démo	72	6,60 %	0,18
Gratuits avec des fonctionnalités premium	69	6,30 %	0,18
Total de répondants	**1089**		

TABLEAU 8 : Types de logiciels de production de données utilisés

Deux types de logiciels se distinguent : ceux libres et gratuits et ceux payants. Ils regroupent autour de ⅔ des répondants. Les autres types de logiciels sont quant à eux très nettement moins sollicités, comme les logiciels propriétaires partiellement ou complètement gratuits ou bien les logiciels créés pour une recherche ou issus d'une recherche (21,1 % et 16,8 %).

La classe d'âge semble être discriminante dans l'usage des logiciels libres et gratuits. La part des logiciels libres et gratuits progresse avec les classes d'âges les plus jeunes même si l'on peut constater une tendance à la stabilisation autour de 56 % pour les répondants qui ont aujourd'hui 35 ans ou moins.

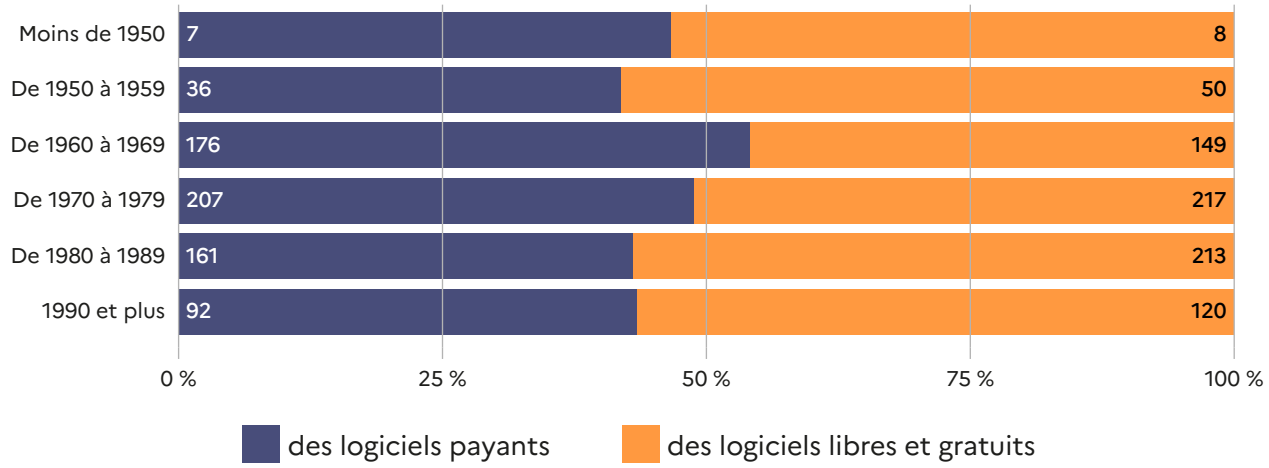


Figure 11: Répartition de l'usage de logiciels payants ou libres et gratuits selon la classe d'âge

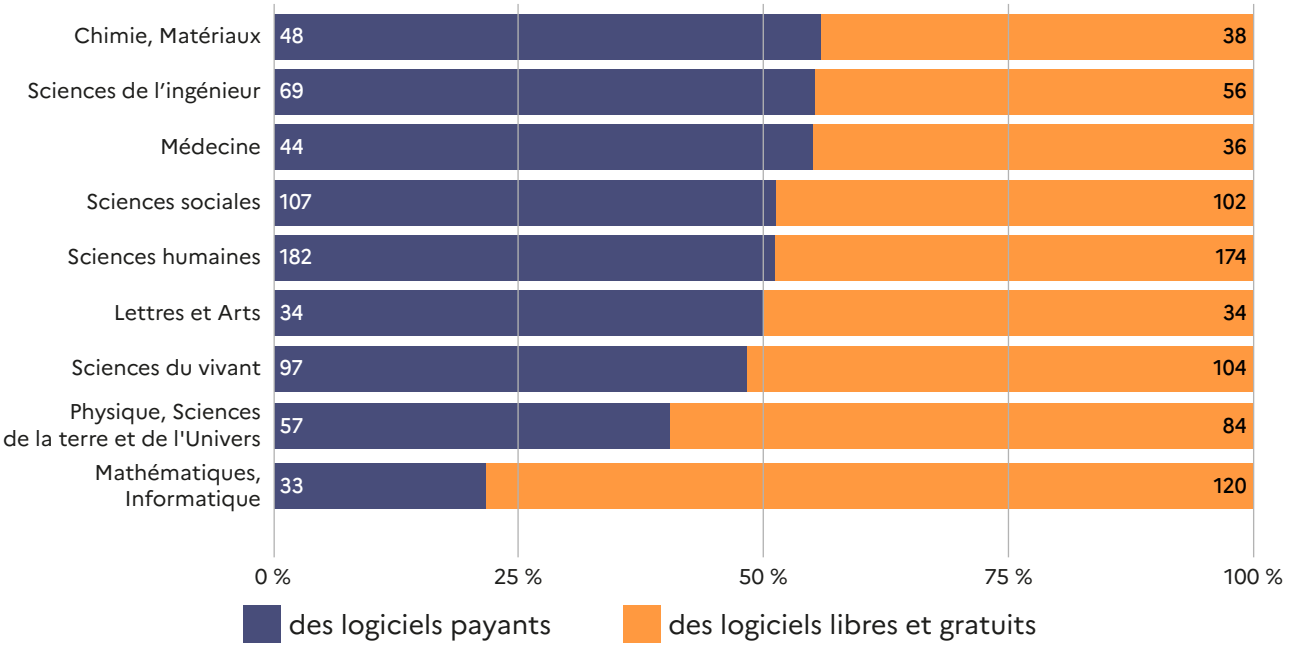


Figure 12: Répartition de l'usage de logiciels payants ou libre et gratuit selon la discipline

Nous constatons également trois tendances disciplinaires:

- La chimie, les sciences de l'ingénieur et la médecine sont les plus enclines à adopter des logiciels payants (autour de 55% en moyenne);
- Les lettres, les SHS et les sciences du vivant se situent dans une répartition équivalente entre les deux types de logiciels mais avec une part d'utilisation des logiciels payants qui est supérieure à la moyenne de la POP;
- La physique et surtout les mathématiques/informatique sont eux résolument du côté de l'usage des logiciels gratuits et libres, avec une pratique qui s'approche des 80% pour les seconds.

RECOMMANDATION : L'appartenance disciplinaire, combinée à la classe d'âge, peut conduire à des comportements assez différents, comportements plus sensibles aux outils qui peuvent s'inscrire dans le Plan National pour la Science Ouverte 2 (2021). Toutefois, les écarts ne sont pas assez marqués pour que l'on considère ces éléments comme les seuls qui puissent influer sur les pratiques des chercheurs.

3. QUELS SONT LES LOGICIELS ET LES LANGAGES DE PROGRAMMATION UTILISÉS ?

HYPOTHÈSE : Face à des études ayant déterminé les principaux logiciels mobilisés pour la recherche et défini des workflows standards, nous allons trouver une diversité de résultats chez les répondants.

La synthèse suivante permet de connaître les logiciels et les langages de programmation qui sont les plus fréquemment utilisés, indépendamment de leur fonction. Le schéma ci-dessous représente le nombre de citations pour les outils cités au moins 10 fois au cours des 4 questions ouvertes portant sur l'utilisation des outils numériques. Ainsi, ce sont 46 outils numériques qui sont cités au moins par 10 répondants.

RÉSULTATS : Les logiciels les plus représentés R et Excel répondent à plusieurs fonctions nécessaires au cours du processus de recherche : production, nettoyage, analyse et visualisation des données.

Ces logiciels sont suivis par des langages de programmation Python et Matlab, ayant une moins grande notoriété mais nous émettons l'hypothèse qu'ils augmenteront en fréquence d'usage.

Par la suite, la synthèse signale majoritairement des logiciels qui sont associés à une seule fonction, à l'exception du cas particulier de la réponse « mes propres logiciels » qui se rencontrent pour les 3 premières fonctions avec une fréquence assez équivalente.

Enfin, nous devons souligner que les caractéristiques des deux premiers outils les plus utilisés, Excel et R, symbolisent les deux grandes tendances en matière d'outils numériques. Le premier est un logiciel commercial ancien qui possède une très forte notoriété et dont l'usage n'est pas spécifique au monde de la recherche. Alors que le second est moins connu, plus récent et son audience et se limite pour l'heure à des spécialistes, particulièrement dans le monde de la recherche. De plus, ce dernier est libre et gratuit contrairement au précédent.

Les outils de production : Excel, Word, R, Python, Matlab, Qgis

Note : pour simplifier la lecture nous avons nommé « outils numériques » les logiciels et langages de programmation donnés par les répondants.

Les outils de productions sont ceux qui ont donné lieu à un nombre plus important de citations avec 492 outils différents signalés pour 2126 citations soit 1,95 outil cité par répondant en moyenne[6].

Les outils signalés au moins 10 fois sont au nombre de 34[7], ce qui signale une pluralité dans les usages. Une scission existe entre les outils libres-gratuits et les outils payants. Le couple Excel/Word représente 17 % des réponses, mais cette proportion est sans doute plus importante en raison de l'exclusion de la réponse « Microsoft Office ». Viennent ensuite des langages de programmation et logiciels libres dont les usages peuvent être qualifiés de plus techniques, comme R, Python et Matlab. Notons également les logiciels conçus par les chercheurs, signalés à 35 reprises et les logiciels propres à l'organisation ou le laboratoire notés par 33 répondants.

Les distinctions disciplinaires dans l'usage des outils de production permettent de définir 4 ensembles distincts :

1. Les sciences physiques, mathématiques, informatique et sciences de l'ingénieur ont une propension à utiliser des logiciels adaptés à leurs besoins (Matlab, Python, LateX, logiciel conçu par moi-même…) ;
2. Les lettres, arts et SHS vont vers des logiciels largement diffusés par les deux grands éditeurs que sont Microsoft et Adobe ;
3. La biologie et la chimie utilisent des logiciels plus confidentiels ou bien plus spécialisés sur les images et les représentations graphiques ;
4. La médecine se caractérise par une faible spécificité en termes d'outils, la représentation graphique les plaçant parmi les outils les plus cités, donc ceux qui sont les plus communément utilisés.

Outils de production[8]			
Nom	Nombre	Nom	Nombre
Excel	219	Arcgis	19
Word	143	Access	18
R	112	C/C++	18
Python	105	SAS	16
Matlab	80	Zotero	15
Qgis	49	Labview	15
Logiciel conçu par moi-même	35	NVivo	14
Libre Office	33	Sphinx	14
Logiciel interne à mon organisation	33	Stata	14
ImageJ	32	Prism	13
FileMaker	27	RStudio	13
LateX	27	Oxygen	12
Origin	26	GraphPad	11
Photoshop	26	Mathematica	10
Lime survey	26	Iramuteq	10
Powerpoint	23	Chemdraw	10
Illustrator	23	Gimp	10
SPSS	22	Inkscape	10

TABLEAU 9 : Liste des logiciels et langages de programmation cités pour la production de données de recherche

6. Nous avons enlevé du décompte des réponses trop vagues comme « Microsoft Office » ou « Adobe ».
7. Si le tableau comprend 37 outils, nous avons mis de côté dans l'analyse, les réponses « Je ne produis pas de données » et « Ne comprends pas la question ».
8. Réponses recodées d'une question ouverte. Le tableau reprend uniquement les outils cités au moins à 10 reprises.

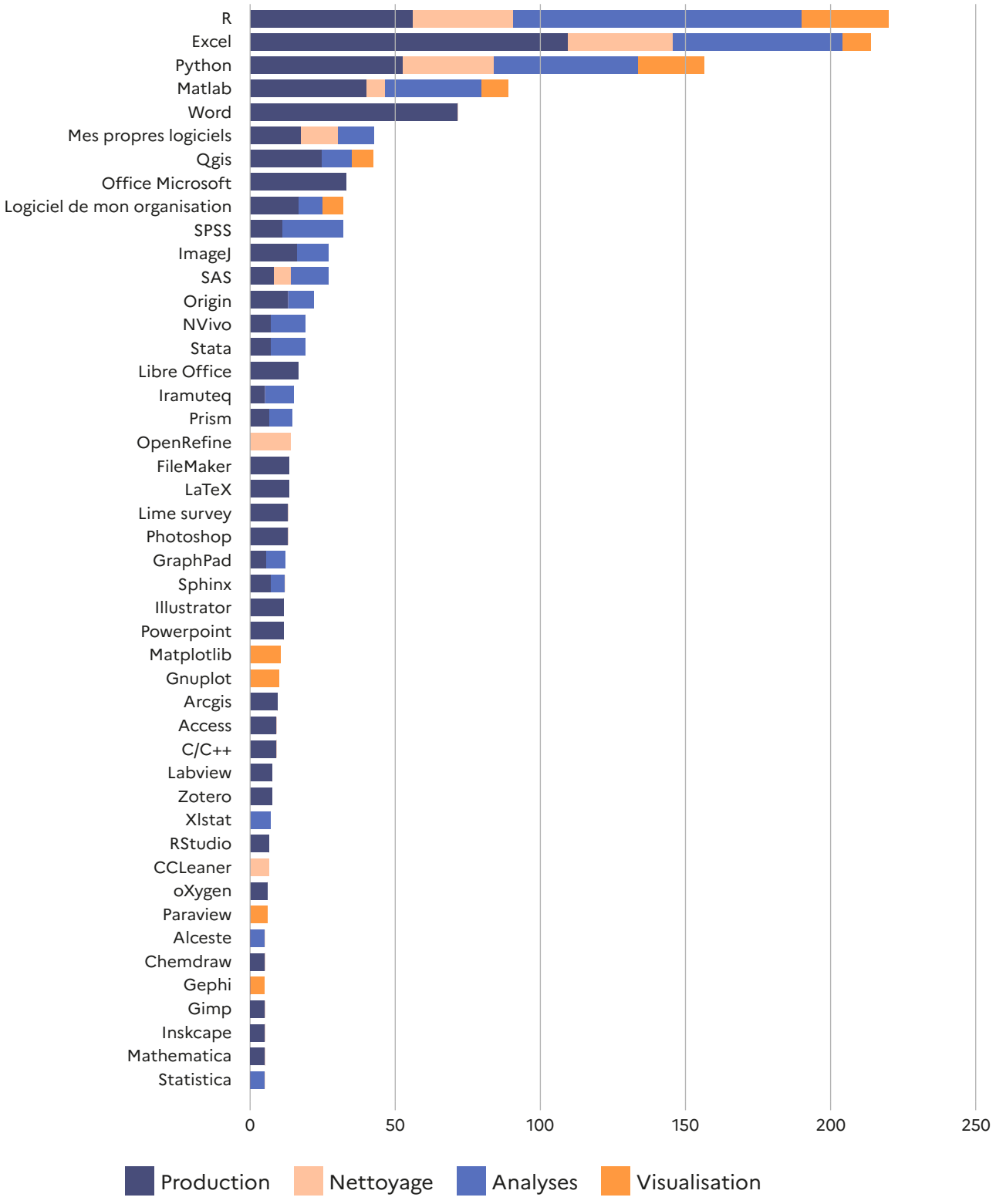


Figure 13: Logiciels, suite et langages de programmation cités les plus fréquemment pour les 4 fonctions: production, nettoyage, analyses et visualisation des données

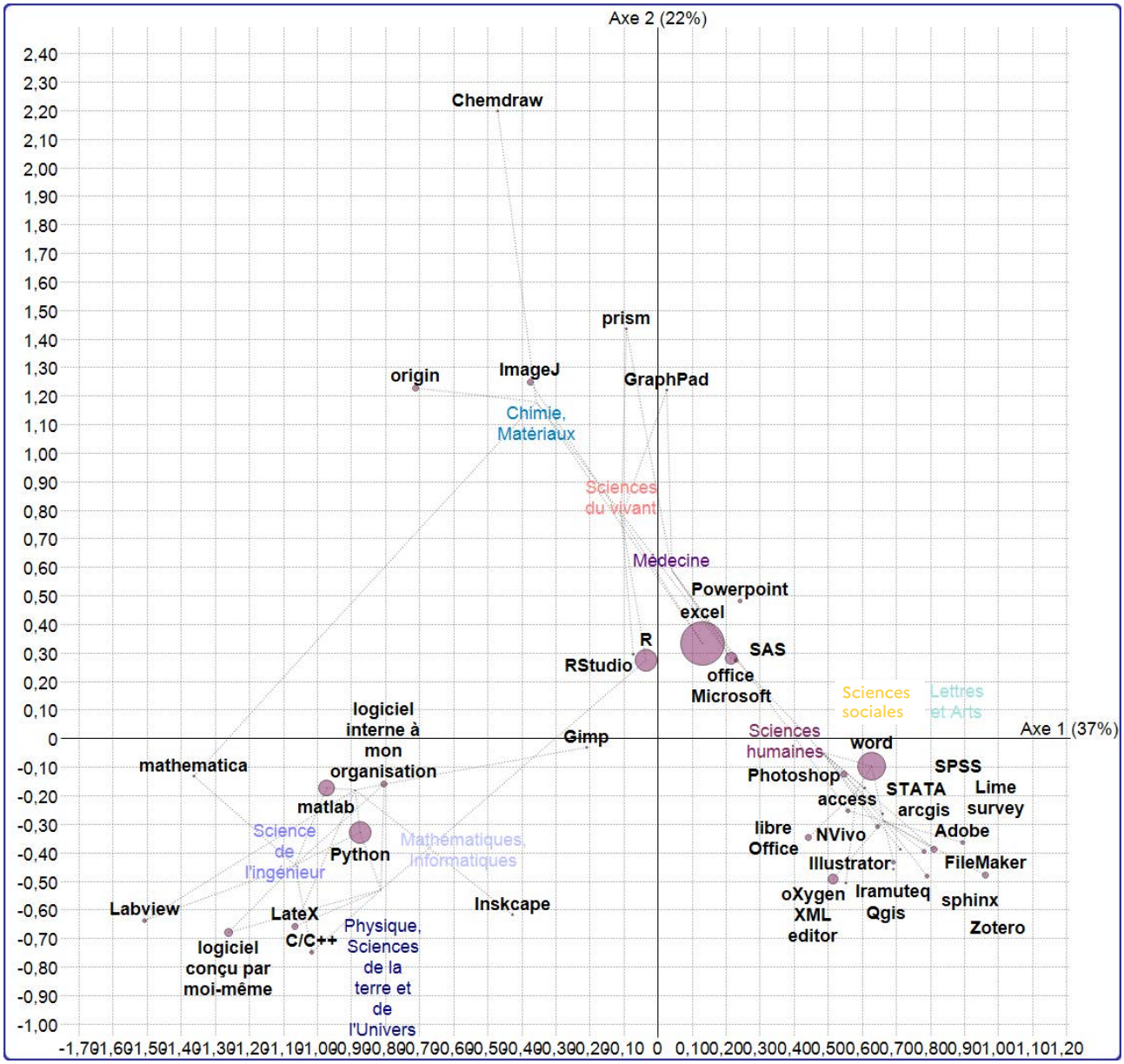


Figure 14: AFC croisant la variable outils de production avec la discipline du répondant

Le contexte de travail marque un clivage dans les usages. L'appartenance à l'université ou aux grands organismes de recherche explique les usages différenciés des outils de production de données de recherche.

Les chercheurs (Directeur de Recherche, DR et Chargé de Recherche, CR) associés aux ingénieurs de Recherche (IR) et, dans une moindre mesure aux ingénieurs d'études (IE), mobilisent des logiciels le plus souvent assez spécifiques, pour certains libres, et parfois spécialement construits pour l'activité de recherche.

En revanche, les professeurs, les maîtres de conférences, les doctorants et les post-doctorants utilisent des outils qui sont moins spécifiques à partir de solutions payantes et propriétaires, ayant donc un caractère résolument plus généraliste, et sont ancrés dans les pratiques (Access, Sphinx, Photoshop).

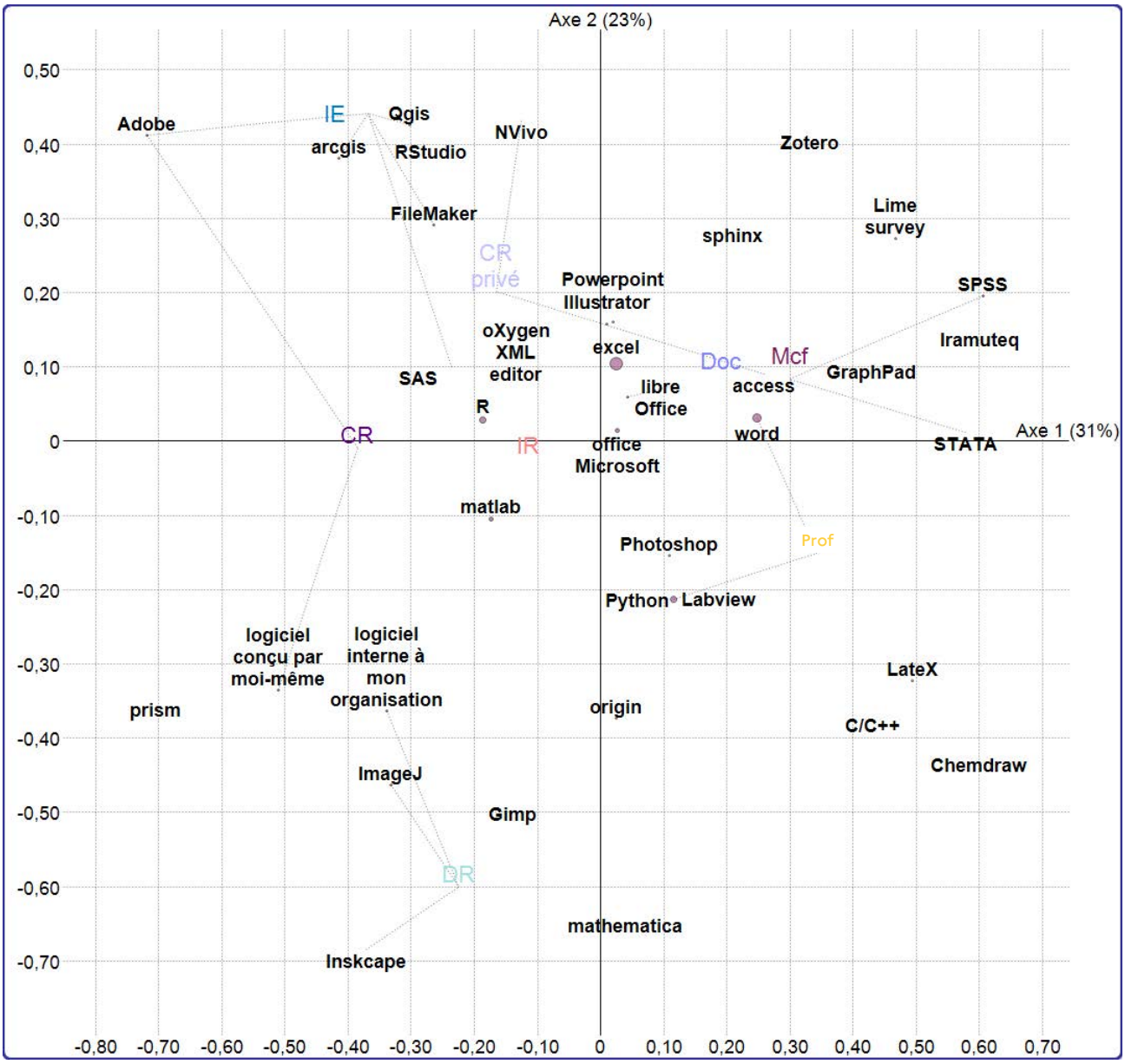


Figure 15: AFC croisant la variable outils de production avec la fonction du répondant

Les outils de nettoyage des données : Excel, R, Python, OpenRefine et mes propres logiciels

La fréquence d'utilisation outils de nettoyage des données est nettement moindre si bien que le principal résultat est qu'ils sont assez peu utilisés. Seulement 27,7 % des répondants se sont positionnés sur cette question, le taux le plus faible pour l'ensemble des questions portant sur les outils numériques.

Les tâches dites de nettoyage de données concernent principalement les chercheurs qui travaillent sur des données numériques. Parmi ceux qui déclarent utiliser un outil de nettoyage (les 15 cités au moins 5 fois), 83,9 % travaillent « souvent » ou « toujours » avec des valeurs numériques, alors qu'ils sont 62 % à le faire dans l'ensemble de la POP.

Les outils d'analyse des données : R, Excel, Python, Matlab, SPSS

Les outils d'analyse des données, les plus cités en volume absolu, donnent lieu à une ou plusieurs réponses dans 55 % de réponses. La liste des 20 outils donnent lieu à 10 réponses au moins, avec une présence de R et d'Excel. Excel se situe ici en recul net par rapport au premier. Ces deux outils d'analyse des données représentent 25,4 % des réponses ce qui est plus important que pour le cas des outils de production de données où ils représent 15,6 % des réponses.

Outils de nettoyage[9]			
Nom	Nombre	Nom	Nombre
Excel	72	Stata	8
R	69	Perl	6
Python	63	Notepad	6
OpenRefine	28	SPSS	5
Mes propres logiciels	25	Sed	5
Matlab	13	QGis	5
CCLeaner	13	Awk	5
SAS	12		

TABLEAU 10 : Liste des 15 logiciels et langages de programmation cités pour le nettoyage des données de recherche

Outils d'analyse[10]			
Nom	Nombre	Nom	Nombre
R	199	QGis	21
Excel	117	Iramuteq	20
Python	99	Origin	18
Matlab	66	Logiciel « maison »	17
SPSS	42	Prism	16
SAS	26	Xlstat	14
Logiciel élaboré par moi-même	25	GraphPad	13
NVivo	24	Alceste	10
Stata	24	Sphinx	10
ImageJ	22	Statistica	10

TABLEAU 11 : Liste des 20 logiciels et langages de programmation cités pour l'analyse des données de recherche

Les distinctions disciplinaires dans l'usage des outils d'analyse permettent de définir 3 ensembles :

- Les SHS se concentrent sur des outils d'analyses ayant une certaine ancienneté et notoriété comme Stata, SPSS, Sphinx ou Alceste, qui sont essentiellement des outils d'analyses statistiques ou d'analyses textuelles ;
- Les sciences du vivant et la médecine semblent attirées par des logiciels qui permettent une analyse schématique comme GraphPad ou ImageJ ;

9. Réponses recodées d'une question ouverte. Le tableau reprend uniquement les outils cités au moins à 5 reprises.

10. Réponses recodées d'une question ouverte. Le tableau reprend uniquement les outils cités au moins à 10 reprises.

- Les sciences et techniques mobilisent des langages de programmation plus récents, comme Python, Matlab ou bien des logiciels élaborés par le ou les chercheurs eux-mêmes.

R, Excel, SAS apparaissent alors plutôt comme des outils transversaux, d'un usage plus fréquent mais moins spécifique à l'une ou l'autre des disciplines.

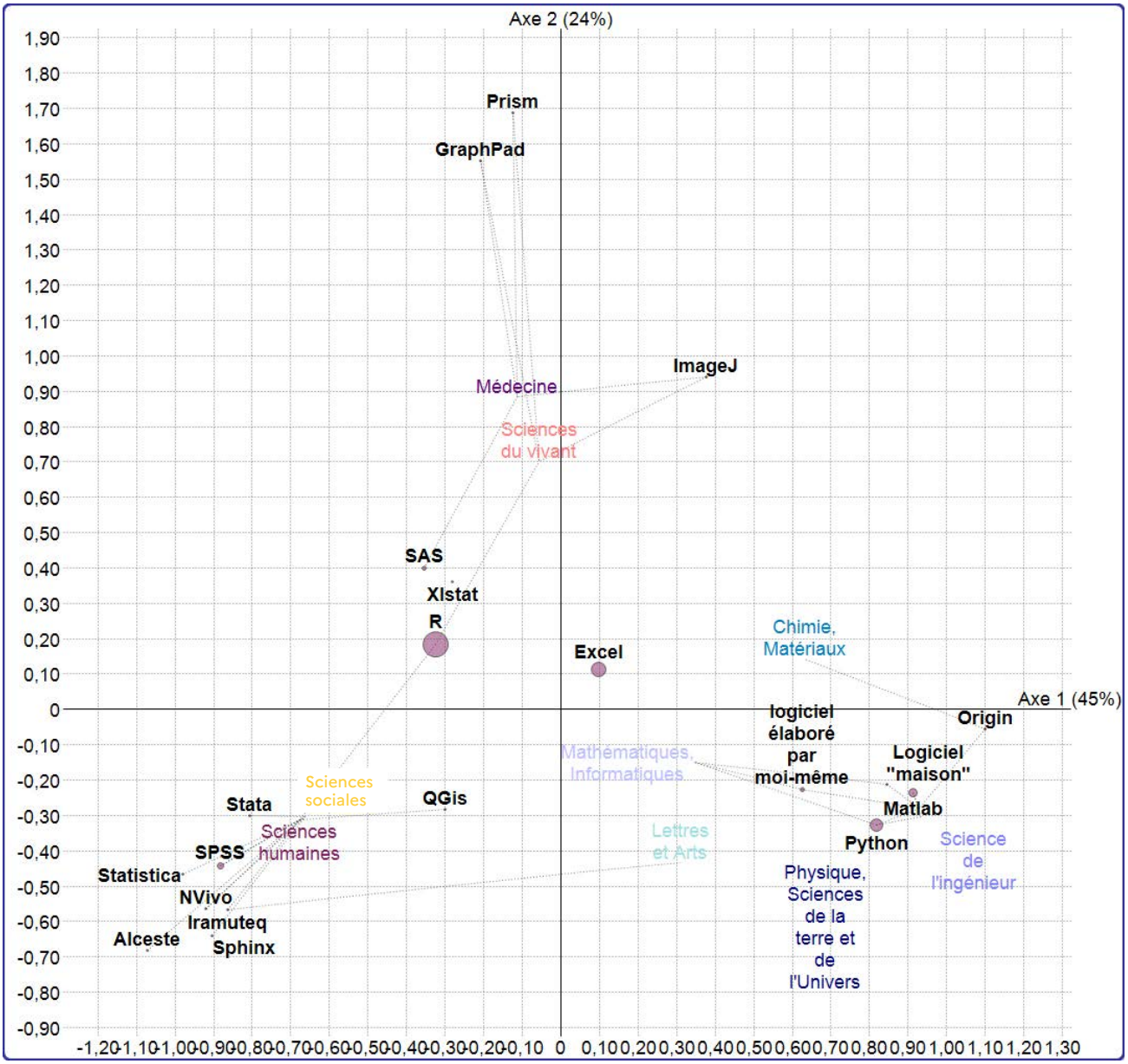


Figure 16: AFC croisant la variable outils d'analyse avec la discipline du répondant

Les outils de visualisation des données

Le taux de réponse de 66,6 % permet de supposer qu'un nombre conséquent de chercheurs utilisent les outils de visualisation des données. Toutefois, il se peut que la fréquence d'usage de ce type d'outil soit en lien avec la présence de fonctions de visualisation dans des logiciels numériques « multifonction » tel qu'Excel. Les répondants n'ont pas spécifié dans leur réponse leur définition de la visualisation des données et il existe des différences d'usages et de production de visualisation entre PowerPoint et matplotlib. La caractéristique de cet ensemble de réponses est que 725 répondants ont cité en moyenne 2 outils de visualisation, signe que ce type d'outils est assez bien identifié.

Deux types d'outils coexistent :

- Les outils dédiés plus spécifiquement à la visualisation tels que PowerPoint, Inkscape, Illustrator, Matplotlib ;
- Les outils plus polyvalents comme Excel et R qui ne supplantent pas les outils spécialisés comme ce peut être le cas pour les outils de production ou d'analyse.

Les outils de visualisation (PowerPoint, Illustrator, Inkscape) regroupent 349 réponses.

Les disciplines mathématiques et informatiques et les sciences de l'ingénieur marquent un usage d'Inkscape tandis que les SHS et lettres sont liées à Illustrator et surtout Photoshop. Les sciences du vivant et la médecine utilisent PowerPoint. Mais la place centrale de ce logiciel permet de comprendre qu'il a une moins forte spécificité disciplinaire que les autres outils numériques précédemment cités.

Outils de visualisation[11]			
Nom	Nombre	Nom	Nombre
PowerPoint	171	Tikz	29
Excel	154	Paint	29
R	106	Origin	25
Inkscape	97	Matplotlib	23
Gimp	90	Gnuplot	23
Illustrator	81	Libre Office	17
Python	65	Chemdraw	13
Photoshop	60	Open Office Draw	12
Matlab	43	ImageJ	11
QGis	43	ArcGis	11
Word	39	Prism	11
Latex	30	Graphpad	10

TABLEAU 12 : Liste des 24 logiciels et langages de programmation cités pour la visualisation des données de recherche

RECOMMANDATION : L'usage des outils libres actuellement minoritaires dans les résultats de notre enquête devrait être observé dans le temps pour montrer s'ils sont d'un usage minoritaire ou si les politiques de science ouverte encouragent leurs usages.

11. Réponses recodées d'une question ouverte. Le tableau reprend uniquement les outils cités au moins à 10 reprises.

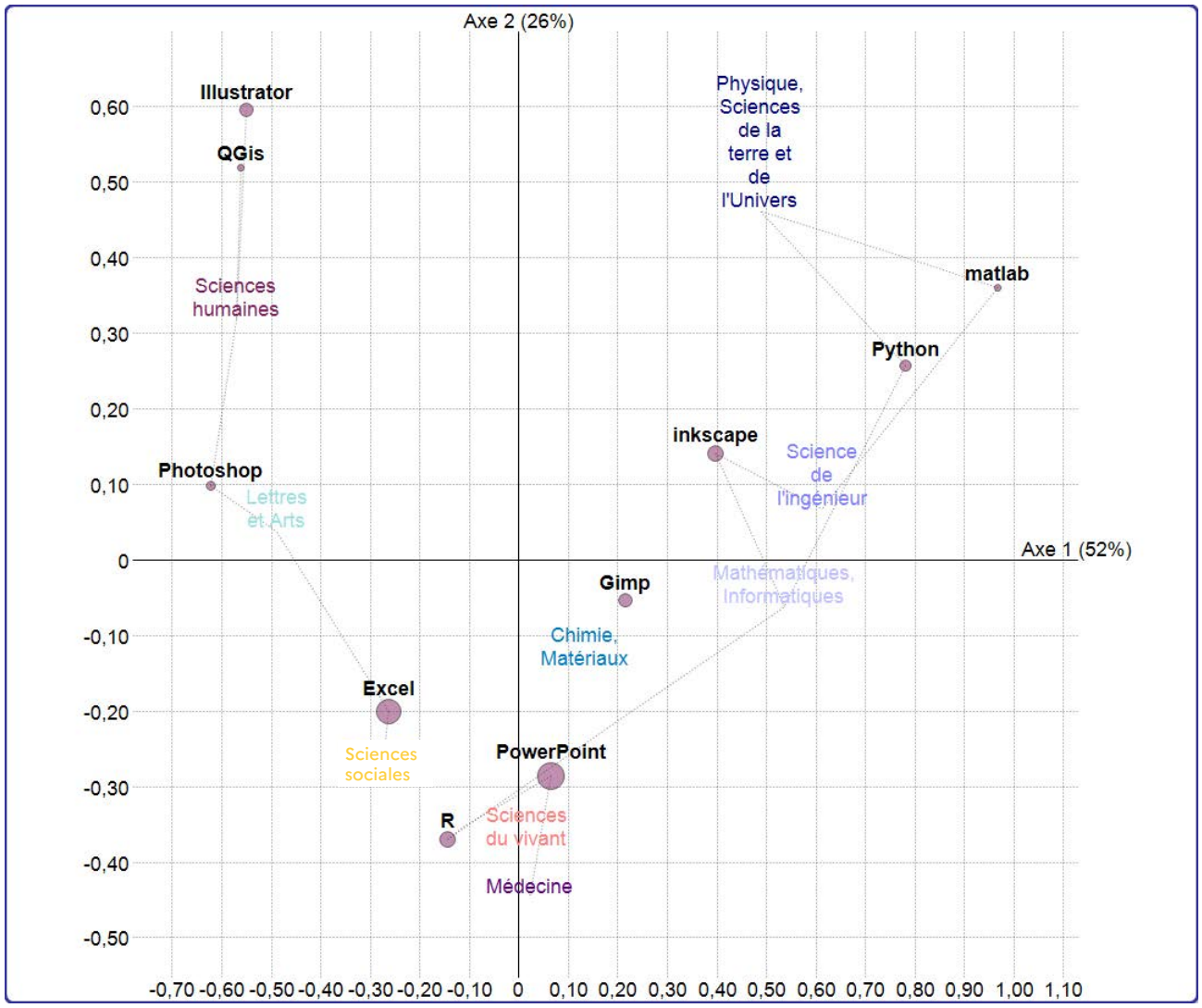


Figure 17 : AFC croisant les 10 premiers outils de visualisation avec la discipline du répondant

4. FOCUS SUR 3 OUTILS NUMÉRIQUES ET 3 LOGIQUES D'INNOVATION

Nous procédons à une analyse plus approfondie des 3 premiers outils les plus cités : Excel, R et Python. Leurs caractéristiques propres en font à la fois des « concurrents » et des outils complémentaires. Notre analyse tente d'évaluer si l'on peut trouver des profils de chercheurs, qui par leurs caractéristiques propres peuvent être associés à chacun de ces trois outils. Nous constatons que nous rencontrons trois configurations. Nous rencontrons l'innovation : en voie d'institutionnalisation (N. Alter, 2015) symbolisée par R ; le logiciel institutionnalisé représenté par Excel ; et la pratique en émergence avec Python.

Les utilisateurs de R (n = 244) : la voie de l'institutionnalisation

Une moyenne d'âge des utilisateurs de R plus jeune

Les utilisateurs de R se caractérisent par une moyenne d'âge et un âge médian inférieur d'environ 4 ans à la POP. L'usage de R est lié à des chercheurs parmi les plus jeunes, les écarts étant sensibles pour les 35-45 ans et nettement plus marqués pour les chercheurs de moins de 35 ans.

RÉSULTATS : L'émergence de ce logiciel à la fin des années 1990 peut expliquer en grande partie la « jeunesse » des utilisateurs, qui ont été formés à ce logiciel durant leurs études universitaires ou en début de carrière.

Moyenne = 1979,32 ; Médiane = 1980	Nombre	% Obs.	Ech.
Moins de 1950	2	0,80 %	1,60 %
De 1950 à 1954	0	0,00 %	1,00 %
De 1955 à 1959	9	3,70 %	5,60 %
De 1960 à 1964	15	6,10 %	9,00 %
De 1965 à 1969	26	10,70 %	12,90 %
De 1970 à 1974	25	10,20 %	14,90 %
De 1975 à 1979	38	15,60 %	14,90 %
De 1980 à 1984	37	15,20 %	14,00 %
De 1985 à 1989	39	16,00 %	11,20 %
1990 et plus	53	21,70 %	15,00 %
Total	**244**	**100,00 %**	**100,00 %**

TABLEAU 13 : Répartition des utilisateurs du logiciel R en fonction de leur classe d'âge

Une distinction disciplinaire

Les distinctions disciplinaires sont significatives, avec une très nette surreprésentation des sciences du vivant, celle moins marquée des mathématiques et de l'informatique et des SHS, voire de la médecine. En revanche, les sciences de l'ingénieur semblent utiliser nettement moins ce logiciel tout comme la chimie et les sciences des matériaux. R est assez multifonction et sied assez mal aux besoins expérimentaux propres à ces disciplines qui nécessitent parfois des logiciels *ad hoc*.

La progression de R ne se fait pas de manière uniforme et surtout ne touche pas forcément les disciplines les plus marquées en STM.

Un contexte de travail en petit collectif

La corrélation entre l'usage de R et un contexte de travail en petit collectif est particulièrement forte. Travailler en équipe de 2 à 5 personnes est un élément favorable dans l'adoption de ce logiciel. L'hypothèse que l'usage d'un outil numérique est lié à un contexte de travail est confirmée indirectement. L'imitation et la contrainte en local de partager les travaux expliquent potentiellement des usages plus développés de R dans des petits collectifs de travail.

Disciplines	Nombre	% Obs.	Ech.
Sciences de l'ingénieur	7	2,90%	8,40%
Sciences humaines	63	26,00%	25,10%
Sciences sociales	39	16,10%	14,80%
Lettres et Arts	2	0,80%	5,30%
Médecine	16	6,60%	5,00%
Sciences du vivant	**64**	**26,40%**	**12,70%**
Chimie, Matériaux	5	2,10%	5,40%
Mathématiques, Informatique	38	15,70%	12,80%
Physique, Sciences de la terre et de l'Univers	8	3,30%	9,50%
Total	**242**	**100,00%**	**98,80%**

TABLEAU 14 : Répartition des utilisateurs du logiciel R en fonction de leur discipline

Environnement de recherche	Nombre	% Obs.	Ech.
Je travaille seul	29	11,90 %	23,80 %
Nous sommes entre 2 et 5 personnes	**167**	**68,40 %**	**57,00 %**
Nous sommes entre 6 et 10 personnes	37	15,20 %	13,80 %
Nous sommes entre 11 et 20 personnes	10	4,10 %	3,90 %
Nous sommes entre 21 et 50 personnes	1	0,40 %	1,00 %
Nous sommes plus de 50 personnes	0	0,00 %	0,60 %
Total	**244**	**100,00 %**	**100,00 %**

TABLEAU 15 : Répartition des utilisateurs du logiciel R en fonction de leur contexte de travail

Les utilisateurs d'Excel (n = 369) : le logiciel institutionnalisé

Des caractéristiques sociodémographiques conformes à la POP pour les utilisateurs d'Excel

Les utilisateurs d'Excel, plus nombreux (n = 369), sont plus conformes, dans leurs caractéristiques sociodémographiques à la POP, avec une moyenne d'âge et une médiane quasi identiques. L'âge n'est pas discriminant.

Les distinctions disciplinaires

Elles sont pour les sciences du vivant qui semblent privilégier ce logiciel au détriment de R. Les mathématiciens et des informaticiens semblent se détourner de ce logiciel, sans doute au profit de langages de programmation (R, Python) ou bien encore des logiciels *ad hoc*.

Disciplines	Nombre	% Obs.	Ech.
Sciences de l'ingénieur	33	9,00 %	8,40 %
Sciences humaines	**101**	**27,60 %**	**25,10 %**
Sciences sociales	59	16,10 %	14,80 %
Lettres et Arts	14	3,80 %	5,30 %
Médecine	27	7,40 %	5,00 %
Sciences du vivant	70	19,10 %	12,70 %
Chimie, Matériaux	27	7,40 %	5,40 %
Mathématiques, Informatique	14	3,80 %	12,80 %
Physique, Sciences de la terre et de l'Univers	21	5,70 %	9,50 %
Total	**366**	**100,00 %**	**98,80 %**

TABLEAU 16 : Répartition des utilisateurs du logiciel Excel en fonction de leur discipline

Le contexte de travail

Concernant le **contexte de travail**, la corrélation est bien différente par rapport au logiciel R. Le contexte de travail collectif influence assez peu l'utilisation d'Excel, avec des écarts qui peuvent difficilement être considérés comme significatifs.

Les utilisateurs d'Excel ne se distinguent pas par rapport à la POP. La notoriété de ce logiciel semble se suffire à elle-même pour induire son usage chez les chercheurs.

Les utilisateurs de Python (n = 170) : une pratique émergente

Les utilisateurs de Python sont une population assez restreinte soit 15,6 % de la POP.

Ils se caractérisent par une moyenne d'âge nettement plus jeune que la POP comme celle des utilisateurs de R (avec une médiane à 40 ans). Les 35-40 ans sont surreprésentés. Chez les plus jeunes générations, la surreprésentation est moins marquée. À l'inverse, même s'il existe des utilisateurs de ce langage nés avant 1980, ils sont globalement sous-représentés au point qu'on peut les considérer comme des précurseurs.

Les distinctions disciplinaires montrent que de très nombreuses disciplines ont très peu d'utilisateurs de Python. Contrairement à R, Python demeure une pratique spécialisée, assez émergente, (dans le sens où cette pratique concerne surtout les jeunes générations). Les utilisateurs de ce langage se trouvent massivement dans les disciplines les plus techniques. La moitié d'entre eux (51,8 %) figure dans les mathématiques, l'informatique, la physique, les sciences de la terre et de l'univers, une part plus de deux fois supérieure à leur poids dans la POP. Si on les associe aux utilisateurs issus des sciences de l'ingénieur, la part se monte à 7 répondants sur 10.

Moyenne = 1978,72 ; Médiane = 1980,00 ; Ecart-type = 10,82	Nombre	% Obs.	Ech.
Moins de 1960	5	2,90 %	8,20 %
De 1960 à 1964	12	7,10 %	9,00 %
De 1965 à 1969	14	8,20 %	12,90 %
De 1970 à 1974	29	17,10 %	14,90 %
De 1975 à 1979	23	13,50 %	14,90 %
De 1980 à 1984	**35**	**20,60 %**	**14,00 %**
De 1985 à 1989	18	10,60 %	11,20 %
De 1990 à 1994	25	14,70 %	11,80 %
1995 et plus	9	5,30 %	3,20 %
Total	**170**	**100,00 %**	**100,00 %**

TABLEAU 17 : Répartition des utilisateurs de Python selon leur classe d'âge

Disciplines	Nombre	% Obs.	Ech.
Sciences de l'ingénieur	33	19,40%	8,40%
Sciences humaines	13	7,60%	25,10%
Sciences sociales	9	5,30%	14,80%
Lettres et Arts	4	2,40%	5,30%
Médecine	6	3,50%	5,00%
Sciences du vivant	13	7,60%	12,70%
Chimie, Matériaux	4	2,40%	5,40%
Mathématiques, Informatique	43	25,30%	12,80%
Physique, Sciences de la terre et de l'Univers	**45**	**26,50%**	**9,50%**
Total	**366**	**100,00%**	**98,80%**

TABLEAU 18 : Répartition des utilisateurs de Python selon leur discipline

Environnement de travail collectif

Cette sous-population se distingue également – et très nettement – **par sa propension à travailler dans des environnements collectifs** : ils ne sont que 8,8% à travailler seuls et 70% à travailler dans des petits collectifs (2 à 5 personnes). Cette population, nettement plus jeune que la POP ne comprend pas plus de doctorants, dont l'activité est assez souvent solitaire, notamment en SHS (Chao et al, 2015).

Environnement de recherche	Nombre	% Obs.	Ech.
Je travaille seul	15	8,80%	23,80%
Nous sommes entre 2 et 5 personnes	**119**	**70,00%**	**57,00%**
Nous sommes entre 6 et 10 personnes	25	14,70%	13,80%
Nous sommes entre 11 et 20 personnes	7	4,10%	3,90%
Nous sommes entre 21 et 50 personnes	3	1,80%	1,00%
Nous sommes plus de 50 personnes	1	0,60%	0,60%
Total	**170**	**100,00%**	**100,00%**

TABLEAU 19 : Répartition des utilisateurs de Python selon leur contexte de travail

Les utilisateurs de Python se rencontrent avant tout parmi les chercheurs des plus jeunes générations qui exercent leur mission principalement dans de grands établissements de recherche, et moins au sein des universités. Il semble que la fonction exercée pèse considérablement dans la fréquentation de ce langage. Les universitaires sont moins utilisateurs de Python, surtout les maîtres de conférences (– 7,3 points de %) que les chercheurs, directeurs de recherches (+ 4,3 points de %) selon le Tableau 20.

Environnement de recherche	Nombre	% Obs.	Ech.
CR privé	4	2,40%	2,80%
JE	9	5,30%	6,90%
RI	28	16,50%	11,50%
CE	26	15,30%	12,90%
DR	19	11,20%	9,30%
Prof	24	14,10%	13,70%
MCF	**31**	**18,20%**	**25,20%**
Doc	29	17,10%	17,80%
Total	**170**	**100,00%**	**100,00%**

TABLEAU 20 : Répartition des utilisateurs de Python selon leur fonction

Pour finir, trois autres résultats peuvent être notés pour renforcer cette analyse. En premier lieu, ces utilisateurs de Python sont pour près de la moitié d'entre eux (49 %) des utilisateurs de Linux (contre 22,5 % dans la POP), orientent massivement (91,2 %) leur choix vers des logiciels libres et gratuits (contre 69,5 % dans la POP) et sont pour 68,2 % d'entre eux des hommes alors que ces derniers sont juste majoritaires dans la POP (51,9 %).

Système d'exploitation	Nombre	% Obs.	Ech.
Windows	69	26,40%	48,70%
Mon Cos	46	17,60%	23,80%
Linux	**128**	**49,00%**	**22,50%**
Autres Unix	4	1,50%	0,80%
Autres OS (Android…)	14	5,40%	4,20%
Total	**261**	**100,00%**	**100,00%**

TABLEAU 21 : Répartition des utilisateurs de Python selon leur système d'exploitation

Logiciels	Nombre	% Obs.	Ech.
Libres et gratuits	155	91,20 %	69,50 %
Créés à cette fin (seul ou avec des collègues)	78	45,90 %	21,10 %
Payants	71	41,80 %	62,40 %
Issus d'un projet de recherche	54	31,80 %	16,80 %
Partiellement ou complètement gratuits mais propriétaires	38	22,40 %	26,80 %
Internes à mon organisation	29	17,10 %	15,90 %
Payants « crackés »	11	6,50 %	9,40 %
Version d'essai ou démo	5	2,90 %	6,60 %
Gratuits avec des fonctionnalités premium	4	2,40 %	6,30 %

TABLEAU 22 : Répartition des utilisateurs de Python selon leur préférence de type de logiciels

Genre	Nombre	% Obs.
Un homme	116	68,20 %
Une femme	44	25,90 %
Je ne souhaite pas répondre	10	5,90 %
Total	**170**	**100,00 %**

TABLEAU 23 : Répartition des utilisateurs de Python selon leur genre

L'usage de Python est émergent et encore spécialisé. Cette étude ne peut conclure sur son usage futur. Les caractéristiques de la POP d'utilisateurs permettent de la distinguer des utilisateurs de R. Ce dernier semble s'être plus largement diffusé, au point d'apparaître comme une innovation en voie d'institutionnalisation, ce qui n'est pas le cas pour le langage Python, tout au moins si l'on raisonne sur les communautés de recherche en France.

RECOMMANDATION 1 : Les variables des logiciels libres et gratuits et systèmes d'exploitation peuvent confirmer des usages liés à la science ouverte qu'il conviendra d'affiner.

RECOMMANDATION 2 : La multiplication des supports (ouvrages, formations URFIST, enseignements) et les attentes des employeurs devraient être étudiées pour rendre compte de son institutionnalisation dans les pratiques.

5. LES DONNÉES DE LA RECHERCHE

Nous avons posé une série de 6 questions qui ont trait aux pratiques liées aux données de la recherche. N'ayant pas opté pour une définition (Borgman, 2020 ; Leonelli, 2019) *a priori* de ce qu'est une donnée de recherche, nous avons la possibilité d'avoir des retours sur des conceptions différentes de ces données de la recherche. Nous avons préféré utiliser des termes vernaculaires comme « nombre », « valeur numérique » ou « texte », tant il est difficile pour les répondants de se situer dans des nomenclatures spécialisées[12]. De même, l'objectif était de connaître la fréquentation de ces différents types de données (d'où la proposition de fréquence), plutôt que d'évaluer la connaissance du répondant à propos de ces nomenclatures techniques et/ou théoriques.

Quel type de données de recherche travaillées ?

En premier lieu, nous souhaitions évaluer sur quel type de jeux de données les répondants travaillent. Les réponses dressent un schéma en deux pôles : d'un côté les données utilisées assez fréquemment voire très fréquemment, de l'autre celles qui sont soit une spécialité, soient qui viennent en appoint des autres.

Les nombres, valeurs numériques (62,1 % d'usage fréquent) et les textes (61,9 % d'usage fréquent) sont les données les plus fréquemment utilisées par les chercheurs. Les 4 autres types de données (Images, vidéos, 3D et sons) sont nettement moins présentes.

Il existe une différence entre les données images et les données vidéos au niveau des usages (près de 40 %/à peine 10 %).

Le résultat est faible pour les données 3D comme pour les sons (moins de 10 % de fréquence).

Les données « sons » sont assez régulièrement utilisées en sciences sociales, lors d'interviews semi-directifs par exemple avant leur retranscription.

Ces données « sons » peuvent avoir un statut de données « texte » lorsqu'elles sont transcrites en texte.

En croisant avec les disciplines, les usages sont plus variés, surtout les valeurs numériques et les données textes.

- Sons : Sciences sociales ;
- Données 3D : Sciences de l'Ingénieur ;
- Données textes : disciplines littéraires et artistiques ;
- Données images : l'usage fréquent concerne entre 15 et 25 % des répondants d'une discipline et ne dépend pas d'une discipline.

Finalement il ressort une diversité du lien entre données et discipline, plutôt qu'une forme de spécialisation. S'il peut exister des tendances, il apparaît que bon nombre de types de données sont partagées par l'ensemble des chercheurs et dans l'ensemble des groupes disciplinaires.

12. Difficultés signalées dans de nombreuses enquêtes portant sur les données de la recherche, comme celle d'Alexandre Serres (dir.), Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet, *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*, 2017, qui avait donné lieu à la distinction entre « données sources » et « données résultats ».

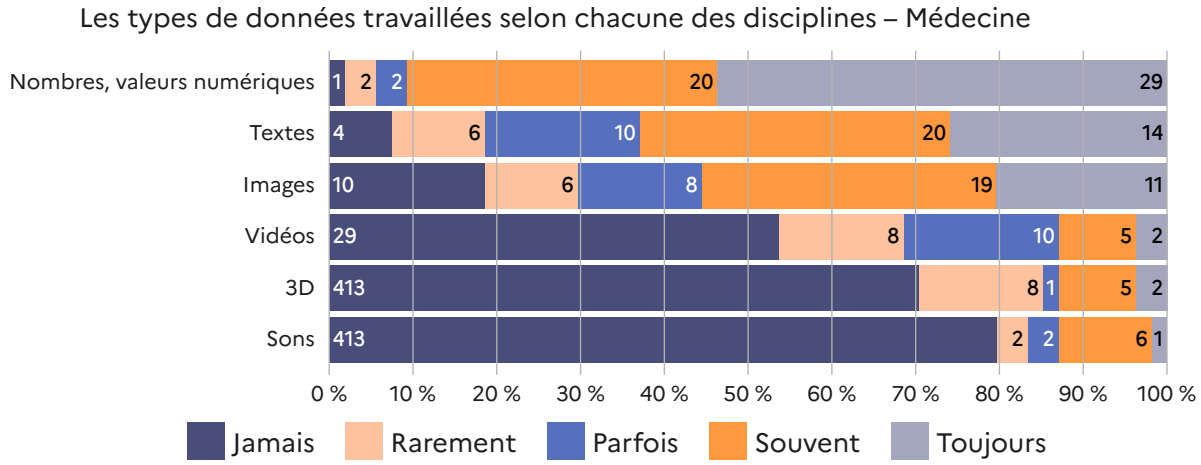


Figure 18: Fréquence d'usage des types de données travaillées

RECOMMANDATION: Pour s'inscrire dans des standards de reproductibilité de ce questionnaire, peut-être faudrait-il reprendre des typologies établies notamment par Christine Borgman. Mais il faut se poser la question de l'acculturation des publics dans le cadre de la passation d'un questionnaire assez long où les définitions sont susceptibles de dérouter le répondant.

Le volume des données

QUESTION 12BIS: quelle est la quantité de ces données dans une de vos recherches ?

Cette question a posé problème aux répondants, tant dans sa formulation que dans leur capacité à estimer le volume des données qu'ils utilisent. Le premier indice tient au nombre conséquent de non-réponse (157) et de réponses inexploitables (263). Le second indice réside dans les 125 répondants qui déclarent ne pas avoir compris le sens de la question et enfin les 151 autres qui déclarent ne pas savoir estimer le volume de leurs données. Pour les premiers, cela signifie probablement que les problématiques autour des données de la recherche sont éloignées voire inconnues. Ces répondants appartiennent majoritairement (58,4 %) aux SHS, un tiers d'entre eux travaille seuls (contre 23,8 % pour la POP) avec une présence marquée des MCF.

Au total on parvient à 696 (63,9 %) répondants dont la réponse n'a pas pu être intégrée aux résultats. De plus, cela signifie que pour la moitié des répondants (545), cette question n'a pas véritablement de sens, comme si le fait de connaître le volume de données travaillées manipulées était peu pertinent. Dès lors, les réponses à cette question sont à analyser avec prudence.

Pour un peu plus d'un tiers (35,4 %) le volume de données se situe entre 1 et 10 Go, un volume courant dans les capacités informatiques. Environ 1 chercheur sur 5 utilise moins d'un 1 Go de données et 45 % au-dessus du seuil des 10 Go. Le Téraoctet est mentionné par environ 1 répondant sur 5.

RECOMMANDATION: Cette question libre puis modifiée avec une échelle de réponses montre que proposer des réponses induit une évaluation du volume, mais il n'est pas possible de dire si les répondants ont compris la question. Le mieux serait donc d'ajouter des items « n'a pas compris la question », « ne peut pas évaluer » ou « libre », en plus d'une échelle de réponses.

Système d'exploitation (OS)	Nombre	% (/POP)
1 à 10 Mo	34	3,1%
10 à 100 Mo	14	1,3%
100 à 1000 Mo	29	2,7%
1 à 10 Go	139	12,8%
10 à 100 Go	53	4,9%
100 à 1000 Go	36	3,3%
1 à 10 To	73	6,7%
> 10 To	15	1,4%
NSP	151	13,9%
Non réponse	157	14,4%
N'a pas compris la question	125	11,5%
Réponse inexploitable	263	24,2%
Total	**1089**	100,0%

TABLEAU 24 : Le volume de données estimé par le chercheur

La pratique de réutilisation des données

La pratique de réutilisation des données est globalement peu fréquente. Elle concerne seulement 28 % des chercheurs (ceux qui ont répondu « souvent » + « toujours »).

La pratique de réutilisation des données n'est pas partagée par tous les chercheurs et la marge de progression est importante, étant donné la part importante de « parfois » (38,4 %) chez les répondants.

Les distinctions disciplinaires sont présentes. Les spécialistes des lettres et arts se détachent, car ils sont moins d'un sur cinq à ne pas être dans la pratique de réutilisation. En sciences humaines et sociales, mathématiques, informatiques et physique, la pratique de réutilisation existe réellement pour un peu moins d'un tiers des répondants de ces disciplines. Enfin, en chimie, sciences du vivant, médecine, la réutilisation des données est limitée voire marginale (pour les sciences de l'ingénieur).

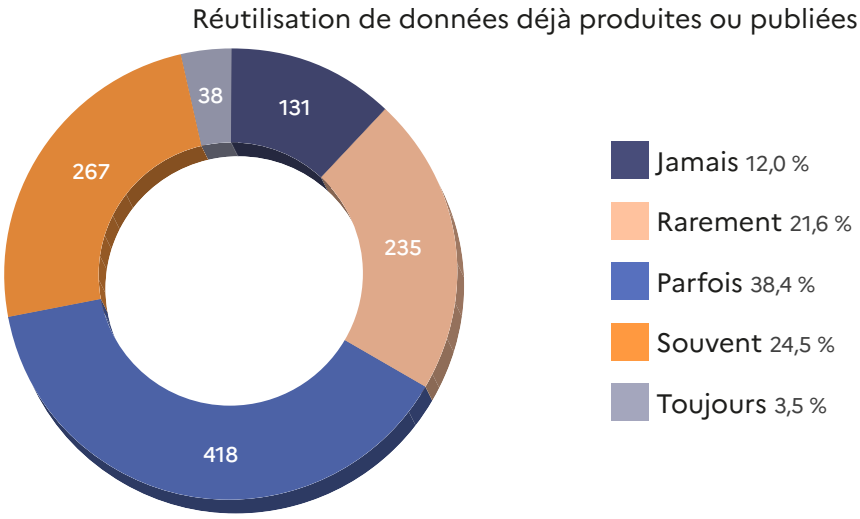


Figure 19 : La pratique de réutilisation des données déjà produites ou publiées

La notion de réutilisation renvoie à des réalités différentes selon les disciplines. Les lettres et arts, les sciences humaines qui intègrent l'histoire réutilisent des documents ou des textes produits par d'autres, parfois dans des périodes reculées, alors que la réutilisation chez les chimistes ou les spécialistes des matériaux, en physique correspond à la réutilisation de données produites plus récemment, par des instruments de mesures et qui sont largement diffusées, parfois à l'échelle internationale. En sciences techniques et appliquées, l'existence de coopérations de recherche académiques ou de partenariats public-privé est également susceptible de moduler les conditions d'accès et de réutilisation des données à l'égard d'autres communautés de recherche. La signification de la notion de réutilisation demande à être affinée et précisée. Elle est probablement différente selon le type et le contexte de travail de recherche réalisé. Cela suggère qu'il existe une pratique de réutilisation plus large que nos résultats ne le laissent apparaître.

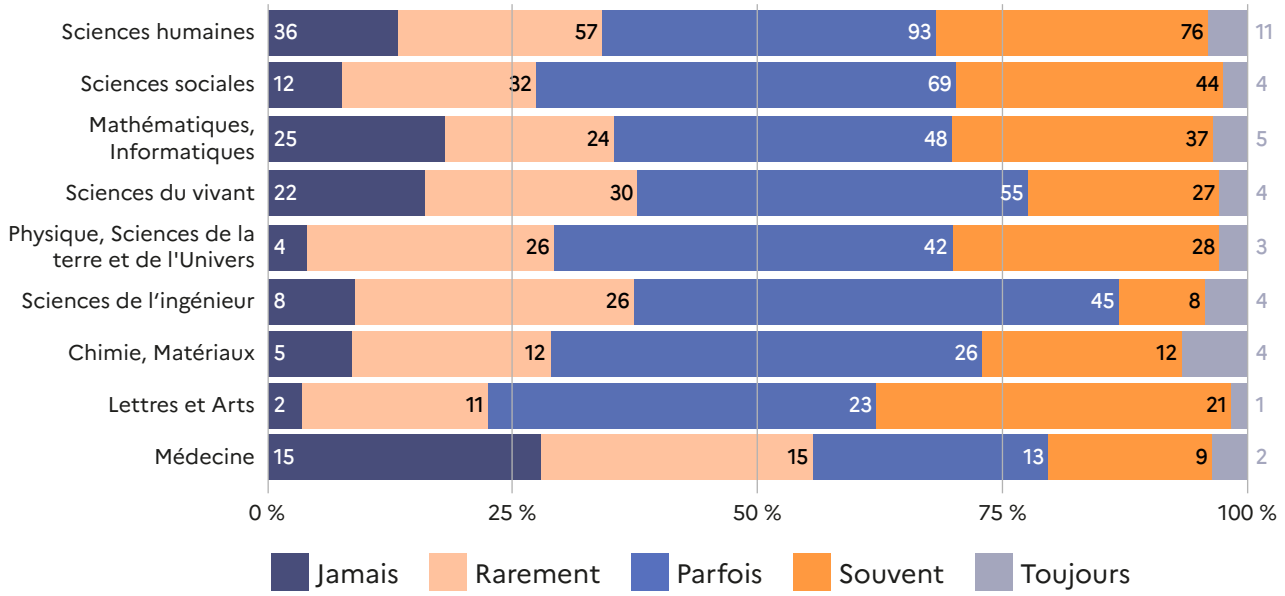


Figure 20: La pratique de réutilisation des données déjà produites ou publiées selon la discipline

RECOMMANDATION : Interroger la nature et le contexte des données réutilisées. Il semble, selon les résultats obtenus, que les lettres, arts, sciences humaines travaillent à partir d'archives qu'il faudrait distinguer de données produites à partir d'instrument de mesures.

Les pratiques d'archivage des données à long terme en fonction des moyens d'accès aux infrastructures

QUESTION 15: Archivez-vous (conservation à long terme) ces données ?

Nous admettons ici que nous n'avons pas respecté les nomenclatures habituellement développées par les professionnels de l'information pour essayer de ne pas nous éloigner des pratiques des chercheurs[13]. Nous n'avons pas communiqué de définition, estimant que le temps de réponse au questionnaire était déjà long.

HYPOTHÈSE : Le contexte de travail influe sur l'accès à l'archivage des données.

Les pratiques d'archivage des données à long terme subissent de multiples variations selon les caractéristiques sociodémographiques du répondant alors qu'un chercheur sur deux développe cette pratique.

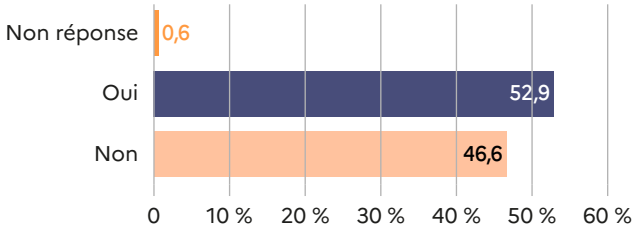


Figure 21: La pratique d'archivage (conservation à long terme) des données

Les distinctions disciplinaires interviennent avec une pratique d'archivage à long terme assez peu développée dans les mathématiques et l'informatique, les sciences de l'ingénieur et les sciences humaines. À l'inverse, avec plus 60 % de répondants qui pratiquent l'archivage, la médecine, les sciences du vivant et la chimie sont des disciplines où le souci de l'archive est plus marqué souvent sous la pression d'une règlementation spécifique (recherche impliquant la personne humaine, biodiversité…) ou dans le cadre partenarial, par exemple lors d'une opération de transfert de technologie où les données doivent rester disponibles pour que la technologie-résultat puisse être exploitée (au titre de savoir-faire associé à un brevet par exemple).

L'hypothèse doit être poursuivie, surtout au niveau la valorisation économique des données sans écarter / oublier l'acculturation progressive aux principes FAIR, qui constituent un des piliers majeurs de la science ouverte.

En mathématiques et informatiques, les codes sources des logiciels pourraient être à l'origine d'une sous-pratique. Roberto Di Cosmo et Stefano Zacchiroli ont souligné la valeur des codes sources et la nécessité de les sauvegarder *via l'organisation Software heritage*[14] (patrimoine logiciel). Il est probable que la problématique soit assez proche pour les sciences de l'ingénieur. Pour les autres disciplines, qui ont une pratique de sauvegarde à long terme supérieure à la moyenne, sans doute que des considérations institutionnelles interfèrent

13. Par exemple, Lina Sbeih, Fanny Dedet, Patrick Moreau, Esther Dzale. L'archivage des données de la recherche à l'Inra. Élément de réflexion, démarche et perspectives. *Cahier des Techniques de l'INRA*, INRA, 2020. ffhal-02861909f

14. Roberto Di Cosmo et Stefano Zacchiroli. 2016. « Software Heritage: Why and How to Preserve Software Source Code ». In Proceedings of 14th International Conference on Digital Preservation, Kyoto, *Japan*, *September* 2017.

comme le renforcement des politiques de sécurité des systèmes d'information (dans le cadre de la protection du patrimoine immatériel de l'état, PPST (protection du potentiel scientifique et technique de la nation)…). La question n'est pas posée dans notre enquête, mais la nécessité de répondre aux principes FAIR portés par les financeurs nationaux et européens, toutes disciplines confondues, pour permettre la reproductibilité de la science en SO et le recensement de ce qui existe pour pouvoir le protéger, le valoriser, le partager via la rédaction de Plan de Gestion de données[15] par exemple peut et pourra influer sur ces pratiques. De même, on peut penser que les enjeux liés aux résultats des recherches (en médecine, dans les sciences du vivant) conduisent à avoir des pratiques de sauvegarde plus attentives.

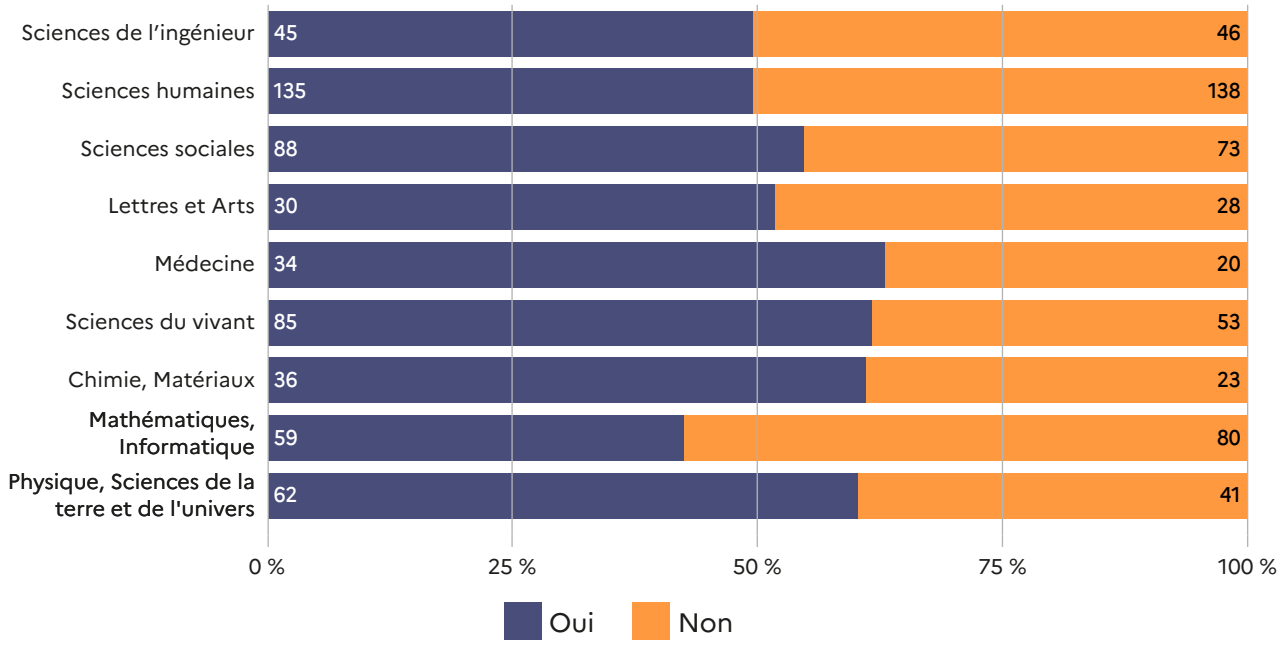


Figure 22: La pratique d'archivage à long terme selon la discipline

RECOMMANDATION : Inclure une question sur l'expérience dans la rédaction d'un plan de gestion de données et à quelle fin. Le contexte de travail a un impact assez net sur la pratique d'archivage à long terme.

Le travail collectif influe favorablement sur cette pratique. Le développement d'infrastructures de recherche ou leur accessibilité dans des contextes collectifs ont un effet d'entraînement sur cette pratique, au point que le recours à l'archivage se produit dans 8 cas sur 10 lorsque le répondant travaille dans un contexte regroupant plus de 20 personnes.

15. L'Agence Nationale de Recherche a rendu obligatoire l'élaboration d'un Plan de Gestion des Données depuis 2019, ce qui contraint les chercheurs à déterminer, entre autres, les moyens d'archivage des données produites lors d'une recherche financée par l'ANR.

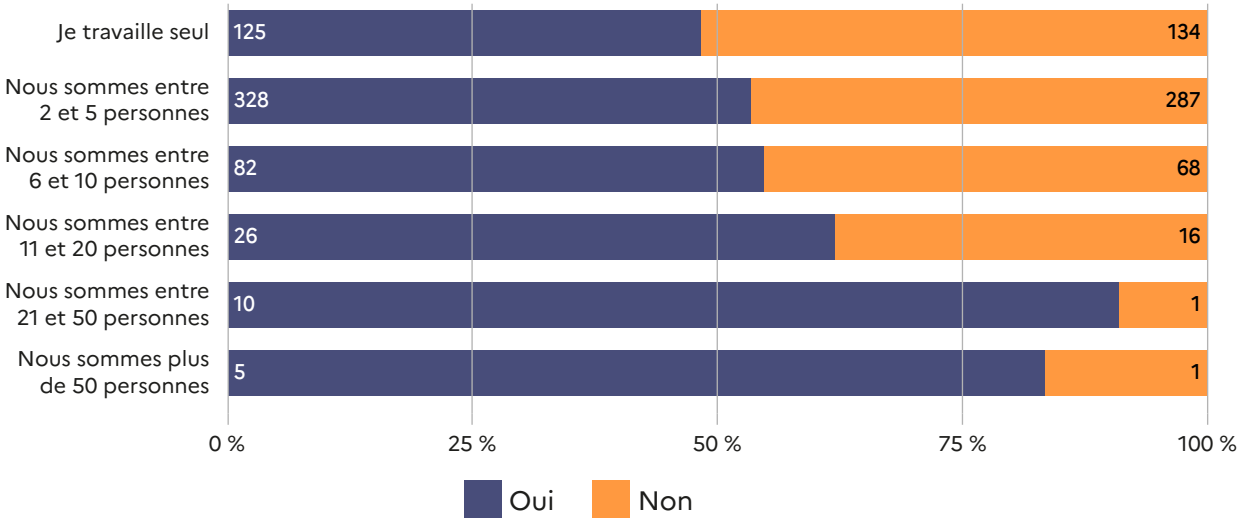


Figure 23: La pratique d'archivage à long terme selon le contexte de travail

La fonction exercée par le répondant est prégnante. Les chercheurs ayant des responsabilités (notamment en termes d'animation de la recherche comme les professeurs et les directeurs de recherches) sont plus sensibles à la question de l'archivage à long terme.

Leur position d'encadrement et de responsabilités dans les opérations de recherches les confronte plus fréquemment aux problématiques de l'archivage des données, sachant que les recommandations ou les obligations institutionnelles dans ce domaine sont croissantes.

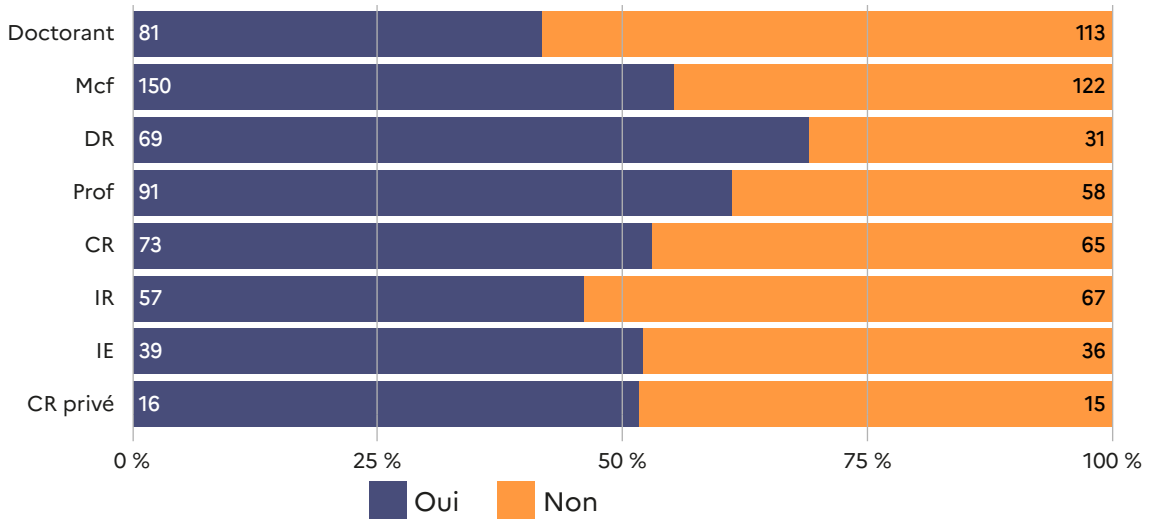


Figure 24: La pratique d'archivage à long terme selon la fonction

Un statut stable et permanent peut être un élément favorable à la pratique de l'archivage à long terme (58,7 % de réponse positive). Les répondants aux statuts non-permanents ont une pratique d'archivage à long terme minoritaire (49,3 %) et éloigné de 4,6 points de la moyenne de la POP. L'écart est encore plus conséquent pour ceux qui bénéficient d'un contrat doctoral (37,4 %), car environ 4 sur 10 disent pratiquer ce type d'archivage, alors que ce sont des chercheurs qui entrent dans la carrière de chercheur, un moment propice à la diffusion de cette pratique.

Sans doute les résultats sont influencés par un autre variable, celle de la classe d'âge du répondant, minorant le poids de la variable statutaire. Mais, le tableau croisant l'année de naissance du répondant à sa pratique d'archivage fait ressortir que cette tendance ne s'applique pas uniquement à la convergence entre un statut et un âge. Car plus le répondant fait partie des jeunes générations de chercheurs, plus sa pratique en matière d'archivage s'affaiblit, au point qu'elle finit par ne concerner que 37,1% de ceux nés après 1995 (dont les ¾ sont des doctorants). À l'inverse, cette pratique devient majoritaire à partir des générations nées avant 1975.

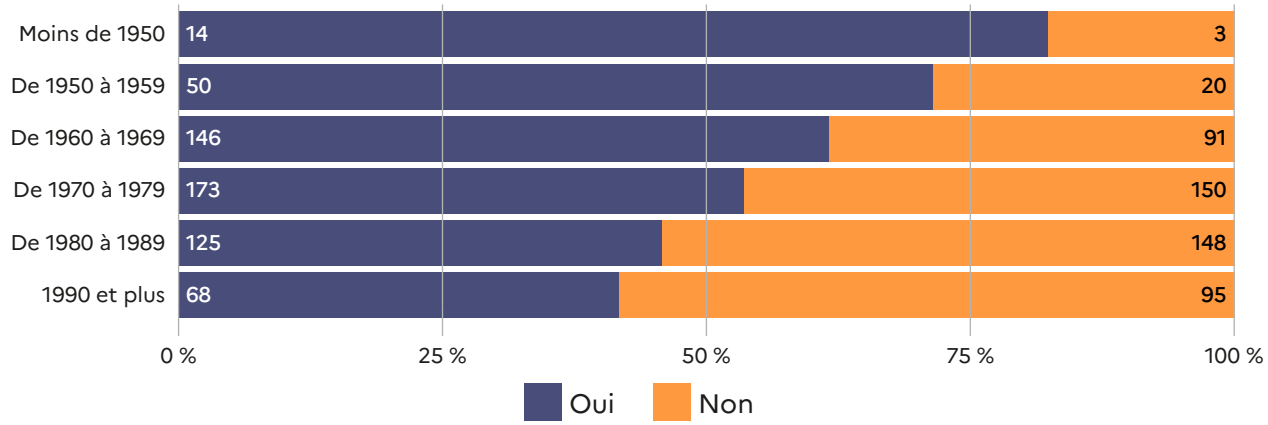


Figure 25: La pratique d'archivage à long terme selon la classe d'âge

Dans la population étudiée, la pratique d'archivage à long terme est associée à des chercheurs installés, statutairement et du point de vue de la carrière. On peut alors douter que la pratique d'archivage se diffuse avec les jeunes générations pris dans une course à la recherche de financements.

RECOMMANDATION : développer une sensibilisation, une information et des formations auprès des jeunes chercheurs.

Deux interprétations doivent être explorées:

- La question du partage et de la valorisation académique et économique des données;
- La question de l'accessibilité (informationnelle et matérielle) aux infrastructures.

La pratique de l'archivage à long terme doit beaucoup à la fois au contexte de travail et au statut du chercheur, pas uniquement à son adhésion ou sa proximité avec les valeurs de la science ouverte. La question touche à celles des moyens mis à dispositions des chercheurs mais également aux canaux par lesquels ceux-ci prennent connaissance des moyens pour enregistrer à long terme et de manière sûre leurs données.Si ces données peuvent déboucher sur des valorisations économiques ou académiques, il est possible que cette sensibilité soit accrue chez le chercheur.

La mobilisation des langages de description des données

QUESTION 22: Utilisez-vous des outils ou des langages pour qualifier ou décrire vos données (ex. Thésaurus, référentiels, métadonnées)?

Dans le cadre des principes FAIR, cette question permet de connaître l'usage des métadonnées qui permettent de décrire les données de recherche.

La présence de 65,7 % des répondants qui ne le fait jamais montre que cette pratique n'est pas majoritaire. Mais avec 13,1 % d'usage fréquent, il n'est pas possible d'indiquer si les répondants ont une familiarité avec cette pratique de documentation des données de recherche.

RECOMMANDATION : reproduire cette question pour voir si les usages se développent et décliner la question sur les standards de métadonnées usités pour essayer de voir si les répondants les utilisent selon des distinctions disciplinaires.

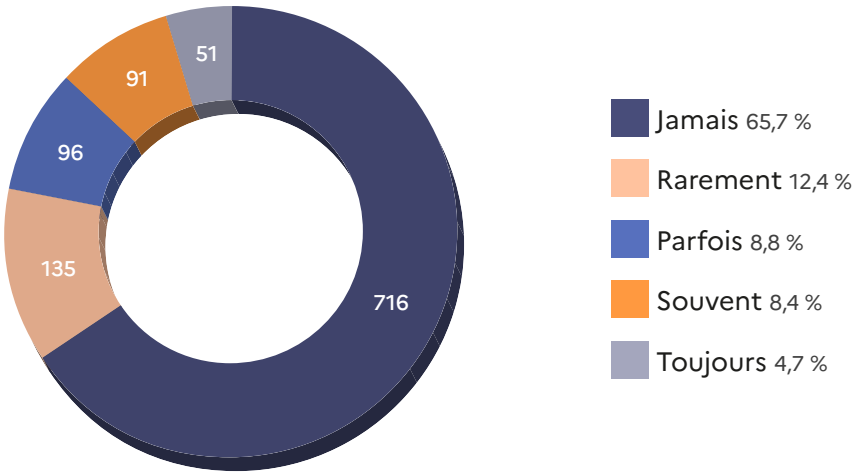


Figure 26: Répartition de l'usage des métadonnées

Le tiers de confiance pour l'archivage des données de travail du chercheur

QUESTION 15BIS: Si oui, à quel tiers les confiez-vous ?

HYPOTHÈSE : l'archivage des données ne touche pas à un sentiment de propriété mais est liée à des valeurs.

Sur les 576 personnes qui ont répondu « oui » à la question sur la pratique de l'archivage des données à long terme, 491 ont renseigné la question du tiers d'archivage, soit 85 %. Les 15 % restant correspondent à des non-réponses.

Il apparaît 4 types de tiers :

- Le tiers d'archivage qui n'en est pas un, dans le sens où c'est le chercheur ou son matériel qui est le lieu de sauvegarde : clef USB, ordinateur, disque dur externe, etc. Cela représente 150 réponses soit 30,5 %. On pourrait y ajouter les 62 réponses qui expliquent ne pas recourir à un tiers et dont on peut penser qu'il s'agit de personnes qui réalisent des sauvegardes personnelles. On parvient alors à une pratique qui se situe à 43,1 % des répondants ;
- Vient ensuite le tiers d'archivage qui s'oriente vers les solutions institutionnelles proposées par les universités ou les grands organismes de recherche : 34 %. Il faudrait peut-être y ajouter les

répondants qui déclarent se tourner vers les solutions proposées plus localement, par un réseau interne au laboratoire, mais également par un système de sauvegarde collaboratif : 17,7 %. L'utilisation de solutions institutionnelles (de proximité variable) est autour de 1 chercheur sur 2 (51,7 %) ;

- La part des recours à d'autres solutions, externes aux institutions ou à son environnement de travail, s'élèvent à 11 % ;

Enfin, une pratique assez rare mais tout de même notable consiste à recourir aux services de son entourage familial ou amical (2,4 %).

Tiers évoqués par les 491 répondants (taux de réponse : 45,1%)	Nombre	% Obs.
Institution	167	34,00 %
Équipe de recherche ou laboratoire	87	17,7 %
Conservation personnelle	150	30,5 %
Aucun tiers (supposée ici personnelle)	62	12,60 %
Service de sauvegarde de sociétés privés (Google Drive, Claudia, DropBox.)	54	11,00 %
Réponses inexploitables	25	5,10 %
Amis, famille, parents	12	2,40 %
NAS	8	1,60 %
Zenodo	7	1,40 %
GitHub	4	0,80 %

TABLEAU 25 : Le tiers de confiance pour l'archivage des données de travail du chercheur

% Obs.	Permanents	Doctorant/ Post-docrant	Non-permanents	Travailleurs indépendants
Institution	33,4 %	17,00 %	39,60 %	28,80 %
Conservation personnelle	29 %	34 %	18,8 %	42,9 %
Équipe de recherche ou laboratoire	16,1 %	17 %	20,8 %	14,3 %
Aucun tiers (supposée ici personnelle)	13,4 %	2,70 %	10,40 %	0,00 %
Service de sauvegarde de sociétés privés (Google Drive, Claudia, DropBox.)	8 %	28,30 %	10,40 %	14,30 %

TABLEAU 26 : Tiers d'archivage mobilisé en fonction du statut

Reste que ces résultats révèlent de grandes tendances à propos de l'archivage des données à long terme.

La conservation personnelle est assez répandue, un peu moins d'un tiers de ceux qui déclarent pratiquer l'archivage. Une part non négligeable des répondants déclare archiver à long terme dans un espace social et technologique réduit. Élargie à l'environnement direct du chercheur, son laboratoire, la pratique d'archivage de proximité regroupe 48,2 % des répondants. L'archivage à long terme ne passe pas par l'institution pour une majorité de ceux qui déclarent avoir cette pratique. Ainsi, l'archivage qui utilise des outils ou des clouds institutionnels représente une forte minorité de ceux qui sauvegardent, avec 23,3 % des répondants qui le font (moins d'un quart de la POP).

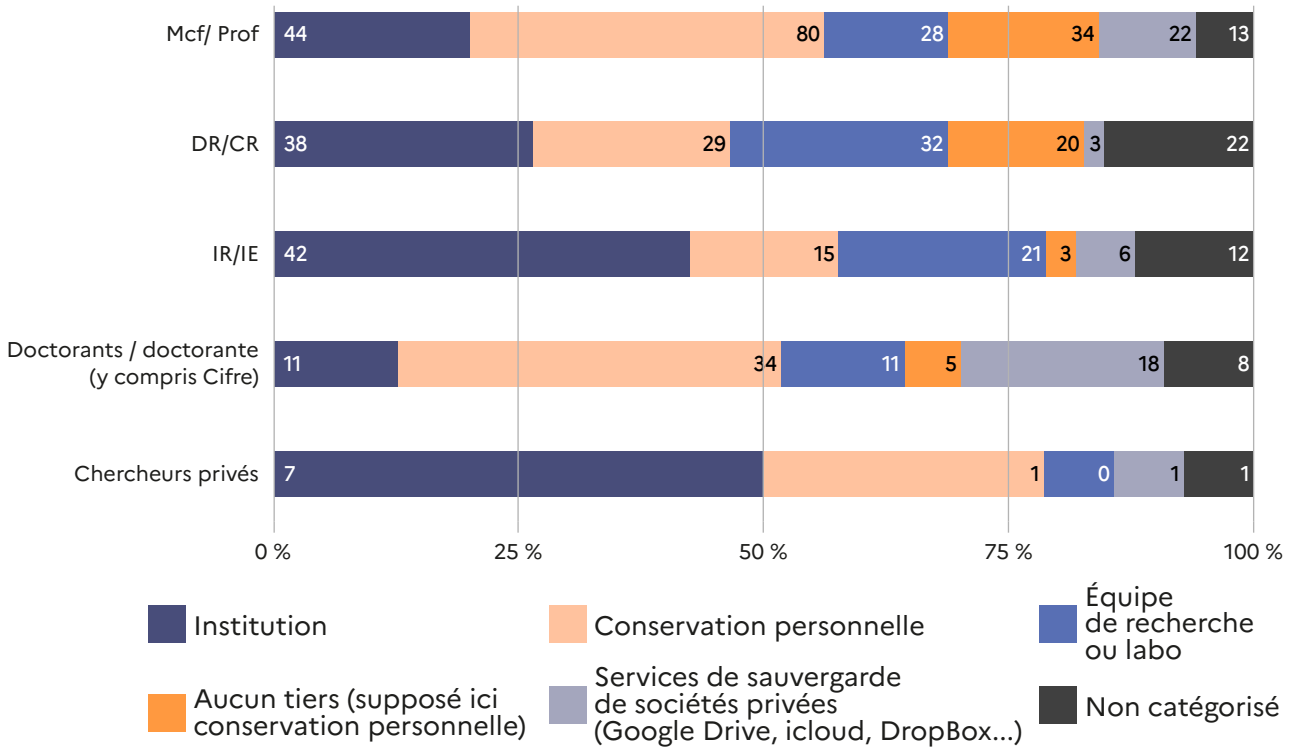


Figure 27: Tiers d'archivage utilisé en rapport avec la fonction

Contexte de travail

L'usage d'archivage institutionnel est nettement plus pratiqué par les non-permanents et les chercheurs stables, mais bien moins par les travailleurs indépendants et surtout les doctorants et post-doctorants. Ils utilisent des services de sauvegarde de sociétés privées et commerciales.

Il existe visiblement une distinction entre les secteurs de la recherche, l'université d'une part, et les chercheurs appartenant aux grands organismes de recherche et les chercheurs du privé d'autre part. Les premiers font moins appel aux ressources institutionnelles et optent plus souvent pour une conservation personnelle ou sans tiers : un tiers des universitaires opte pour une conservation personnelle. Les doctorants et post-doctorants ajoutent une appétence particulière supplémentaire pour les solutions de sauvegardes proposées par des sociétés privées et commerciales.

Les DR/CR utilisent les ressources de leurs laboratoires. Les IR et IE font encore plus confiance aux ressources institutionnelles, même si une part non négligeable d'entre eux fonctionne sur des ressources personnelles ou des prestataires externes. Enfin, les chercheurs du privé font nettement confiance aux ressources de leur entreprise ; il faut sans doute penser que parfois il s'agit d'une obligation contractuelle.

Ce graphique laisse supposer que les ressources institutionnelles pourraient être plus utilisées par les chercheurs. Pour les universitaires, il se peut qu'il soit question de moyens à disposition et/ou celui de la connaissance et de l'accessibilité à ces moyens (accompagnement dans l'installation, machine connectée au réseau local). Travailler dans un laboratoire du CNRS ou d'un grand organisme de recherche est lié à une pratique plus assidue de l'archivage à long terme.

Être plus ou moins isolé a un impact fort sur la pratique d'archivage à long terme lorsqu'elle existe. Plus le contexte est collectif (et ce collectif important) plus la sauvegarde institutionnelle est forte. À l'inverse, plus la personne travaille de manière isolée, plus elle aura une tendance à la conservation personnelle et le cas échéant à utiliser des prestataires commerciaux ou à avoir recours à un réseau social de proximité.

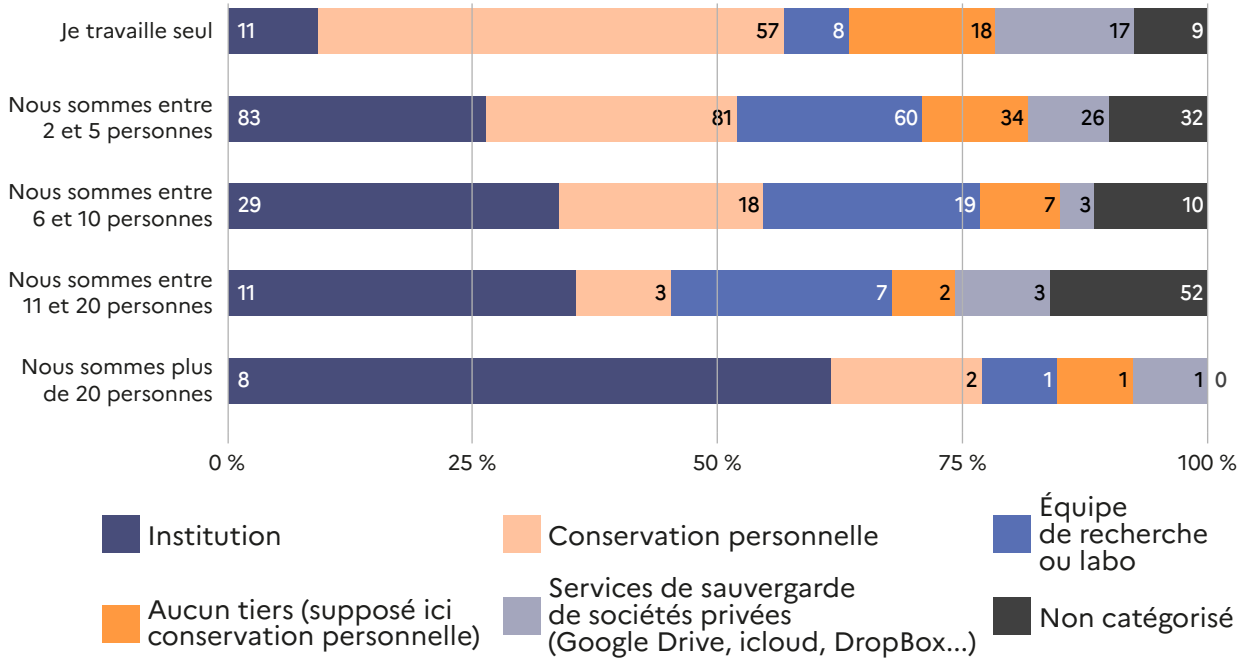


Figure 28 : Tiers d'archivage utilisé en rapport avec le contexte de travail

Distinctions disciplinaires

Des distinctions disciplinaires apparaissent. Les chercheurs des lettres, des sciences humaines et des sciences de l'ingénieur utilisent assez faiblement les ressources institutionnelles au contraire des biologistes, médecins, mathématiciens et informaticiens qui le font nettement plus souvent. La conservation personnelle est très présente chez les spécialistes des sciences humaines, mais elle semble surtout s'affaiblir lorsque les ressources institutionnelles sont préférées (sciences du vivant, médecine, mathématiques et informatique). Les solutions des sociétés

privées semblent avoir plus d'échos chez les spécialistes des lettres et des sciences de l'ingénieur. Les solutions proposées par les laboratoires (qui sont sollicités par 17,7 % des répondants à cette question) sont visiblement appréciées par les physiciens et les biologistes mais très ignorées par les SHS.

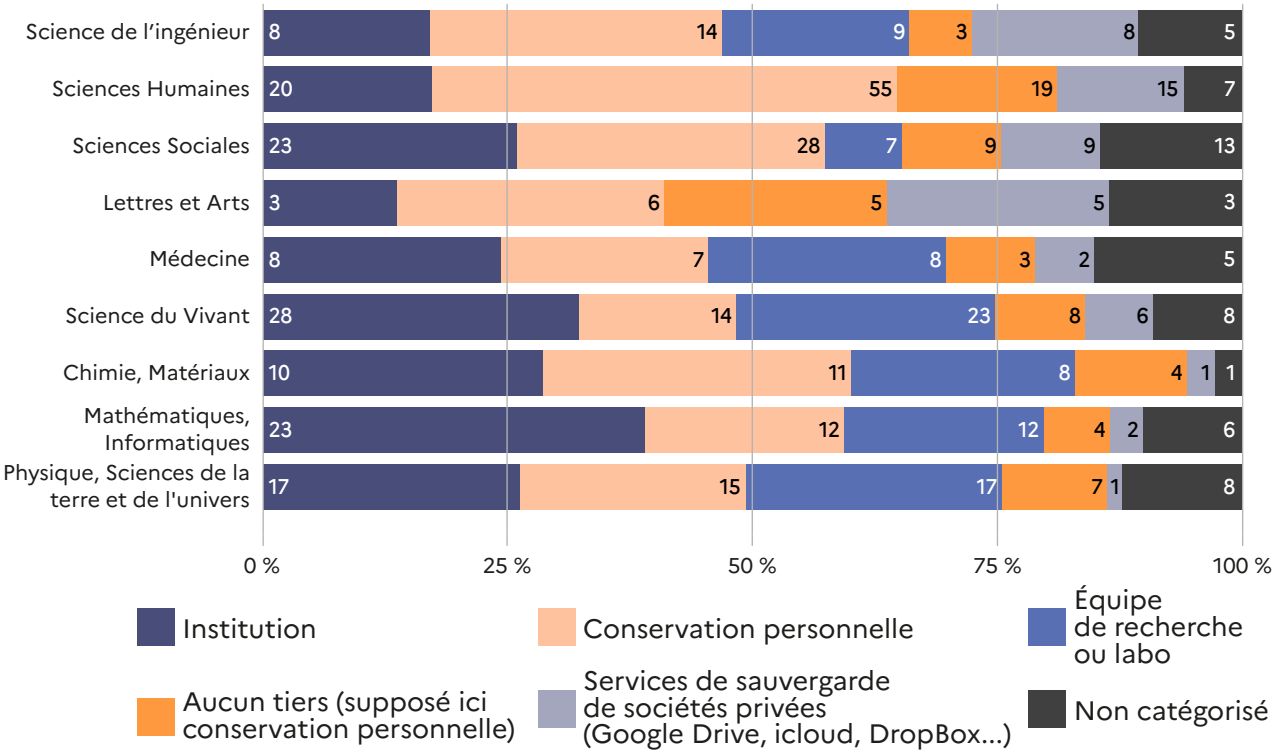


Figure 29: Tiers d'archivage utilisé en rapport avec la discipline

L'utilisation de l'archivage est le fait de personnes intégrées ou en souhait d'intégration dans les institutions de recherche. Mais sur ce point il y a une différence entre le monde de l'université et celui des organismes de recherche, le monde universitaire semblant en retrait dans cette pratique.

RECOMMANDATION : veiller à l'accompagnement dans l'installation, au-delà de la formation et de l'information.

Les pratiques d'enregistrement et de sauvegarde des données

HYPOTHÈSE : les statuts influent sur l'accès aux solutions institutionnelles d'enregistrement et de sauvegarde des données.

Concernant l'enregistrement et la sauvegarde des données, les résultats sont assez conformes avec les différentes études et recherches que l'on peut consulter à ce propos (Serres et al., 2017). Le principal support de sauvegarde demeure le disque dur qu'il soit interne et/ou externe suivi des supports amovibles. Ces résultats laissent entendre que la question de la pérennité des sauvegardes réalisées par de nombreux

chercheurs n'est pas la première considération prise en compte. Le disque dur est devenu l'élément de sauvegarde basique associé à d'autres solutions qui viennent en complément ou en soutien.

L'usage des solutions de sauvegarde proposés par les institutions de recherche est déclaré par 48,1% des répondants. Ce taux paraît assez faible, compte tenu des efforts consentis par les institutions (universités, grands organismes de recherche public) pour offrir des solutions à leurs chercheurs (Banat-Berger et al., 2009). L'interprétation n'est pas aisée. S'agit-il d'une méconnaissance, d'une difficulté d'accès ou d'installation ? Le chercheur avec un statut permanent déclare un usage légèrement plus important des enregistrements et sauvegardes (49,8%) *via* un réseau informatique local proposé par les universités ou les grands organismes de recherches, ce qui ne va pas dans le sens d'un rejet. À l'inverse, les chercheurs relevant d'un statut non-permanent sont eux moins enclins à cet usage (41,4%). [L'hypothèse d'un manque d'intégration (ou d'accès à l'information) envers les ressources proposées par leurs institutions peut s'expliquer par une différence de statut.]

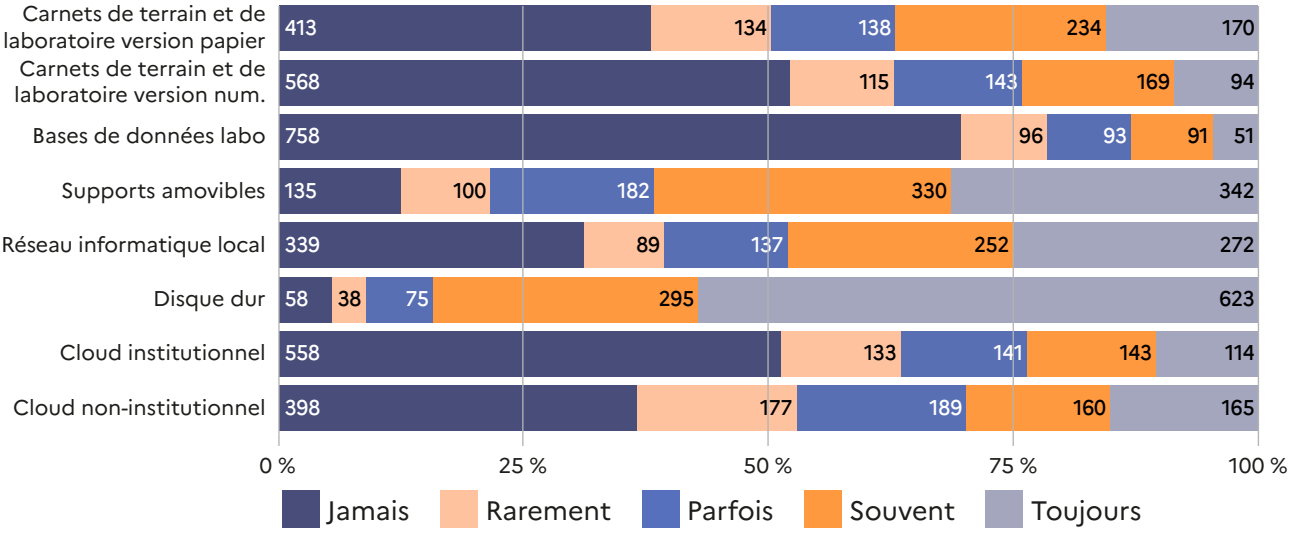


Figure 30: Répartition des outils d'enregistrement et de sauvegarde des données

Les autres formes de sauvegarde sont nettement plus confidentielles. Elles ne parviennent pas à une fréquence («toujours» et «souvent») supérieure à 40%: les carnets de laboratoire numérique et papier, les bases de laboratoire et les *clouds*. Notons que les *clouds* institutionnels sont moins utilisés par les chercheurs que leurs homologues non institutionnels (GoogleDrive, DropBox…).

L'usage des *clouds* non institutionnels est discutable en termes de protection des données, mais ils sont globalement légèrement plus utilisés que les *clouds* institutionnels. Leur usage fluctue selon le statut du répondant : si pour 28,8% des chercheurs bénéficiant d'un statut permanent le recours à ces *clouds* est fréquent, ce sont 30,7% des chercheurs au statut non-permanents et 38,4% des bénéficiaires d'un contrat doctoral qui ont la même pratique. Concernant l'utilisation des *clouds* institutionnels[16], si la forte fréquence d'usage est moindre, elle est

16. Les chercheurs peuvent utiliser à la fois des *clouds* institutionnels et non institutionnels.

d'autant plus faible que le répondant est dans l'un de ces deux derniers statuts (11 % d'usage pour les contrats doctoraux contre 1 chercheur stable sur 4). Les écarts sont également d'environ 10 points de pourcentage à propos de l'usage fréquent des réseaux informatiques locaux. Dès lors, tout se passe comme si le fait d'appartenir à un statut non-permanent ou dans une situation de contrat doctoral éloignait d'un usage des outils d'archivage que proposent les institutions de recherche au profit de solutions non institutionnelles.

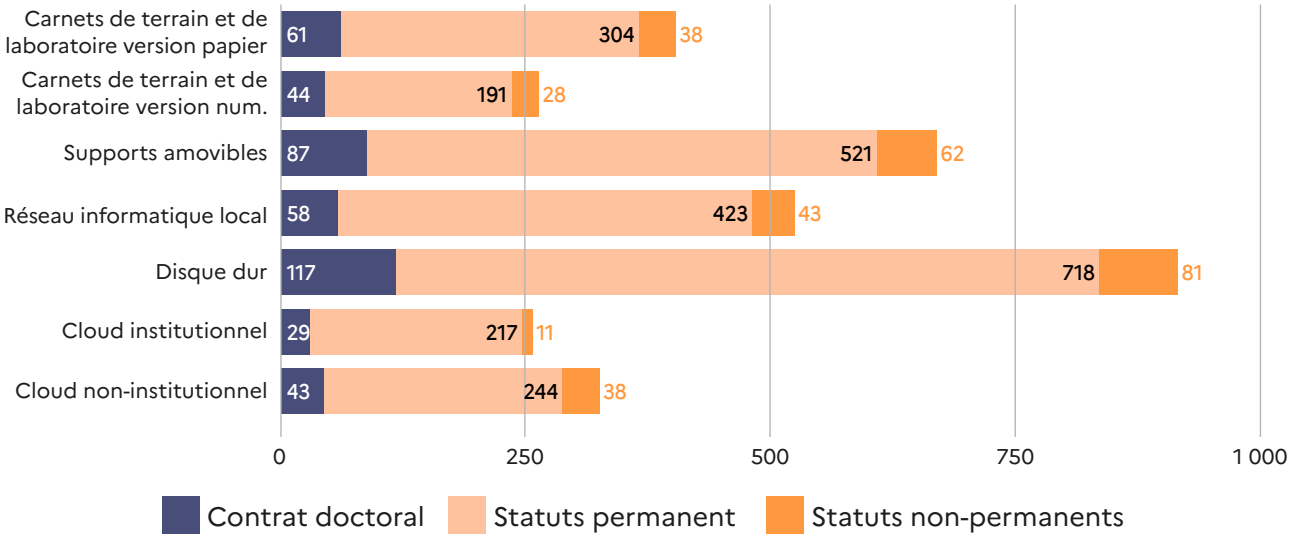


Figure 31: Répartition de la pratique fréquente des outils d'enregistrement et de sauvegarde selon le statut

6. PARTAGE DES DONNÉES ET DES RÉSULTATS DE LA RECHERCHE

QUESTION 34 : Selon vous, est-il souhaitable que les données de recherche que vous avez produites ou contribué à produire soient partagées ?

HYPOTHÈSE : il existe différents positionnements possibles envers les pratiques de partage des données et des résultats de la recherche.

Dans le tableau récapitulatif ci-dessous, les différentes modalités représentent des sphères de partage de plus en plus large, allant du collègue avec lequel je travaille à l'ensemble des citoyens. L'échelle permet alors de situer jusqu'à quel niveau le répondant adhère à ce partage.

Nous pouvons constater une fracture assez nette. La grande majorité des répondants s'est positionnée favorablement pour le partage au sein de la sphère académique (3 premières modalités) : le taux est toujours supérieur à 80 % d'opinion favorable et atteint même 97,8 % pour la première modalité.

Partage des données produites	% Cit. Oui	% Cit. Non
Entre chercheurs avec qui j'ai un partenariat	97,80 %	2,20 %
Entre chercheurs de la même discipline ou domaine de recherche	92,20 %	7,80 %
Entre chercheurs du même pays ou du même continent	83,10 %	16,90 %
Avec des membres du domaine économique et médiatique	63,60 %	36,40 %
Avec des membres du secteur associatif et/ou caritatif	70,20 %	29,80 %
Avec tout citoyen, sans restriction professionnelle ni géographique	67,40 %	32,60 %

TABLEAU 27 : Souhait de partage des données et résultats de la recherche

Description de la sous-population avec une opinion défavorable (n = 457)

Pas une distinction disciplinaire

La discipline ne participe pas à une opinion défavorable. Les répondants avec une opinion défavorable quant à une diffusion large se répartissent de manière très proche avec celle de la population totale, à la légère exception des mathématiques et informatique. Cette sous-population est assez hétérogène du point de leurs caractéristiques sociologiques.

Focus

Pour comprendre et connaître un peu mieux cette sous-population (n = 457), nous l'avons divisée en deux parties : d'un côté celles et ceux qui ont répondu négativement à l'une des 3 premières modalités et de l'autre celles et ceux qui se sont positionnés négativement sur l'une des 3 dernières modalités. Nous nommerons les premiers les « défavorables au partage au sein du monde de la recherche » (n = 16), les seconds étant dénommés les « défavorables au partage en dehors du monde de la recherche » (n = 267). Les « défavorables au partage au sein du monde de la recherche » sont faibles et nous ne les décrivons pas.

Notre description porte sur l'opinion défavorable du partage avec des personnes qui n'appartiennent pas au monde académique que nous allons observer selon trois variables : âge, statut, disciplines et critères dans les choix de publication.

Il est à noter une surreprésentation des 50-55 ans et une sous-représentation des moins de 35 ans (Tableau 29). Cette tendance se confirme logiquement avec le critère du statut : les répondants sont plus souvent permanents, les doctorants, post-doctorants et non-permanents étant quant à eux moins présents (Tableau 30). Cette population bien ancrée dans le monde de la recherche craint le partage de la recherche à l'extérieur du monde académique. La distinction selon les disciplines n'apporte pas d'explication plus précise, la répartition entre cette sous-population et l'échantillon total étant assez proche (Tableau 31). Certes, une présence moindre des chercheurs des sciences humaines ou des mathématiques/informatiques renforce une crainte légèrement plus marquée pour les sciences sociales ou la physique mais les écarts (de quelques points de pourcentage) conduisent à manier ces résultats avec prudence, surtout que les liens ou conditions communes explicatives qui rassemblent les chercheurs de ces disciplines sont difficiles à saisir.

Répartition de la sous-population de répondants défavorables au partage des données de la recherche

Modalités	Défavorable au partage au sein du monde de la recherche	Défavorable au partage en dehors du monde de la recherche	Défavorable au partage	Défavorable exprimé au moins une fois
Entre chercheurs avec qui j'ai un partenariat	16 (1,5 %)			
Entre chercheurs de la même discipline ou domaine de recherche				
Entre chercheurs du même pays ou du même continent				
Avec des membres du domaine économique et médiatique		267 (24,5 %)	14 (1,3 %)	457 (41,9 %)
avec des membres du secteur associatif et/ou caritatif				
avec tout citoyen, sans restriction professionnelle ni géographique				

TABLEAU 28 : Répartition de la sous-population de répondants défavorables au partage des données de la recherche

Il apparaît que la solution ou l'explication réside, en partie, dans les contraintes de publication, voire d'évaluation qui s'imposent aux chercheurs, sachant que nous avons surtout des enseignants-chercheurs et des chercheurs (Tableau 32). En matière de publications, nous remarquons que la préférence des chercheurs étudiés (réponses **QUESTION 26**) va vers des revues à fort impact (*Impact factor*, SIGAPS…). Ce positionnement se double d'une réticence à l'égard de la science ouverte et des incitations à publier en Open Access.

- Dans les questions ouvertes certains acteurs ont exprimé une tension entre la volonté de partager des résultats au-delà du monde de la recherche et les critères qui interviennent dans l'évaluation des productions de la recherche (essentiellement les publications). Cette tension aboutit souvent au refus de partager des données de recherche.

- Les répondants de cette sous-population expriment sans doute une crainte plus qu'une position de principe. Ils interrogent les conditions dans lesquelles et

par lesquelles un partage peut se réaliser, conditions qui ne sont pas toujours clairement connues ni cadrées. De plus, ces conditions viennent se heurter à des exigences, où l'évaluation du travail de recherche valorise peu les pratiques (notamment de publications) qui s'appuient et favorisent un partage de principe des résultats de la recherche le plus large possible.

RECOMMANDATION : Il serait intéressant de pouvoir distinguer les résultats des répondants selon leur appartenance à un établissement ayant signé la déclaration San Francisco *Declaration on Research Assessment*; DORA.

Année de naissance (Moyenne = 1974,27; Médiane = 1974)	Nombre	% Obs.	Ech.
Moins de 1950	5	1,90%	1,60%
De 1950 à 1954	3	1,10%	1,00%
De 1955 à 1959	12	4,50%	5,60%
De 1960 à 1964	29	10,90%	9,00%
De 1965 à 1969	51	19,10%	12,90%
De 1970 à 1974	40	15,00%	14,90%
De 1975 à 1979	39	14,60%	14,90%
De 1980 à 1984	42	15,70%	14,00%
De 1985 à 1989	14	5,20%	11,20%
De 1990 à 1994	23	8,60%	11,80%
1995 et plus	9	3,40%	3,20%
Total	**267**	**100,00%**	**100,00%**

TABLEAU 29 : Répartition par date de naissance de la sous-population de répondants défavorables au partage des données de la recherche

Statut	Nombre	% Obs.	Ech.
Doctorants/Post-doctorants	24	9,00%	12,30%
Non-permanents	18	6,70%	9,40%
Permanent	221	82,80%	75,80%
Travailleur indépendant	3	1,10%	1,80%
Total	**263**	**99,60%**	**99,40%**

TABLEAU 30 : Répartition par statut de la sous-population de répondants défavorables au partage des données de la recherche

Disciplines	Nombre	% Obs.	Ech.
Sciences humaines	55	20,90%	25,10%
Sciences sociales	43	16,30%	14,80%
Sciences du vivant	42	16,00%	12,70%
Physique, Sciences de la terre et de l'univers	32	12,20%	9,50%
Sciences de l'ingénieur	26	9,90%	8,40%
Mathématiques, Informatique	25	9,50%	12,80%
Chimie, Matériaux	18	6,80%	5,40%
Médecine	14	5,30%	5,00%
Lettres et Arts	8	3,00%	5,30%
Total	**263**	**100,00%**	**98,80%**

TABLEAU 31 : Disciplines des répondants (n = 263) défavorables au partage des données de la recherche

Critères choix de publication	Nombre	% Obs.	Imp.	Ech.
Notoriété de la revue, de l'éditeur, du congrès dans ma discipline	249	93,30%	3,38	91,60%
Indicateurs chiffrés de la revue (Facteur d'impact, JCR, SIGAPS…)	141	**52,80%**	1,53	43,20%
Critères HCERES, FNGE, CNRS, CNU…	70	26,20%	0,69	24,50%
Frais de publications (Article Processing Charges)	89	33,30%	0,73	31,60%
Transparence du processus d'évaluation	78	29,20%	0,68	34,20%
Diffusion en accès libre (choix personnel)	62	23,20%	0,52	39,50%
Refus des revues prédatrices	95	35,60%	0,76	33,80%
Diffusion en accès libre (imposé par mon institution, financeur)	27	10,10%	0,21	13,60%
Total	**267**			

TABLEAU 32 : Critères de choix pour soumettre une publication ou une communication **(QUESTION 26)** pour les avis défavorables à l'ouverture des données de recherche

7. DIFFUSION ET VALORISATION DES RÉSULTATS ET DES DONNÉES DE LA RECHERCHE

Les résultats sont issus des parties « Écrire », « Publier », « Diffuser, valoriser » les résultats de la recherche sous différentes formes ainsi que les actions ou supports de valorisation du travail de recherche.

Nous proposons d'aborder successivement :

- Les outils de publication ;
- Les formes de publication ;
- Les critères de publication ;
- La pratique des *preprints* ;
- La diffusion des données de la recherche en lien avec les résultats de la recherche ;
- La valorisation des résultats des recherches ;
- Les formes d'évaluation émergentes des publications.

Les outils de rédaction ou d'aide à la rédaction : le poids des outils de traduction en ligne

> **QUESTION 23** : Pour rédiger vos documents de recherche et publications, quels outils utilisez-vous ?

Les outils de traduction de type DeepL sont les troisièmes outils utilisés fréquemment après les outils de traitement de texte en local et les outils de références bibliographiques, avec 33 % d'usage fréquent « toujours et souvent) et 25,7 % d'usage intermittent (parfois). Il est intéressant de voir que les outils de traduction arrivent en 3e position, marquant de fait un manque de compétences linguistiques (des répondants) pourtant exigées par la « compétition internationale » qui se joue dans la science, et plus certainement dans l'évaluation de la recherche.

Les formes et outils de publication : l'hégémonie de l'article dans la revue à comité de lecture

Dans le monde académique, la publication reconnue est l'article dans une revue à comité de lecture. Elle est suivie de près par la communication dans un colloque et un congrès. En termes de fréquence d'usage, ces deux formats de publication regroupent ¾ des répondants pour la première et les ⅔ pour la seconde.

Les autres formats de publication sont nettement plus confidentiels. L'ouvrage regroupe à peine 1 répondant sur 10 pour un usage fréquent, un peu moins de 20 % pour le chapitre d'ouvrage. Les formats électroniques ou hors de revues scientifiques introduits dans le monde académique de manière plus récente ne sont pas plus utilisés. Écrire des billets de blog ou réaliser des articles dans la presse représentent au mieux un répondant sur 20 dans notre enquête. Le format d'article dans des revues avec évaluation par les pairs est hégémonique.

La constance des formes de publication malgré la transition numérique interroge, du moins montre une fois encore la prégnance du modèle traditionnel qui fait prévaloir l'évaluation quantitative des résultats de recherche par la notoriété du support à la dissémination plus large, plus ouverte des contenus comme sur un git (Tableau 36).

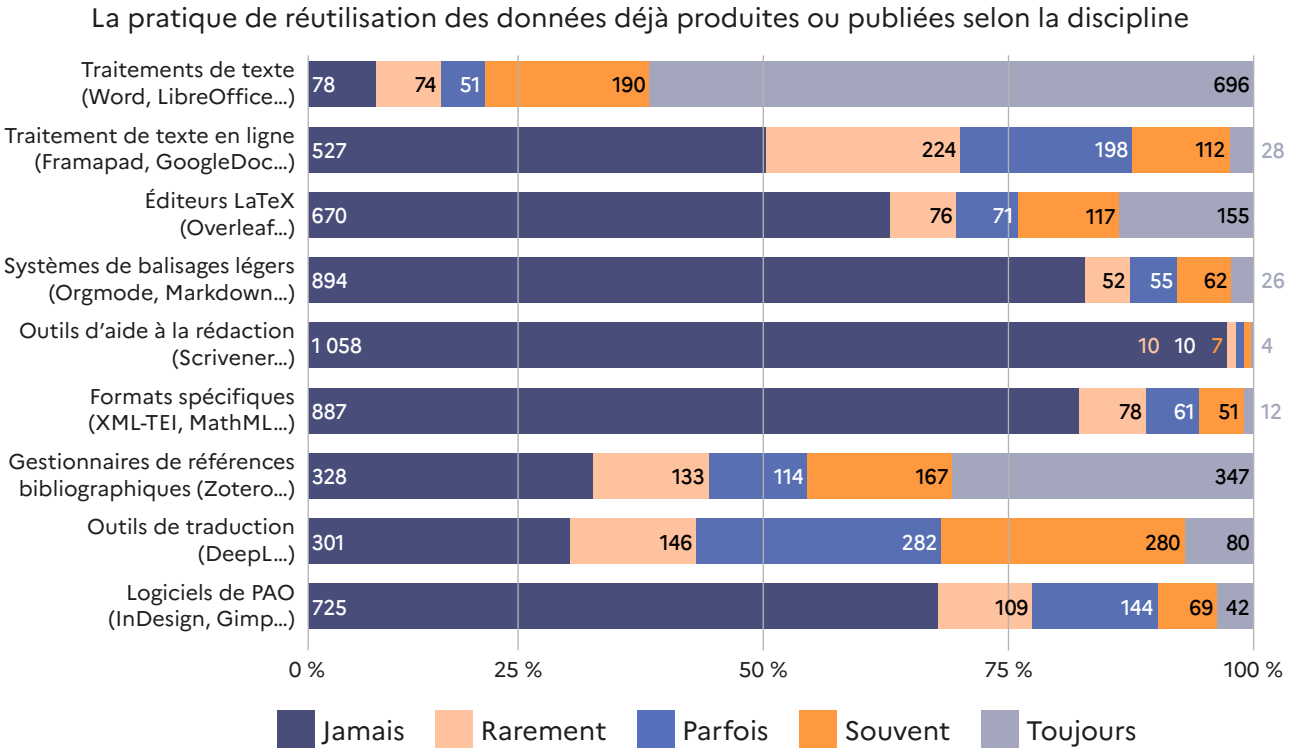


Figure 32: Les outils de rédaction ou d'aide à la rédaction des publications

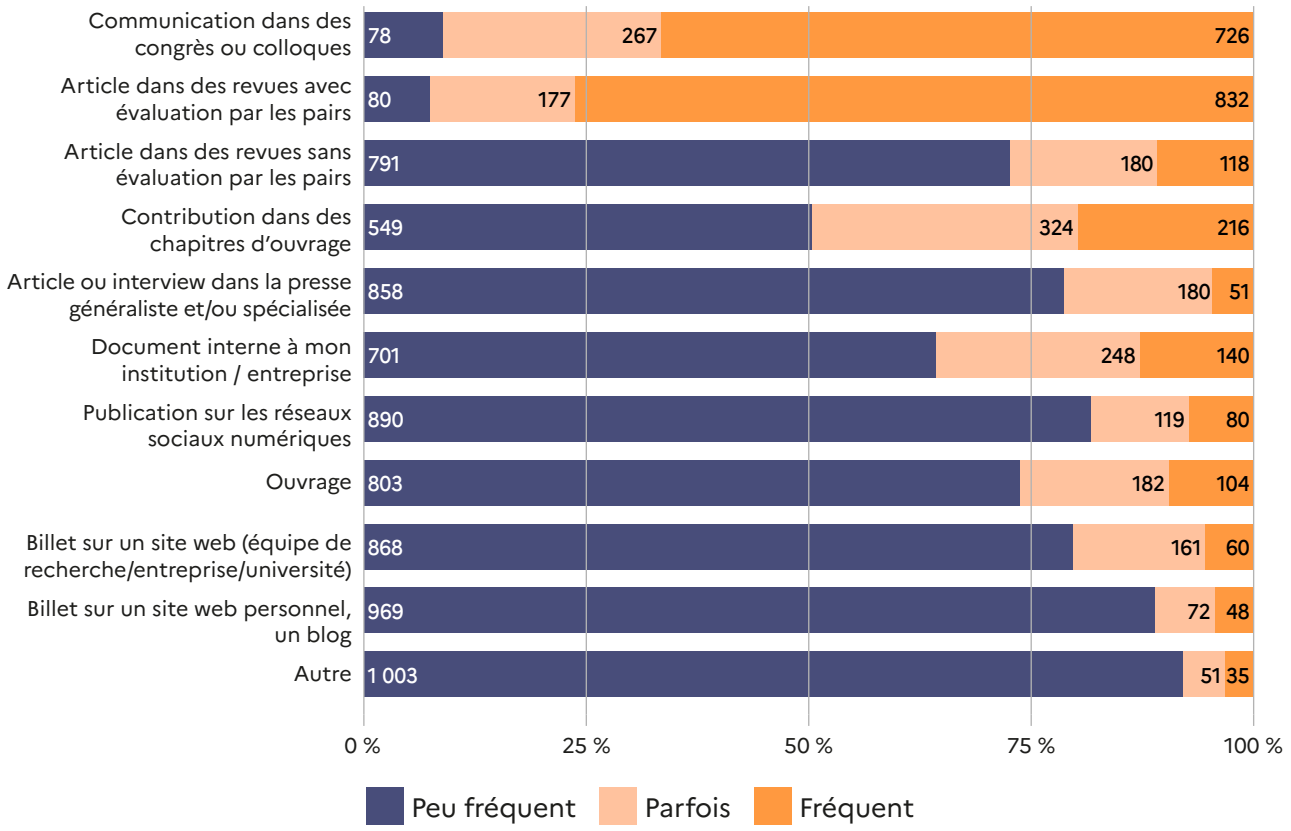


Figure 33: Choix des formats de publication des résultats de recherche

Discipline	% Cit. Peu Fréquent	% Cit. Parfois	% Cit. Fréquent
Sciences sociales	10,60%	24,20%	65,20%
Sciences humaines	9,90%	19,40%	70,70%
Sciences de l'ingénieur	15,40%	13,20%	71,40%
Lettres et Arts	1,70%	24,10%	74,10%
Mathématiques, Informatique	4,30%	14,40%	81,30%
Médecine	1,90%	13,00%	85,20%
Physique, Sciences de la terre et de l'univers	2,90%	11,70%	85,40%
Sciences du vivant	5,10%	9,40%	85,50%
Chimie, Matériaux	5,10%	5,10%	89,80%

TABLEAU 33: Fréquence de la publication dans une revue à comité de lecture selon la discipline

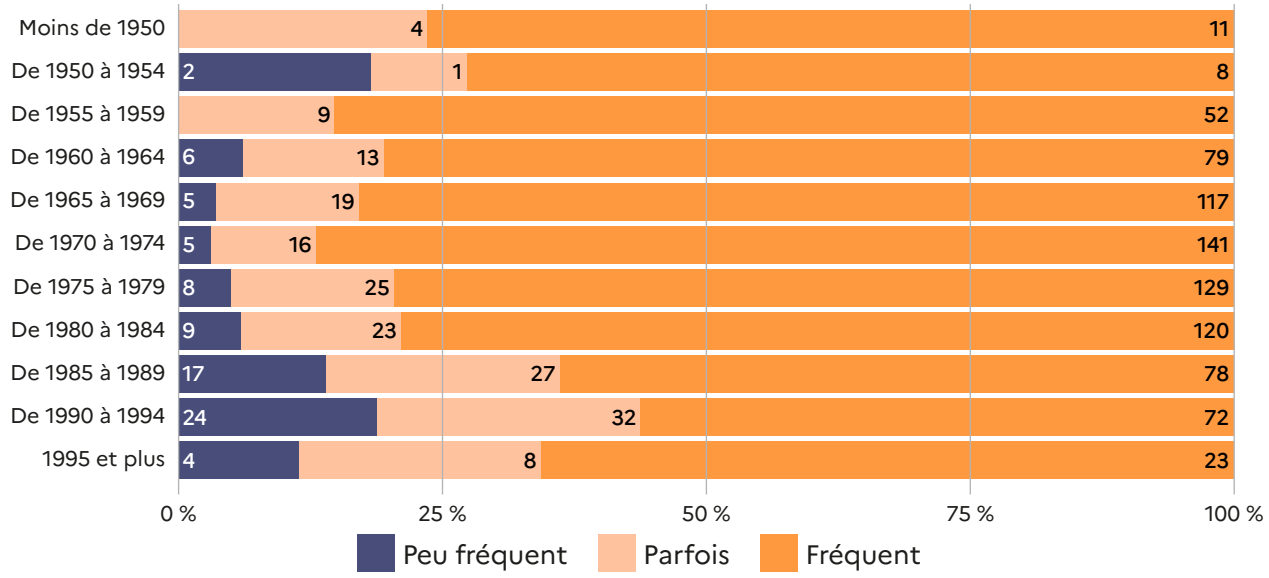


Figure 34: Fréquence de la publication dans une revue à comité de lecture selon la classe d'âge

Les distinctions disciplinaires sont présentes.

Les SHS, les lettres et les sciences de l'ingénieur se situent dans une utilisation moins marquée de ce format de diffusion, (en deçà de la moyenne de 76,4%, avec un écart de plus de 10 points pour les sciences sociales). Les autres disciplines sont plus en lien avec ce canal de diffusion qui est même quasi incontournable dans des disciplines comme la chimie, où on atteint des taux d'utilisation de près de 90%.

Le croisement avec l'année de naissance du répondant apporte quelques éléments explicatifs. Les plus jeunes générations

(moins de 35 ans) pratiquent moins ce type de diffusion. Ce constat peut traduire la difficulté que peuvent avoir ces jeunes générations à publier dans des revues à comité de lecture. Mes les résultats peuvent également traduire une évolution dans ces jeunes générations vers une moindre concentration envers ce type de publication.

La diffusion ne s'oriente pas vers les réseaux sociaux numériques. Par exemple, si l'on effectue une comparaison avec la pratique de publication sur les réseaux sociaux numériques, une répartition assez hétérogène en fonction des classes d'âge, avec une fréquence bien moindre chez les plus jeunes générations (rappelons que la moyenne pour notre population est de 7,3 % de pratique fréquente) est relevée.

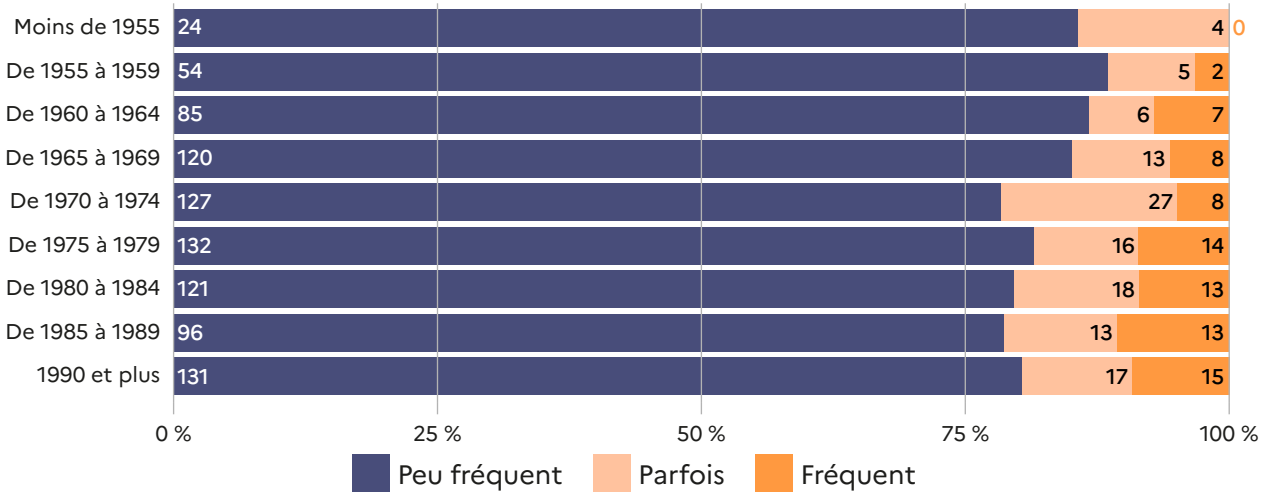


Figure 35: Fréquence de la publication sur les réseaux sociaux numériques selon la classe d'âge

Les critères de publication

> **QUESTION 26**: Quels sont vos critères de choix pour soumettre une publication ou une communication ?

Les critères de diffusion des résultats de recherche sont de deux ordres. Le premier critère constaté dans notre enquête tient à la notoriété du support (de la revue généralement) qui apparaît comme le critère essentiel (cité par 91,6 % des répondants). Tout se passe comme si, sans validation de ce critère, le support n'est pas pris en considération. Certes, la notoriété peut être considérée comme un critère issu de représentations partagées par un public (Alloing, 2017); mais il ne l'est pas pour les répondants qui souhaitent diffuser, qui connaissent les revues et qui répondent favorablement à ce critère. Le second critère est en réalité constitué d'un ensemble de critères, visiblement d'importance assez proche, à savoir 1) les indicateurs chiffrés et 2) la diffusion en libre accès de l'article (environ 40 %). Or ces deux critères sont assez distincts par leur nature, l'un révélant une adhésion à l'évaluation quantitative et algorithmique des travaux de recherche, l'autre à la philosophie de la Science Ouverte.

Enfin, d'autres critères revêtent une importance moindre que les précédents sans occulter leur capacité à créer des métriques aisément duplicables pour

les comparaisons de performance. Nous soulignons la présence des critères institutionnels quantifiables (HCERES, CNU…) qui ne semblent concerner qu'un quart des répondants et la faible adhésion à l'accès libre lorsqu'il est imposé par l'institution (13,6 %).

Critères	% Cit. Peu Fréquent	% Cit. Parfois	% Cit. Fréquent
Notoriété de la revue, de l'éditeur, du congrès dans ma discipline	997	91,60 %	3,28
Indicateurs chiffrés de la revue (Facteur d'impact, JCR, SIGAPS…)	470	43,20 %	1,22
Critères HCERES, FNGE, CNRS, CNU…	267	24,50 %	0,64
Frais de publications (Article *Processing Charges*)	344	31,60 %	0,69
Transparence du processus d'évaluation	372	34,20 %	0,79
Diffusion en accès libre (Choix personnel)	430	39,50 %	1
Refus des revues prédatrices	368	33,80 %	0,75
Diffusion en accès libre (Imposé par mon institution, financeur)	148	13,60 %	0,3
Total	**1089**		

TABLEAU 34 : Critères de choix pour soumettre une publication ou une communication

La pratique des *preprints* ou prépublication

QUESTION 27 : Diffusez-vous des prépublications (*preprints*) non évalués de vos travaux ?

La pratique des *preprints*, *working papers* ou prépublication a été importante durant la crise Covid-19 (Fraser et al, 2020 ; Bordignon et al, 2021). Pendant la crise Covid-19, les prépublications ont été déposées sur des serveurs (par exemple arXiv) notamment par des auteurs qui ne le pratiquaient pas précédemment (Fraser et al, 2020). La diffusion des *preprints* est une pratique qui produit des débats, que nous ne pouvons résumer ici, tout en étant une solution pour publiciser plus rapidement ces résultats.

Au regard de nos résultats, cette pratique demeure très limitée : au mieux un répondant sur cinq diffuse fréquemment sur des plateformes *not for profit, de type serveurs de preprint comme ArXiv, OSF, HAL, RePec, Zenodo…* Elle est encore plus confidentielle sur des réseaux sociaux académiques ou personnels.

Canal de diffusion	% Cit. peu fréquent	% Cit. parfois	% Cit. fréquent
Sur des plateformes « *not for profit* »	68,60 %	11,20 %	20,20 %
Sur des réseaux sociaux numériques académiques	81,50 %	10,40 %	8,10 %
Sur un web site personnel	87,10 %	4,30 %	8,60 %

TABLEAU 35 : Canaux de diffusion des prépublications non-évaluées (*preprint*)

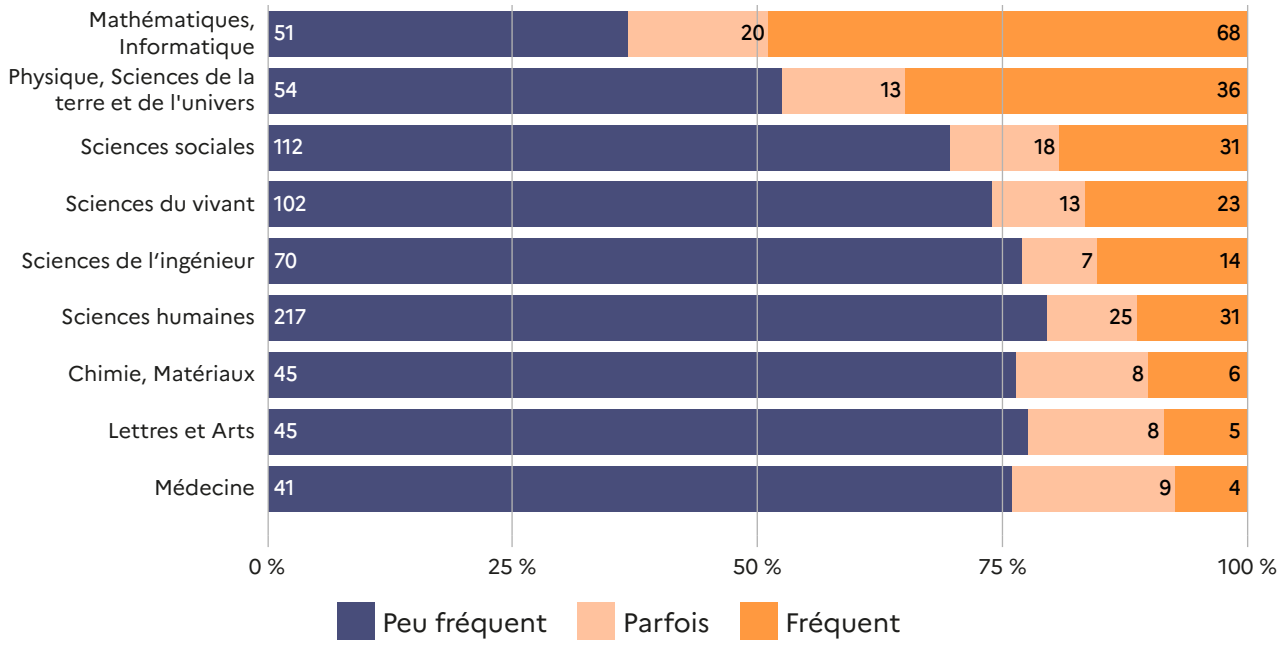


Figure 36 : Pratique de diffusion des prépublications sur des plateformes « *non for profit* » selon la discipline

Cette pratique est surtout le fait de chercheurs travaillant **dans un contexte collectif voire collaboratif**, ce qui met en relief l'importance d'être entouré dans son activité de recherche pour la diffusion de nouvelles pratiques. Les appartenances disciplinaires sont également assez déterminantes, ce qui reflète des pratiques antérieures à notre enquête et au développement de politiques en matière de science ouverte (Pontille, D. & Torny, D., 2013). Ainsi, diffuser des *preprints* est une pratique quasi majoritaire en mathématiques et informatiques, bien répandue en physique et sciences de la terre et de l'univers.

Elle l'est aussi en sciences sociales, notamment avec le serveur RePec dans le domaine des sciences économiques.

RECOMMANDATION : Le « plan S » rend obligatoire la publication en libre accès, dans une revue ou sur un serveur, à partir de 2021 pour tout cofinancement par une agence. Cette pratique de publication devra être étudiée selon des distinctions disciplinaires, des contextes de travail et le développement de plateforme comme *Peer Community In*.

Diffuser les données en lien avec les résultats de ses recherches

QUESTION 28: Diffusez-vous des données produites lors de vos recherches ?

La diffusion des données de la recherche en lien avec la publication ou la diffusion des résultats est une pratique très limitée. Au mieux, un chercheur sur dix le fait fréquemment ou parfois.

La diffusion des données de la recherche en lien avec une publication

Les faibles usages fréquents de diffusion des données de la recherche en lien avec une publication rend difficile l'analyse. Dans ce domaine, l'usage de plateformes proposées par les éditeurs, sous la forme compléments à l'article (*supplementary materials*) est la méthode la plus fréquente (11,3 % et autant pour parfois 14,1 %). Les *data paper* (2,80 % d'usage fréquent, 6,5 % de parfois), les dépôts dans des entrepôts de données généralistes (3,4 % d'usage fréquent, 5,1 % de parfois) ou disciplinaires (5 % d'usage fréquent, 4,8 % de parfois) ainsi que la diffusion de résultats négatifs (moins de 1 % d'usage fréquent et parfois) demeurent des usages marginaux de publication ou valorisation des résultats de recherche.

Distinctions disciplinaires

Les plateformes d'éditeurs qui proposent de placer les données en complément de l'article sont une pratique peu fréquente (25,4 % fréquent ou parfois), mais liée à la discipline du répondant. Cette pratique commence à être intégrée en chimie et sciences des matériaux, dans les sciences du vivant et dans une moindre mesure la médecine et la physique. À l'inverse, les lettres et SHS sont encore peu sensibles à ce type de dépôt de données.

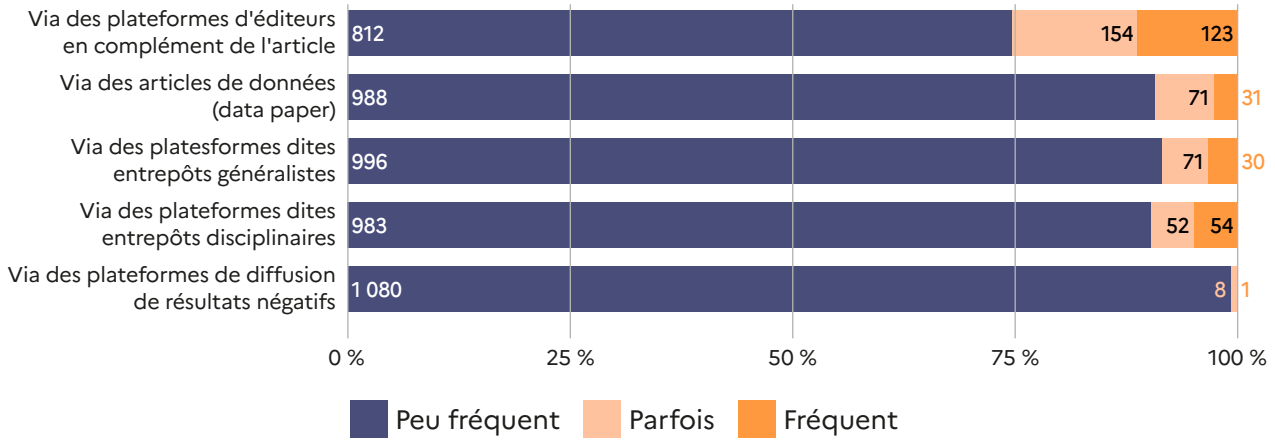


Figure 37: Usages des canaux de diffusion des données de recherches produites

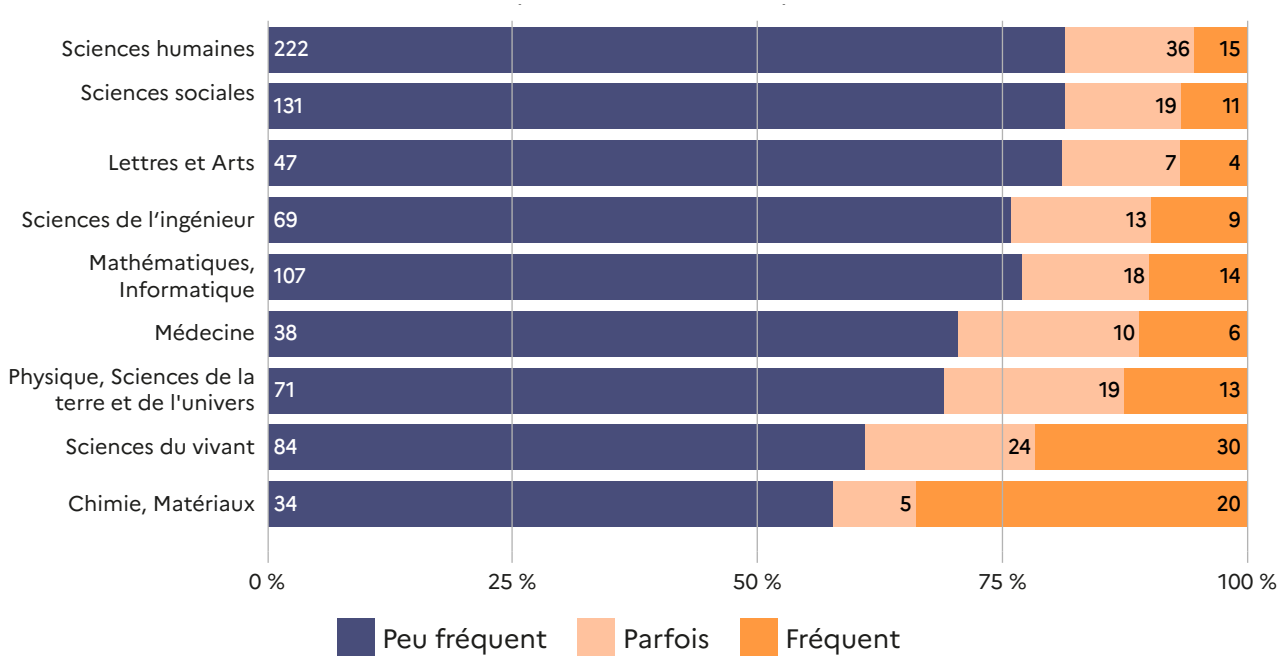


Figure 38: Diffusion des données sur des plateformes d'éditeurs en complément de l'article selon la discipline

La diffusion des données de la recherche en lien avec d'autres productions de la recherche

QUESTION 29 : Diffusez-vous d'autres types de productions scientifiques ?

Les données incluses dans d'autres types de productions de la recherche, comme les projets de recherches, les *notebooks*, les *grant proposals*, les rapports de recherche ou encore les posters, sont d'une diffusion très variable. Le code lié aux recherches et les rapports de recherche sont diffusés par environ ⅓ des répondants et près de la moitié en font tout autant pour leurs posters. Les autres productions semblent bien plus délaissées, notamment tous les éléments qui ont trait au déroulement du processus de recherche (*notebooks*, cahiers de recherche, projets de recherche, etc.). Cette faiblesse des résultats conduit à s'interroger sur la considération des répondants envers ces éléments de production issus des recherches. L'absence de diffusion sous forme d'article dans des revues à comité de lecture peut donc nuire à ce type diffusion. Comme le soulignent Gruson-Daniel et al. (2021) au sujet des codes sources, la problématique est double, d'une part l'évaluation et la carrière des chercheurs qui publient ces codes sources sans reconnaissance ou selon une obligation administrative et d'autre part, la reproductibilité des résultats de recherche.

Il semble donc qu'en la matière, l'appartenance disciplinaire, de même que l'âge du répondant, n'explique que partiellement les usages. L'appartenance disciplinaire fonctionne pour la diffusion des codes chez les mathématiciens, les informaticiens et les physiciens, mais elle reste peu opérante pour les autres formes. La diffusion de documents issus de la recherche autres que ceux qui sont de l'ordre des publications est non seulement assez peu effective, mais il n'existe pas de variable assez déterminante, sans doute en raison de la faiblesse relative de cet usage.

	N oui	% cit. Oui	N non	% cit. Non
Les posters	534	49,00 %	555	51,00 %
Les rapports de recherche	391	35,90 %	698	64,10 %
Les codes (GitHub, Git, Software Heritage…)	347	31,90 %	742	68,10 %
Les projets de recherche (OSF…)	171	15,70 %	918	84,30 %
Les prototypes	133	12,20 %	956	87,80 %
Les notebooks ou cahiers de recherche (OpenNotebookscience…)	98	9,00 %	991	91,00 %
Les protocoles et workflow (Protocols.io, MyExperiment, Scientific protocols, DMPOpidor…)	90	8,30 %	999	91,70 %
Les *grant proposals* (RIO…)	76	7,00 %	1013	93,00 %
Du matériel d'accompagnement d'articles scientifiques (Kudos, JOVE…)	68	6,20 %	1021	93,80 %

TABLEAU 36 : Répartition de la diffusion des autres types de production de la recherche

Les rapports de recherche et la discipline

Nous avons réalisé une analyse croisée à propos des rapports de recherche, document souvent plus dense qu'un article et qui comporte souvent d'autres données qui peuvent ne pas figurer dans d'autres publications plus condensées. On notera que ce type de diffusion semble être surtout celle de génération plutôt ancienne et moins marquée chez les moins de 30 ans. De plus il apparaît que ce mode de diffusion est délaissé dans certaines disciplines, pas uniquement les STEM puisqu'elle est sous-représentée chez les répondants des lettres et arts et surreprésentée en mathématiques et informatiques.

RECOMMANDATION : Maintenir cette question mais lier ce questionnement dans le questionnaire aux questions d'évaluation, de carrière et de reproductibilité.

Une valorisation des résultats des recherches encore balbutiante

QUESTION 31 : Comment valorisez-vous vos résultats de recherche ?

Une fois la publication réalisée, se pose la question des actions des répondants pour la valoriser si nous reprenons les impératifs tels que « *promote or perish* ». Nous avons proposé aux chercheurs de nous indiquer quels sont les moyens qu'ils utilisent pour valoriser leurs propres recherches sur les réseaux sociaux numériques spécialisés ou généralistes, dans les médias en lignes collaboratifs, lors de conférences, etc.

Les réponses permettent de distinguer 4 ensembles de formats de valorisation d'usage très différents :

1. En premier lieu, les conférences demeurent une pratique assez bien répandue (près

de 55 %). Cette dernière demeure le mode de valorisation le plus pratiqué (en dehors des publications académiques);

2. Ensuite, trois pratiques se distinguent par une fréquence moyenne sans être rares, autour de 33 %: la valorisation dans les réseaux sociaux numériques académiques et professionnels (ResarchGate, Academia, LinkedIn, etc.), la présence numérique et le référencement sur les moteurs de recherche;
3. Viennent ensuite les réseaux sociaux numériques généralistes à hauteur de 12,5 % de fréquence (« toujours » et « souvent »);
4. Enfin, nous constatons que les autres formes de valorisation sont marginales voire rares. Ces dernières touchent surtout à des formes de valorisations qui s'adressent à des espaces sociaux externes aux mondes académiques : médias en ligne (TheConversation, AOC…), outils multimédia (dont podcast, Youtube, CanalU…), sciences citoyennes ou participatives (Zooniverse…).

Il semble que la valorisation a une frontière, celle de la sphère académique. La valorisation donne lieu à des pratiques, du moment où elle est destinée à l'espace académique, au sens large. Dès que cette valorisation rime avec une diffusion des résultats de la recherche à des personnes ou des institutions qui ne sont pas en lien avec le monde académique, elle est nettement moins fréquente.

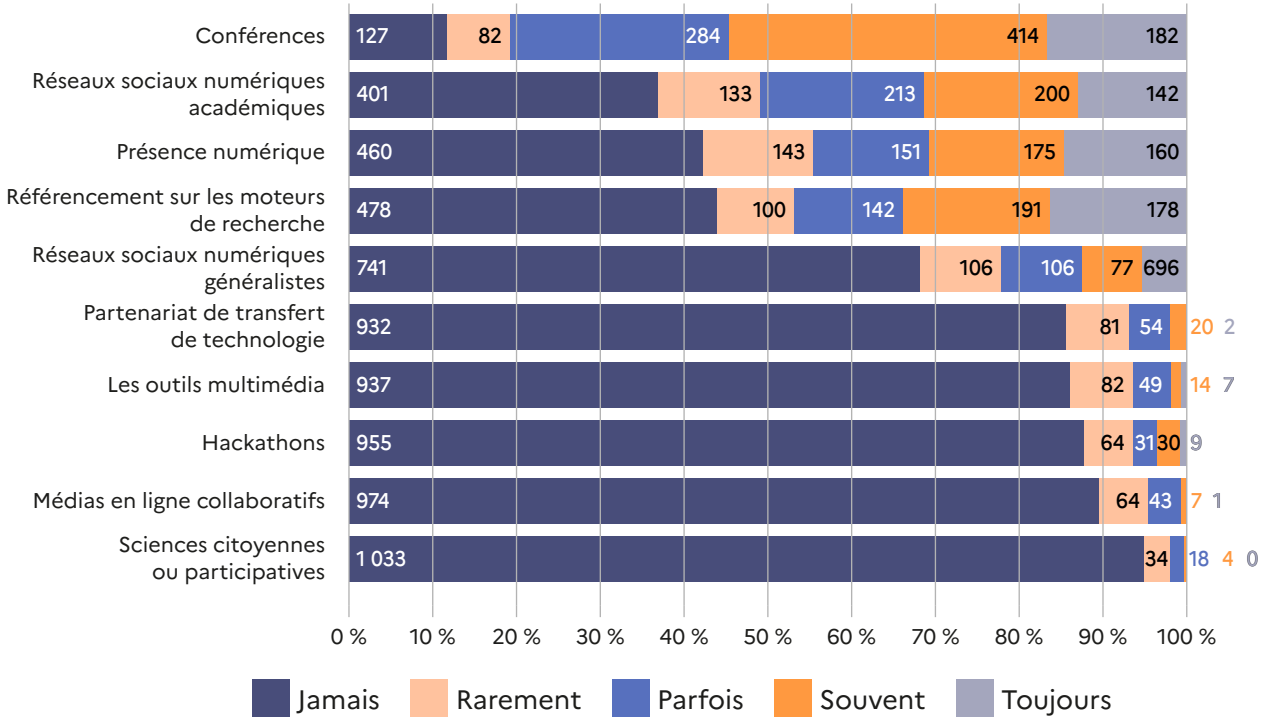


Figure 39: Formats de valorisation des résultats de la recherche utilisés

Précédemment une question sur la production des données élargissait également vers la société civile. La **QUESTION 16** [Avez-vous recours au crowdsourcing (Amazon Mechanical Turk, OpenStreetMap, Vigie-Nature…) ?] a reçu 7,7 % de réponses positives. Ce taux de réponse est faible sans être non négligeable. Il serait donc

intéressant de lier les deux pendants, entre la production des données et la valorisation des résultats de recherche dans une perceptive d'ouverture à la société civile ou de formats moins reconnus académiquement. L'appui sur la société civile pour la production des données pourrait voir augmenter l'usage des formats de sciences participatives ou citoyennes.

Il semble intéressant de prolonger l'étude sur ce point d'autant que le Plan National pour la Science Ouverte 2 (2021) inscrit la recherche participative dans ses actions.

RECOMMANDATION : les résultats peuvent être biaisés. Nous avons des répondants qui ont eu de nombreuses questions sur les données de recherche. Il est possible que ces pratiques, notamment dites de vulgarisation scientifique soient plus valorisées par des politiques d'établissement ou par des chercheurs n'ayant pas participé à notre enquête.

Les formes d'évaluation des publications émergentes peu connues

QUESTION 32 : Connaissez-vous des plateformes d'évaluation ouverte par les pairs ?

Dans le débat académique, en lien avec les questions autour de la Science Ouverte, de nouvelles formes d'évaluation des articles (ou de *peer-review*) émergent, comme l'initiative Peer Community In, une plateforme d'évaluation ouverte par les pairs, développée en 2016 par trois chercheurs de l'INRAE.

Si une majorité des répondants ne les connaît pas du tout (57 %), 8,2 % les connaît parfaitement et 34,8 % en a déjà entendu parler. Étant donné la récente arrivée de ces modes d'évaluation, un taux de réponse de connaissance à 43 % n'est pas négligeable ; il souligne un intérêt des répondants. Mais nous ne quantifions pas ici une pratique.

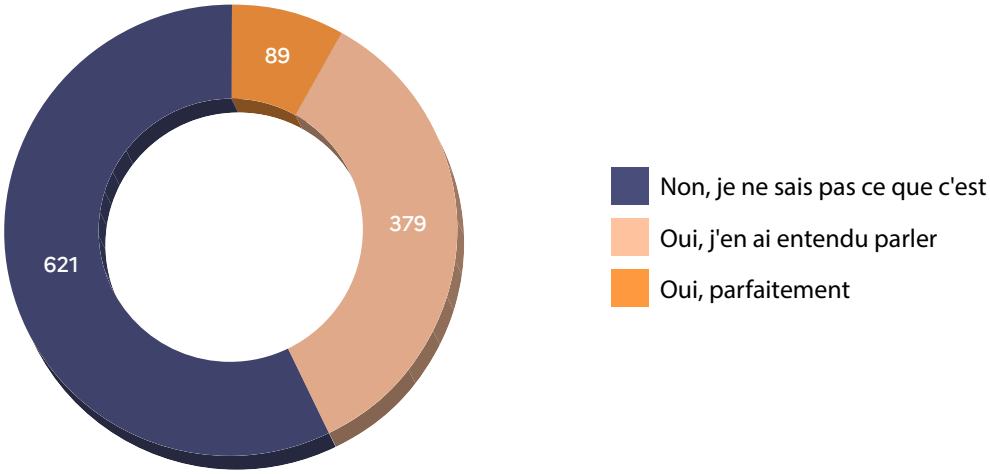


Figure 40 : Connaissance des plateformes d'évaluation ouverte par les pairs ?

Lorsque le répondant a déclaré connaître ce type de support, nous lui avons proposé de nous signaler quel était ce type de support **(QUESTION 33)**. Les plateformes outil ou communautaires majoritairement connues sont Publons, Pubpeer. Le chercheur y dépose ses prépublications, ses publications, sa biographie et effectue des commentaires qualitatifs sur les textes post-publications de ses pairs. Bien entendu il aurait fallu distinguer plus en détail ces outils, mais cela indique globalement que la communauté scientifique s'intéresse, voire semble prête à se saisir de plateformes d'*open peer review* en dehors des procédures classiques des revues traditionnelles.

De même le nombre de répondants qui citent les *registred reports* n'est pas anodin, il va dans le même sens : une évaluation qualitative du contenu avec le moins de biais possibles dans la publication des résultats fructueux ou infructueux.

Ainsi loin des réseaux sociaux numériques académiques et des indicateurs de la publication commerciale (facteur d'impact, comités restreints, absence de transparence dans l'évaluation…), la moitié des chercheurs interrogés connaissent l'*open peer review*. Y sont-ils favorables ? Sont-ils prêts à changer de modèle ? Ces questionnements dépassent notre enquête.

Si oui lesquels ? (Taux de réponse 40,7 %)	Nombre	% Obs
Outils de *peer review* (PubPeer, ScienceOpen, Publons, Peer Community In, Self Journal of Science…)	310	70,00 %
Journaux *open peer review*/évaluation ouverte (pré-publication) (F1000, PeerJ…)	189	42,70 %
Pre-registering (PLOS, OSF, AsPredicted…)	182	41,10 %
Journaux *open peer review*/évaluation ouverte (post-publication) (Winnower…)	47	10,60 %

TABLEAU 37 : Connaissance des plateformes d'évaluation ouverte par les pairs

RECOMMANDATION : lier la connaissance à la pratique dans de prochaines enquêtes, même si la distinction disciplinaire pourrait venir pallier les résultats.

5 DIFFUSION DES RÉSULTATS DES RECHERCHES, DES DONNÉES ET PLACE DU CHERCHEUR DANS LA SOCIÉTÉ

QUESTION 35: le partage et la diffusion des données vous paraissent-ils compatibles avec les activités de valorisation ?

La question placée en fin de questionnaire, porte sur la compatibilité entre le partage et la diffusion des données et les activités de valorisation. Elle permet indirectement de fournir des réponses sur le débat autour de la science ouverte et les obstacles ou réussites rencontrés pour la diffusion des données de la recherche que nous avons formalisées à travers une matrice de SWOT adaptée en 4 items : forces/faiblesses/opportunités/risques. Cette matrice de SWOT a pour objectif de clarifier les enjeux stratégiques qui s'affichent dans les réponses. Elle est ici utilisée comme une grille de lecture.

- **Méthode adoptée :** Nous avons opéré un traitement manuel des réponses, puis procédé à une analyse textuelle via le logiciel Iramuteq qui a permis de valider les thématiques identifiées manuellement.

Analyse textuelle avec Iramuteq

Le nombre de réponses sur cette question s'élève à 970, pourtant, seul 338 feront l'objet d'une analyse textuelle avec Iramuteq[1]. Les 632 réponses manquantes ont une moyenne de 2,13 occurrences (ou mots) qui ne se prête pas à ce type d'analyse. La majorité des répondants se sont contentés de réponses courtes ou

1. L'analyse textuelle via Iramuteq (http://www.iramuteq.org/) a été réalisée par Lucie Loubère, elle est venue compléter nos analyses qualitatives et réalisée précédemment avec Sphinx.

très courtes (oui/non ou des formes équivalentes). La classification montre une classe de discours anglophone, composée de 28 réponses, elles tendent vers une forte adhésion à la compatibilité (24 sur les 28).

L'analyse Reinert réalisée via Iramuteq avec 6 classes vient renforcer les quatre lignes de tensions qui émergent de l'étude qualitative des réponses ouvertes :

- Un questionnement éthique et de valeurs (forces);
- Un questionnement pratique et technique (faiblesses);
- Un questionnement légal et réglementaire (risques);
- Un questionnement épistémologique et méthodologique (opportunités).

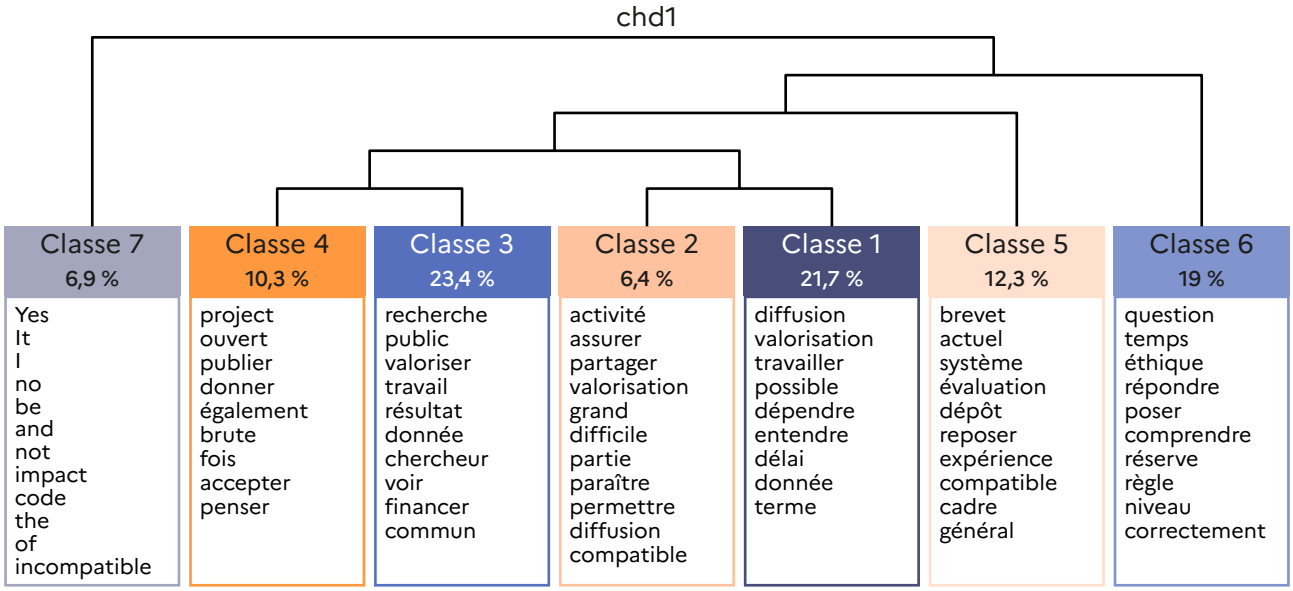


Figure 41: Dendrogramme des 338 réponses à la question de compatibilité (Iramuteq, Lucie Loubère)

La date de la première publication (catégorisée) semble avoir un effet sur la question concernant l'évolution du partage des données. Les « jeunes chercheurs » sont surreprésentés dans les classes de type témoignage et les « anciens chercheurs » sur les classes de type pratique.

Les jeunes chercheurs sont également surreprésentés dans le discours cernant la recherche publique comme bien commun.

Selon la variable statut, les doctorants (avec ou sans contrat) sont surreprésentés dans le discours sur l'évolution possible de ses propres pratiques.

Les fonctionnaires ou indépendants sont par contre surreprésentés dans la classe 1 de la question sur la conciliation où est discutée la valorisation économique des données et le besoin d'équilibrer cela avec leur diffusion.

Matrice de SWOT

Ces questionnements ont été placés dans une matrice de SWOT adaptée en forces/faiblesses/opportunités/risques comprenant 2 axes avec chacun deux pôles. D'une part, il y a bien évidemment la valence de la polarité, autrement dit le fait que le répondant conçoit les arguments qu'ils avancent comme des éléments positifs ou à l'inverse négatifs, des éléments d'une dynamique de progression ou bien ceux qui apparaissent comme des obstacles. D'autre part, il existe un deuxième axe concernant la perspective adoptée par le répondant, à savoir une perspective interne traduite dans le graphique soit avec la position individuelle du répondant (interne), ou bien à l'inverse la connotation collective où l'individu s'insère dans un tout (externe), un collectif auquel il adhère et dont il estime être membre.

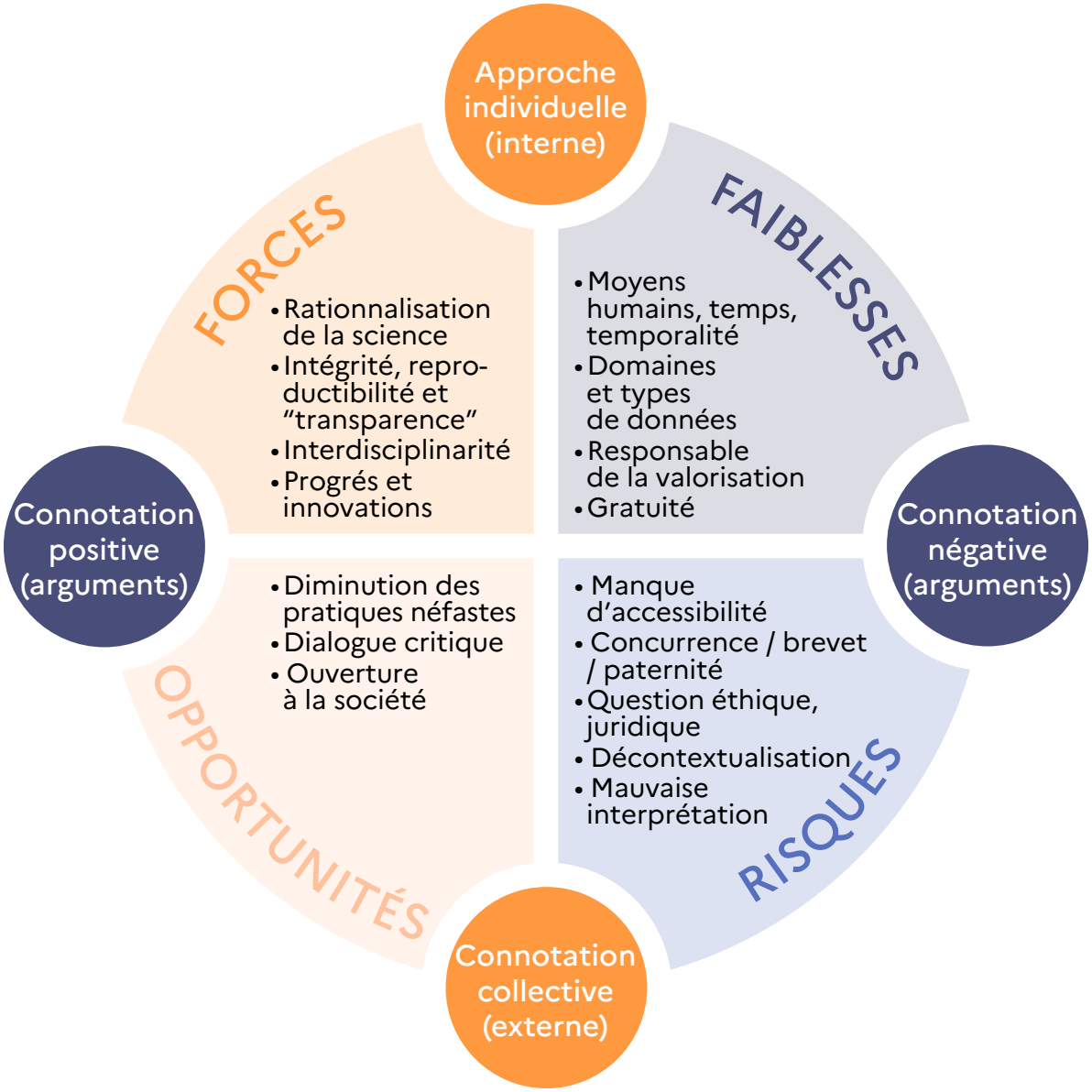


Figure 42: Matrice d'interprétation des questionnements autour de la compatibilité diffusion-valorisation des données de recherche (excalidraw)

Forces : les apports positifs du partage des données pour sa propre activité

Cet ensemble synthétise les réponses des chercheurs qui positionnent le débat autour de considérations personnelles tout en percevant positivement le lien entre partage et valorisation.

La pratique de diffusion des données et des résultats, tout comme leur valorisation, sont perçues comme compatibles : avancées de la science, progrès, capacité à innover. La pratique scientifique devient plus ouverte, d'où la convocation de l'interdisciplinarité :

> « Ce ne sont pas des activités de valorisation. C'est une contribution à l'avancée de la science, quelles qu'elles soient. La fonction du partage assure d'un déploiement de l'interdisciplinarité indispensable à une diffusion des savoirs et des connaissances. »

La reproductibilité des résultats est possible une fois que la diffusion des données a été réalisée par les chercheurs, pour qu'elles puissent être exploitées par d'autres chercheurs, voire des entreprises privées. La commercialisation des données et le dépôt de brevet rencontrent également la notion de bien commun (classe 3 du dendrogramme).

> « C'est une bonne façon d'assurer la reproductibilité des résultats pour des travaux dont la complexité va croissante, et qui peuvent de moins en moins être décrits de façon exhaustive dans un article de 8 pages. »

> « La valorisation de la recherche est une présentation des résultats, de la démarche, des idées, bref une manière d'interpréter des données de recherches. Celles-ci, (en sciences sociales du moins) sont des résultats en soi et le fait de les rendre accessible est aussi **un bien commun à la société**. Les deux sont nécessaires afin que quiconque souhaite réaliser un autre travail de recherche critique à partir des données déjà produites. Les deux permettent de produire des connaissances. »

> « Oui, après publication scientifique, et si les données ne sont pas susceptibles d'être indispensables à une forme **de brevetabilité**. Il faut réfléchir à l'amont si les données sont susceptibles d'être valorisées dans le cadre de développement d'un produit commercial, et choisir alors des partenaires pouvant financer la recherche. Si non, mieux vaut mettre les données en libre accès, au cas où cela inspirerait quelqu'un pour un autre travail scientifique ou commercial (autre valorisation), plutôt que de les perdre en les laissant inconnues. »

Faiblesses : les obstacles de la diffusion des données pour sa propre activité

Les chercheurs ne sont pas opposés au lien entre partage des données et valorisation, ni aux enjeux de la science ouverte. Ils interrogent la dimension pratique et matérielle de leur mise en œuvre et soulignent les obstacles auxquels ils sont confrontés. Les réponses nombreuses montrent que des éléments matériels peuvent grever l'adhésion de principe à la science ouverte.

Le travail de valorisation est défini comme un travail de vulgarisation de la science déléguée à des tiers compétents :

> « Dans l'absolu oui. Mais la vulgarisation ne s'improvise pas. C'est un métier que personnellement je ne sais pas faire. On l'a vu avec le COVID, il y a beaucoup d'a priori négatifs sur la science en France, et cela est dû à un échec de la façon dont la science est enseignée. On ne peut pas sereinement diffuser l'info au grand public comme cela. Il faut des vrais vulgarisateurs. Valoriser passe par expliquer comme il faut. »

Au-delà de la compétence, le temps pour réaliser ces tâches est convoqué.

> « Oui, mais cela peut être très chronophage si l'on veut que la diffusion des données soit effectué correctement. Les données doivent être bien documentées avec des métadonnées et expliquées. La question du temps pour organiser les données mais aussi savoir où publier les données, connaître les normes des métadonnées (parfois très voire trop complexe), les entrepôts, la rédaction éventuelle d'un data paper est à prendre en compte. Les projets de recherche sont rarement dimensionnés pour permettre de prendre ce temps ou avoir des ressources pour permettre de faire ce travail comme on le souhaiterait. »

Les questions en suspens ont également trait aux types de données concernées. Quelles sont les données partageables et selon quelles rationalités ? Les répondants suggèrent de différencier les types de données comme dans l'enquête menée à l'Université de Rennes 2 entre données sources et données résultats :

> « On fait la part des données brutes (sorties de l'instrument), des données validées (calibrées, etc.), des données travaillées. »

Dans le même temps, les domaines de recherche (ou les disciplines) connaissent des pratiques différentes en termes de diffusion des données. Dans la réponse ci-dessous, les pratiques de la science ouverte sont développées sans que l'enjeu soit celui des données de recherche :

> « *In my field of theoretical physics, we most generate equations (not data or codes). Those equations are published in our scientific publications, which we systematically put on arxiv (open access). So «data» sharing/dissemination works via the traditional route of scientific publications. If I write a significantly useful code, I would put it on github.* »

Le principe de diffusion gratuite rencontre des obstacles qui peuvent décourager, car il peut devenir une fragilité pour le chercheur qui diffuse largement et gratuitement ses données de recherche :

> « [...]Le travail de collecte est chronophage, complexe et impliquant, coûteux en temps et en argent, demander à un chercheur de partager gratuitement ses données est abusif car cela laisse la porte ouverte à d'autres pour publier sans fournir d'efforts. Sans oublier que certaines personnes peu scrupuleuses publient des résultats de recherche (dont les résultats sont présentés en conférence) dans des ouvrages notamment avant même qu'un article ait été publié par le chercheur à l'origine de la recherche. De même, présenter des résultats de recherche avec le détail de la base de données ouvre la porte au « vol » de sujet par des chercheurs qui ont des moyens importants pour collecter des données très rapidement (chaires, budget de recherche colossal, armées de doctorants) et de « griller » la priorité au chercheur à l'origine du travail préparatoire. »

Plusieurs obstacles émergent : les moyens humains et en temps disponible, la rationalisation des données à diffuser. Un certain nombre de chercheurs a le sentiment que les réponses leur incombent directement voire individuellement, alors qu'ils sont généralement attachés à une représentation de la recherche comme une activité collective. Des tensions, voire des oppositions peuvent alors de mettre en place.

Risques : les éléments négatifs de la diffusion et de la valorisation des données

La dimension prise en compte est l'environnement professionnel, ou plus largement la science.

En premier lieu, vient le thème de la concurrence et de la question de la paternité des résultats. D'une part il se peut qu'un partenariat intervienne et qu'il restreigne

considérablement les possibilités de partage et même oriente la valorisation elle-même à se limiter au dépôt d'un brevet :

> « Oui on peut valoriser puis partager. Il faut savoir aussi ce qu'on entend par partage des données. Si les données sont produites en partenariat avec un industriel ou si un industriel « achète » les données, la valorisation se fera par le dépôt d'un brevet et les données ne pourront pas être diffusées ou partagées dans leur intégralité. »

D'autre part, à côté de la concurrence avec le monde de la recherche privée, partager des données peut signifier que les chercheurs se mettent en difficulté, ainsi que leur équipe, leur institution :

> « […] Par ailleurs, l'évaluation se faisant sur la publication, partager des données au risque que des concurrents publient des résultats avant vous n'a pas de sens, surtout si les données ont été produites avec les moyens d'un laboratoire ou d'une institution : la priorité de publication doit appartenir à ceux qui ont fourni les moyens et le travail : quelle motivation sinon pour le chercheur ? Produire des données pour les autres ? Devenir la cheville ouvrière de la recherche étrangère ? »

Outre la concurrence, la dimension scientifique émerge au travers de réponses qui évoquent les possibles mésusages, mésinterprétation des données diffusées. Les risques d'interprétations erronées, partielles voire fallacieuses sont craints. De plus, la focalisation sur les résultats noie en quelque sorte tout le processus de développement de la recherche, parfois long, qui serait réduit alors à la portion congrue, voire rendu invisible du travail des données (Denis, 2018).

> « Mettre en commun des données auprès de non scientifiques ne peut que conduire à de mauvaises interprétations, c'est parfois déjà le cas entre scientifiques. Une donnée n'a de sens que si on connait parfaitement son origine, comment elle a été obtenue et dans quelles conditions. Déjà dans de nombreux articles les données apportées posent problèmes au niveau de l'interprétation alors ouvert à tous… en ces temps de communication extrême où tout doit faire Buzz et gazouillis je crains le pire rien de plus facile avec un jeu de donnée de lui faire dire tout et son contraire (ex : facile d'établir une corrélation positive entre le port du short et un non état grippal (hors covid). Alors valorisation oui, mais avec précaution. »

La classe 6 du dendogramme (Figure 38) est composée de plusieurs types de discours dont les difficultés des répondants à se positionner sur ce débat. Souvent peu concernés par les données, ils n'ont pas de visibilité dessus, certains regrettent **que la dimension éthique n'ait pas été abordée dans le** questionnaire.

> « Bref j'ai du mal à répondre à votre question 34 j'ai mis oui partout par principe, mais il manque la dimension éthique par exemple puis je rende universellement accessible les infos publiques trouvées en archives sur telle personne publique »

Le cadrage juridique n'est sans doute pas assez connu, d'autant que pendant de nombreuses décennies les aspects juridiques en matière de diffusion de résultats pesaient faiblement sur l'activité des chercheurs. Avec l'introduction de nouveaux règlements sur les données relayées par les grandes institutions de recherches (université et organismes publics de recherche), les chercheurs y sont plus régulièrement confrontés. Les questions éthiques relèvent du même raisonnement (en matière de traitement des données à caractère personnel, notamment). Émergentes, ces questions prennent appui sur un débat où l'éthique peut relever de la pratique du chercheur ou de la conception même de la démarche de recherche, qui devrait suivre certaines postures morales. Ces deux conceptions se retrouvent dans la réponse reprise ci-dessous :

> « Oui à condition de respecter certaines règles d'éthique (anonymat des personnes enquêtées, etc.). »

La valorisation par la diffusion des données de la recherche pose des questions éthiques aux chercheurs sans que les éléments du débat ne soient identiques pour l'ensemble des chercheurs au-delà des distinctions disciplinaires. La question est de savoir à quel niveau se situe la réalisation d'une démarche éthique : au niveau du chercheur ou au niveau institutionnel (Le Béchec, 2021).

En fin de compte, une diffusion plus large des pratiques doit dépasser deux types d'obstacles structurels et individuels.

Les obstacles structurels rendent compte des tensions qui sont défavorables à un investissement plus important de nombre de chercheur dans la science ouverte, comme la détermination des types de données partageables, la tension entre la diffusion des données et la mise en concurrence des chercheurs et des institutions de recherche entre elles, l'absence de moyens (humains et en temps) dédiés à ces activités. Les obstacles individuels, se lient aux premiers tout en ouvrant la porte à leur dépassement. À titre d'exemple, le manque de compétences sur les techniques de diffusion conduit des chercheurs à ne pas s'investir par manque de savoir-faire plus que par opposition de principe.

Les opportunités : les apports positifs de la diffusion et de la valorisation des données de recherche

Les positions incluent le collectif avant d'être marquées (et perçues) dans la pratique individuelle.

Trois questionnements ressortent particulièrement. Le premier concerne l'opportunité qu'offre le partage des données à une large échelle pour discuter un fonctionnement de la valorisation des recherches jugée négativement, parce qu'il mènerait à l'atomisation de l'activité de recherche, la personnalisation à outrance des résultats de la recherche, et à la privatisation, par les grands éditeurs, de résultats de recherches produits sur des fonds publics :

> « La mainmise des éditeurs incontournables pour assurer une valorisation effective (évaluation des chercheurs, rayonnement international), soit en accès par abonnement/*pay per view*, soit en *open access* moyennant des frais de publication importants, freinent le partage des connaissances, notamment pour la recherche publique qui devrait être accessible à tous car principalement financée par des fonds publics. »

Un autre répondant insiste, lui, sur le fait que le partage systématique permettrait d'assainir, de mettre en exergue des pratiques jugées « néfastes », voire de les réduire considérablement :

> « Oui dans la mesure où le cadre général de la valorisation a besoin d'être réformé. Les pratiques de "valorisation" incompatibles avec la diffusion publique des données sont pour beaucoup des pratiques néfastes (secret industriel, dépôt abusif de brevets, embargo sur des données pourtant produites sur des fonds publics comme des crédits impôt recherche…). »

Ce renversement s'accompagnerait, selon d'autres répondants, par la mise en œuvre d'un dialogue critique au cœur même de l'activité scientifique, activité qui pourtant lui est consubstantielle. Cependant, il semble que pour certains le partage des données renforce cette activité scientifique de controverse critique, ne serait-ce que parce qu'elle permet un accès aisé aux données qui ont permis de construire les résultats :

> « Oui. Je traite les données récoltées selon mon approche et mon point de vue. Une diffusion et un partage peuvent permettre à d'autres personnes d'analyser les données sous un angle différent et/ou complémentaire. »

Ainsi ces chercheurs sont-ils favorables au partage des données parce qu'il renforcerait la capacité à dialoguer entre chercheurs, sans doute également au-delà des frontières disciplinaires. Il s'agit là d'une manière de rappeler que le dialogue critique fait partie de la démarche scientifique.

Enfin, le troisième aspect de ce dialogue critique porte sur un thème assez proche du précédent, celui du lien entre la science et la société. Là encore, il s'agit de renforcer une mission du chercheur, du scientifique, en tant qu'entité collective.

> « Aujourd'hui, non, mais c'est indispensable à mon sens pour que la science ne s'isole pas de la vie civile et que cette dernière puisse s'approprier également la ressource. »

Ainsi l'enjeu en cause est alors le lien entre science et société. La valorisation des données de la recherche, la diffusion des résultats pourrait (devrait ?) être une part conséquente de ce lien, part qui dépasse l'activité individuelle du chercheur, car elle concerne toutes les recherches scientifiques. Ainsi, cette position appelle à faire de la science ouverte un appui de l'activité scientifique actuelle et future.

RECOMMANDATION : inclure des questions liées à l'éthique et à la législation (par exemple le RGPD) qui influent sur l'ouverture des données de recherche

Les évolutions des pratiques numériques

QUESTION 36: Selon vous, vos pratiques numériques vont-elles évoluer à l'avenir et de quelles manières ?

Cette dernière question avait pour but d'évaluer l'opinion de répondants à propos de l'évolution de leurs pratiques numériques. Cette question permettait au répondant de se positionner à la fois dans une perspective personnelle, et à la fois dans une perspective plus large et collective.

Par une étude qualitative, 5 grandes thématiques se distinguent :

- Le contexte (moyens, pression institutionnelle, évaluation, opportunités…);
- Les outils et les pratiques (publications, données de la recherche, plateformes, visioconférence…);
- Les acteurs (la place de la France dans le développement des outils numériques, les institutions, les chercheurs, les éditeurs et leur contournement…);
- Les valeurs (la transparence, le partage, l'éco-frugalité, les modèles économiques);
- Les moyens (numériques, formations…).

HYPOTHÈSE : il existe des différences de valeurs associées à la science ouverte avec une population sensible aux enjeux techniques et une population sensible aux enjeux humains.

Nous avons tenté de déterminer des indicateurs permettant de discerner, parmi la population qui a répondu, celles et ceux qui étaient le plus sensibles aux questions et débats portés autour de la

science ouverte, à partir de quelques indicateurs sélectionnés parmi les questions de l'enquête SOSP. Toutefois, il est possible pour l'exploration des résultats d'avoir deux formes d'adhésion non-exclusives.

Les adhérents à la science ouverte dans sa dimension humaine

La première est définie comme une adhésion à la dimension humaine de la science ouverte (SO), c'est-à-dire les chercheurs qui s'appuient sur les grands principes de la SO. Ainsi, ils satisfont aux critères suivants :

- **QUESTION N° 26** (Quels sont vos critères de choix pour soumettre une publication ou une communication ?) = diffusion en accès libre (choix personnel)

  Il s'agissait de sélectionner des répondants qui positionnent leur pratique envers la SO comme un élément faisant partie de leur philosophie de la recherche.

- **QUESTION N° 30** (Diffusez-vous vos publications scientifiques évaluées par les pairs dans les réseaux sociaux numériques académiques ?) = parfois/souvent/toujours

  Il s'agissait de sélectionner des répondants diffusant régulièrement des publications scientifiques *via* des outils numériques même privés qui dénotent un contournement (Certeau, 2010) des processus classiques de diffusion des articles disponibles dans le monde académique.

- **QUESTION N° 34** (Selon vous, est-il souhaitable que les données de recherche que vous avez produites ou contribué à produire soient partagées ? Modalité de réponse : avec tout citoyen, sans restriction professionnelle ni géographique) = oui

  Il s'agissait de sélectionner des répondants pour qui le partage des données est le plus large possible.

Le croisement de l'ensemble de ces 3 critères permet de distinguer 172 répondants qui ont répondu de la même manière à ces critères, soit 15,8 % des répondants. Cette sous-population est assez restreinte, mais elle se prête tout de même à la comparaison statistique, notamment par rapport à la population générale. On peut affirmer que nous avons un échantillon de répondants assez au fait de la science ouverte, qui ne se limite pas à des pratiques associées (comme c'est le cas pour la sous-population suivante) mais s'étend à une véritable philosophie. Ils peuvent être qualifiés de précurseurs sur les enjeux de la SO.

Le tableau ci-dessous résume les 7 caractéristiques des répondants appartenant à cette sous-population. Nous allons les détailler en insistant sur les éléments qui les distinguent.

Il n'existe pas de discrimination disciplinaire ni de concentration disciplinaire. La répartition est assez similaire à la population répondante, ainsi le facteur disciplinaire ne peut pas expliquer cette propension à être un précurseur de la SO. Les répondants sont plus présents dans les sciences humaines et les lettres, un peu moins dans la médecine ou la physique, mais aucun des 9 groupes disciplinaires n'est absent.

La fonction n'est pas discriminante. Les écarts sont assez faibles, (de 4 points pour ce qui est des chercheurs), et ne peuvent pas signifier une tendance profonde.

Les systèmes d'exploitation utilisés et le genre ne sont pas discriminants.

Variables	Caractéristique de la sous-population
Disciplines	Partagé dans toutes les disciplines
Statut	Assez comparable
Fonction	Assez similaire à la pop totale
Âge	Plus jeune que la pop totale
Sexe	Identique à la population totale
OS	Similaire à la population totale
Contexte de travail	Quasi absent des équipes > 10 personnes

TABLEAU 38 : Tableau des 7 caractéristiques des adhérents à la SO dans sa dimension humaine

Les éléments qui caractérisent cette population sont d'une part une quasi-absence parmi les répondants qui travaillent dans des grandes équipes (> à 10 personnes) et une moyenne d'âge plus jeune. Ce sont surtout les 30-45 ans qui sont surreprésentés, ce qui se traduit par une moyenne d'âge de quasiment 43 ans (44 ans pour la population totale) et surtout une médiane à 41 ans (44 ans pour l'ensemble de la population). En outre, la présence des plus jeunes chercheurs (les moins de 30 ans) n'est pas plus importante.

Finalement, cette population ne se distingue pas spécifiquement et surtout nettement par les caractéristiques sociodémographiques des répondants. Les éléments les plus déterminants pour figurer dans cette sous-population sont à rechercher probablement dans les parcours ou les trajectoires suivis par les chercheurs, mais ces informations ne sont pas disponibles dans l'enquête SOSP.

RECOMMANDATION : chercher à qualifier les parcours ou les trajectoires suivis par les chercheurs pour mieux comprendre le rôle par exemple de la formation de premier et second cycle.

En dehors de ces variables sociodémographiques, cette sous-population se singularise quelque peu à propos des types de données qu'elles travaillent, qui dérivent de la population générale, par une plus forte propension pour l'usage des données texte (67,4 % contre 51,9 % pour la POP qui les utilisent souvent ou toujours) et des données images (41,9 % contre 27,4 % pour la POP) légèrement au détriment des données numériques ou chiffrée (58,1 % contre 62,1 % pour la POP). Il apparaît que ces répondants utilisent un peu plus de données diverses et avec une fréquence plus importante que la POP.

Valorisation des échelons : de 1 (jamais) à 5 (toujours)

Dans le Tableau 39, la moyenne générale dans cette population est plus élevée (2,60 contre 2,45)[2], ce qui signifie que

2. La moyenne est calculée en attribuant un chiffre à chacun des niveaux de fréquence (de 1 pour jamais à 5 pour toujours). Ainsi, une moyenne se situant à 3 signifie que les répondants sur une fréquence assez forte, sachant qu'un résultat de 5 signifie qu'ils ont tous répondu « toujours », et un chiffre de 1 qu'ils sont unanimes à jamais le faire.

l'ensemble des réponses concernant l'utilisation des données de recherche signale une fréquence en moyenne plus importante. Dans le même temps, ces moyennes sont toujours, à l'exception remarquable des nombres et valeurs numériques, supérieures à la moyenne pour l'ensemble de la population, ce qui signale une plus grande propension de cette population à utiliser plusieurs types de données avec une plus grande fréquence.

Types de données	Moyenne	Référence : moyennes pour la « POP »
Nombres, valeurs numériques	3,49	3,57
Textes	3,72	3,51
Images	3,12	2,80
Vidéos	2,01	1,81
3D	1,56	1,44
Sons	1,70	1,55
Total	2,60	2,45

TABLEAU 39 : Moyennes pour les types de données travaillées

Les adhérents à la science ouverte dans sa dimension technique

La deuxième sous-population est assez proche de la précédente et correspond aux chercheurs qui s'appuient sur les outils numériques « estampillés SO ». Ainsi, ils satisfont aux critères suivants :

- **QUESTION N° 10** (Comment accédez-vous aux informations utiles à vos recherches ? modalité : via les archives ouvertes) = parfois/souvent/toujours

  Il s'agissait de sélectionner des répondants qui utilisent des outils numériques objectivement liés à une pratique valorisée dans la SO comme le fait d'accéder à des informations scientifiques via HAL à titre d'exemple.

- **QUESTION N° 18** (concernant les logiciels de production des données, Ces logiciels sont-ils ?) = libres et gratuits

  Il s'agissait de sélectionner des répondants qui non seulement connaissent la différence entre les types de logiciel et dont la pratique s'oriente résolument vers ce type de logiciel, sans présupposer que ce choix soit adossé à une adhésion à la SO.

- **QUESTION N° 29** (Diffusez-vous d'autres types de productions scientifiques ? modalité les codes) = oui

  Il s'agissait de sélectionner des répondants pour qui le partage des données fait partie d'une pratique courante.

Le croisement de l'ensemble de ces 3 critères permet de distinguer 231 personnes qui ont répondu de la même manière à ces critères, soit 21,2 % des répondants. Cette sous-population est un peu moins limitée que la précédente et se prête tout autant à la comparaison statistique, particulièrement

en rapport à la population générale. On peut affirmer que nous avons un échantillon de répondant assez au fait des outils de la science ouverte même s'il n'est pas possible de déterminer si c'est le fait d'une acculturation aux enjeux de la SO.

Variables	Caractéristique de la sous-population
Disciplines	Près de la moitié de math/info et physique, sciences de la terre et de l'univers
Statut	Très proche de la population totale
Fonction	Plus présent chez les IR/IE
Âge	Très proche de la population totale
Sexe	Plutôt des hommes
OS	⅔ de Linux
Contexte de travail	Travaille nettement plus souvent en équipe

TABLEAU 40 : Les 7 caractéristiques des adhérents à la SO dans sa dimension technique

	Nombre	% Obs.	Ech.
Sciences de l'ingénieur	22	9,70 %	8,4 %
Sciences humaines	34	15,00 %	25,1 %
Sciences sociales	17	7,50 %	14,8 %
Lettres et Arts	6	2,60 %	5,3 %
Médecine	6	2,60 %	5 %
Sciences du vivant	27	11,90 %	12,7 %
Chimie, Matériaux	7	3,10 %	5,4 %
Mathématiques, Informatique	71	31,30 %	12,8 %
Physique, Sciences de la terre et de l'univers	37	16,30 %	9,5 %
Total	**227**	**100,00 %**	**98,8 %**

TABLEAU 41 : Répartition disciplinaire des adhérents à la SO dans sa dimension technique

La morphologie de cette population, définie avec les mêmes critères sociodémographiques que la précédente, s'en distingue assez nettement.

Les éléments pour lesquels cette population est proche, dans sa répartition, sont le statut du répondant et la répartition en classe d'âge. Ainsi, contrairement à la sous-population précédente, l'âge ne semble pas influencer la propension à utiliser des outils de la SO.

La discrimination est liée au sexe du répondant, qui révèle une surreprésentation masculine. Mais ce constat s'explique peut-être par un autre élément plus déterminant : un plus grand déséquilibre dans la répartition

des disciplines. Les mathématiques, informatique, physique et sciences de la terre et de l'univers, sont des disciplines où les répondants sont majoritairement des hommes.

À l'inverse, les SHS et les lettres sont nettement sous-représentées, ce qui marque une différence entre le fait d'afficher des principes liés à la SO et d'avoir des pratiques qui s'y réfèrent.

Quant aux fonctions occupées dans la recherche, les distinctions avec la population générale sont assez fortes. Les répondants qui appartiennent au corps des personnels de soutien à la recherche (IE/IR) sont assez nettement surreprésentés (24,7 % contre 18,4 % pour la POP). Ce résultat s'explique sans doute par la propension de ses personnels à être mobilisés, voire à être considérés comme des référents sur les questions des outils numériques. Nous pouvons mentionner par exemple le réseau Mate-SHS[3]. Ce faisant, cette exigence les conduit sans doute à maîtriser un peu mieux que les autres fonctions les outils numériques. Dans cette dimension technique, l'association entre MCF et professeur est sous-représentée. Dans la population totale, les universitaires représentent 38,9 % des répondants, cette part chute à 33,3 % pour cette sous-population. L'écart signale sans doute une moindre présence des universitaires dans la pratique des outils numériques, constat déjà signalé plus haut.

Il est à souligner que l'usage de l'environnement Linux est très présent chez ces chercheurs, au point d'être le premier système d'exploitation signalé (⅔ de ces répondants l'utilisent). À l'inverse, l'environnement Windows est nettement moins utilisé et MacOS beaucoup plus utilisé en SHS et chez les MCF et PR.

Pour finir la description de cette sous-population, elle se caractérise par une nette sous-représentation de chercheurs travaillant seuls (9,1 % contre 23,8 % dans la POP). De plus, 66,2 % d'entre eux travaillent dans une petite équipe, une part supérieure de 9 points par rapport à la population totale (57 % pour la POP). Ainsi, le contexte de travail des chercheurs de cette sous-population est nettement plus collectif, au sein d'équipe d'équipe inférieure à 10 personnes.

Enfin, les répondants proches de la dimension techniques de la SO utilisent beaucoup plus de canaux pour la découverte des outils numériques que l'ensemble de la population. Ils ont donc une démarche plus active et se singularisent par la nette surreprésentation pour la réalisation d'une veille 35,5 % contre 21,4 % pour la POP.

Cette sous-population met en pratique des usages propres aux démarches de la SO, se distingue plus par des caractéristiques sociodémographiques spécifiques que la précédente.

Il s'agit plutôt de chercheurs masculins, plutôt affiliés aux STM, ayant une plus grande propension à travailler collectivement.

3. « Mate-shs est un **réseau métier** initié et porté par des ingénieur.e.s qui travaillent à la **production**, au **traitement**, à l'**analyse** et à la **représentation des données** dans la recherche en Sciences Humaines et Sociales (SHS). », voir https://mate-shs.cnrs.fr/

CONCLUSION

Les résultats de l'enquête SOSP_FR s'appuient sur 1089 répondants à un questionnaire dont la passation a été effectuée entre juin et septembre 2020. L'ensemble des matériaux de recherche sera mis à disposition dans le respect des règles d'anonymisation des données afin de permettre une exploration plus fine des résultats et d'envisager une reproduction de l'enquête. Des aménagements et des choix seront à opérer pour permettre une reconductibilité de l'enquête. Si l'analyse présentée dans le corps du rapport ne porte pas sur l'approche qualitative (entretiens), qui n'était initialement pas envisagée, elle nous paraît indispensable pour contextualiser des pratiques collectives et de nouvelles pistes de recherche sur les pratiques numériques de recherche, comme nous le suggérons en conclusion.

Un questionnement sur les pratiques individuelles des chercheurs

L'enquête SOSP_FR intervient dans une période charnière des pratiques et des cadres de la recherche où les outils numériques sont de plus en plus présents. Si les pratiques liées à l'accès ouvert font partie d'un cadre déjà bien établi (déclaration de Budapest sur l'*open access* en 2002, Plan national pour la science ouverte en 2018), la crise sanitaire de 2020 éclaire d'un jour nouveau les pratiques numériques des chercheurs en 2020. Aussi dans cette enquête, différentes questions renseignent notamment la numérisation des pratiques de recherche au travers de la conduite des recherches, l'environnement de recherche avec les infrastructures de recherche ou encore les nouvelles formes de valorisation des résultats de recherche.

L'enquête SOSP_FR tend à tester moins l'acculturation à des notions ou à une compréhension des enjeux de la science ouverte que d'analyser des pratiques numériques et des usages d'outils. Nous relevons des pratiques individuelles et collectives à l'aune de caractéristiques socio-démographiques. L'un des atouts est donc de souligner et de qualifier l'environnement de travail des chercheurs pour mieux cerner les pratiques numériques de recherche. Le travail solitaire et isolé ou réalisé dans des collectifs influence-t-il les pratiques numériques et la valorisation des résultats de recherche comme les données ? Les pratiques numériques diffèrent-elles selon les statuts et les fonctions dans la recherche ? Ce questionnement a été un pilier de l'enquête.

Les trois hypothèses initiales nous permettent d'aboutir à trois réponses sur :

- La relation entre le statut et la carrière du chercheur et son engagement dans des pratiques numériques liées à la science ouverte ;
- La relation entre l'environnement de recherche et son engagement dans des pratiques numériques liées à la science ouverte ;
- La relation entre les conditions de l'évaluation de la recherche et l'engagement dans des pratiques numériques liées à la science ouverte.

Des éclairages sur le cadre même de la recherche en France en 2020

L'âge ou la génération, la question du genre et de la discipline restent majoritairement peu discriminantes. Les résultats restent encore à discuter avec la communauté de l'enseignement supérieur et de la recherche.

Mais la question de l'accessibilité à l'information et aux infrastructures de recherche pour les jeunes chercheurs et selon leur lieu d'exercice nous semble importante à souligner. Si l'âge n'est pas discriminant dans les résultats, c'est sans doute les conditions de réalisation de l'activité pour les jeunes chercheurs qui les conduisent à développer moins de pratiques d'archivage et à être plus enclins à mobiliser des infrastructures non institutionnelles.

Les usages des outils pouvant résulter d'une question d'accessibilité, nous avons également interroger le sujet de la formation. L'information et la formation des personnels permanents reposent en premier lieu sur les amis et collègues, ce qui questionne le formalisme de formations peut-être éloignées du contexte de travail.

À propos des outils mobilisés par les communautés de recherche, l'enquête SOSP_Fr a permis d'obtenir la mention de 492 outils numériques différents dans les réponses des enquêté.e.s. Il est ainsi difficile d'établir des standards, des workflows stabilisés ou des usages récurrents mais nous relevons une numérisation des pratiques de recherche quelque soit la discipline. Ce corpus est une piste à travailler et à interroger. Avons-nous conscience de faire de la science ouverte lorsque nous utilisons un outil libre maintenu par une communauté de recherche ? L'enquête menée autour de l'ouverture des codes sources au sein de l'Enseignement Supérieur et de la Recherche (Gruson-Daniel, Jean, 2021) souligne la nécessité d'un accompagnement des communautés de pratiques, la nécessité d'une évaluation du développement de tels outils et leur implication dans les communs de la science, notamment en lien avec l'édition numérique. Selon nos travaux en cours et initiés dans ce rapport, des profils d'utilisateurs sont à affiner pour déterminer les indicateurs de réception à la SO avec une distinction dans sa dimension technique (usage de logiciel libre par exemple) ou dans sa dimension humaine (ouverture des résultats de la recherche à l'ensemble de la population).

Enfin, la question de l'évaluation, même si elle n'est pas directement interrogée dans le questionnaire, apparaît notamment dans un croisement de données entre les critères de choix de publication liée à l'évaluation et le degré d'ouverture des données accepté ou envisagé par les répondants. L'évaluation est apparue également dans la phase des questions ouvertes et dans la seconde partie de l'enquête qui n'est pas restituée ici, des entretiens semi-directifs menés avec 30 chercheurs et chercheuses. L'évaluation apparaît comme un des points nodaux pour l'appropriation de la Science Ouverte. Elle constitue la rémunération symbolique du chercheur qui lui permet d'obtenir l'avancement dans sa carrière, facilite l'obtention de financement pour son laboratoire et ses projets de recherche. Elle lui confère sa notoriété, sa place, son statut, son identité de chercheur. C'est du moins ce que perçoit le chercheur. Si collectivement des organismes (Université, Epst) signent la DORA, si le manifeste de Leiden est positivement perçu par les communautés scientifiques, si la CPU propose des recommandations d'évaluation plus qualitative, la suppression ou la restriction

d'une évaluation uniquement quantitative fondée sur les seuls Impact Factor ou H-index reste encore trop aisée, trop ancrée et d'apparence objective pour changer immédiatement les habitudes. Aussi, si l'évaluation des chercheurs prend en considération la qualité de ce qui est publié en *open access* ou en *open research data*, la bibliométrie réintégrera peut-être le domaine de la biblio-économie et quittera le pilotage de la recherche.

Ouverture : quel accompagnement pour la science ouverte ?

La multiplication de compétences numériques nécessaires et la diversification des outils rendent la science ouverte dépendante de formations, de communautés et d'infrastructures. À l'heure du numérique, les chercheurs ont-ils tous les moyens de bien faire les choses ?

La pratique dite de « science ouverte » ne signifie pas que les répondants s'inscrivent dans ce même mouvement. Les répondants n'ont pas nécessairement tous conscience que leurs pratiques s'inscrivent dans le champ de la science ouverte. Quelle part de pragmatisme ou de militantisme explique le recours à des outils libres ?

De même le logiciel libre s'accompagne parfois d'une gratuité qui peut-être un moteur plus large que celui du champ de la science ouverte dans l'ESR. À l'inverse, ne pas s'inscrire dans d'outils numériques maintenus et détenus par la communauté de l'ESR, enferme les chercheurs dans un environnement propriétaire dont ils n'ont parfois pas conscience sauf à vouloir utiliser un nouveau logiciel dépendant d'un système d'exploitation dont il ne dispose pas. La question de la « fracture numérique » ne peut donc être évacuée même si elle n'apparaît pas en tant que telle dans l'enquête SOSP_FR. L'interopérabilité peut résider dans des écosystèmes numériques fermés dont les *clouds* non-institutionnels rendent compte dans nos résultats. Il serait judicieux de comprendre comment des *workflows* de recherche peuvent découler d'entreprises tierces dont les GAFAM. La précarité peut-elle conduire à utiliser des solutions gratuites mais pérennes?.

Il nous semble également intéressant de mieux inclure dans l'analyse des pratiques numériques de recherche la diffusion des outils de recherche à travers les pratiques d'enseignement. Des logiciels comme GitLab, des langages de programmation et des logiciels libres comme R sont présents dans les formations universitaires. Les enseignants-chercheurs développent les formations sur des logiciels libres car les licences sont parfois inaccessibles dans le cadre d'enseignement. Ces outils libres vont devenir les environnements de recherche des futurs jeunes chercheurs. S'ils ne le sont pas encore, il est probablement qu'ils le seront très prochainement et influeront sur les pratiques de publication et de communication des résultats de recherche. Ces outils sont également mis à disposition des communautés scientifiques *via* des infrastructures. Il serait donc judicieux de mieux cartographier les infrastructures de recherche pour quantifier l'accessibilité et les besoins de formations pour les membres de l'ESR ne participant pas à des grands collectifs de recherche ou en début de carrière. Il existe en ce sens différents territoires de la science ouverte et des pratiques numériques qui s'y lient.

La première ressource informationnelle étant le réseau social de proximité, il nous paraît donc important de laisser le temps à l'information de circuler de pair à pair.

L'initiative des Ateliers de la Donnée[1] afin de se situer au plus près des communautés de recherche nous paraît donc pertinente. Il faut donc dépasser les principes d'acculturation aux notions pour rentrer dans les pratiques en accompagnant au plus près les communautés. Car l'effet d'entraînement des collectifs accompagne une meilleure diffusion des pratiques liées à la science ouverte même si les différences disciplinaires sont existantes.

Il reste à déterminer quelle part dans les pratiques numériques dépend de l'offre institutionnelle (outils, infrastructures, formations, communautés) et du libre choix des individus. Il serait utile de mieux qualifier comment les contraintes (qui peuvent s'appliquer à une appréhension de la Science Ouverte lorsqu'elle est perçue comme une injonction administrative sous couvert de transformation numérique et de transparence), interviennent dans le choix des outils et des pratiques. Les pratiques numériques de recherche demandent un investissement en temps pour les équipes et leurs membres qui recherchent de la pérennité tant d'un point de vue technique qu'humain. Entre injonctions institutionnelles, incitations académiques, recommandations par les pairs, disséminations réticulaires, adoptions généralisées, appropriations individuelles… les pratiques numériques répondent à des dynamiques collectives qu'il convient de sonder régulièrement.

1. Dans le cadre du développement de la plateforme de données, Research.data.gouv.fr, un Appel à manifestation d'intérêt « Ateliers de la donnée » a été initié en 2021. https://www.ouvrirlascience.fr/wp-content/uploads/2021/10/2021.10.11_AMI_Ateliers-de-la-donne%CC%81e.pdf

BIBLIOGRAPHIE

Akrich, M. (2013). Les utilisateurs, acteurs de l'innovation. In M. Callon & B. Latour (Éds.), Sociologie de la traduction : Textes fondateurs (p. 253-265). Presses des Mines. http://books.openedition.org/pressesmines/1200

Alloing, C., Pierre, J.-, & Casilli, A. A. P. (2017). Le Web affectif : Une économie numérique des émotions. INA éditions.

Alter, N. (2015). L'innovation ordinaire. Presses Universitaires de France.

Amiel, P., Frontini, F., Lacour, P.-Y., & Robin, A. (2020). Pratiques de gestion des données de la recherche : Une nécessaire acculturation des chercheurs aux enjeux de la science ouverte ? Résultats d'une enquête exploratoire dans le bassin montpelliérain (juin 2018). Cahiers Droit, Sciences & Technologies, 10, 147168. https://doi.org/10.4000/cdst.2061

Andrews Mancilla, H., Teperek, M., Van Dijck, J., Den Heijer, K., Eggermont, R., Plomp, E., Turkyilmaz-van der Velden, Y., & Kurapati, S. (2019). On a Quest for Cultural Change – Surveying Research Data Management Practices at Delft University of Technology. LIBER Quarterly, 29 (1), 1. https://doi.org/10.18352/lq.10287

Appel de Jussieu pour la Science ouverte et la bibliodiversité. (2017). https://jussieucall.org/

Baligand, M. P., Colcanap, G., Harnais, V., Rousseau-Hans, F., & Weil-Miko, C. (2021). Les pratiques de recherche documentaire des chercheurs français en 2020 (p. 57 p.) [Report]. Couperin.org. https://doi.org/10.5281/zenodo.4562180

Banat-Berger, F., Duplouy, L., Huc, C., & France. Direction des archives. (2009). L'archivage numérique à long terme : Les débuts de la maturité ? La Documentation française.

Bauer, B., Ferus, A., Gorraiz, J., Gumpenberger, C., Gründhammer, V., Maly, N., Mühlegger, J. M., Preza, J. L., Solís, B. S., Schmidt, N., & Steineder, C. (2015). Researchers and Their Data. Results of an Austrian Survey – [Report] 2015 (PDF full report EN). https://phaidra.univie.ac.at/o:409318

Baždarić, K., Vrkić, I., Arh, E., Mavrinac, M., Marković, M. G., Bilić-Zulle, L., Stojanovski, J., & Malički, M. (2021). Attitudes and practices of open data, preprinting, and peer-review – A cross sectional study on Croatian scientists. PLOS ONE, 16 (6), e0244529. https://doi.org/10.1371/journal.pone.0244529

Berghmans, S., Cousijn, H., Deakin, G., Meijer, I., Mulligan, A., Plume, A., de Rijcke, S., Rushforth, A., Tatum, C., van Leeuwen, T., & Waltman, L. (2017). Open Data: The researcher perspective – survey and case studies. 1. https://doi.org/10.17632/bwrnfb4bvh.1

Bonneville, A., Tucci, I., Vion, A., & Giglio, L. (2021). Données de la recherche : Pratiques et besoins dans un laboratoire pluridisciplinaire SHS (p. 55) [Research Report]. Laboratoire d'économie et sociologie du travail (LEST). https://hal.archives-ouvertes.fr/hal-03265603

Bordignon, F., Ermakova, L., & Noel, M. (2021). Preprint Abstracts in Times of Crisis : A Comparative Study with the Pre-pandemic Period. In I. Frommholz, P. Mayr, G. Cabanac, & S. Verberne (Éds.), BIR 2021 Bibliometric-enhanced Information Retrieval (Vol. 2847, p. 3744). https://hal-enpc.archives-ouvertes.fr/hal-03187900

Borgman, C. L. (2020). Qu'est-ce que le travail scientifique des données ? : Big data, little data, no data. C. Matoussowsky (Trad.), OpenEdition Press. http://books.openedition.org/oep/14692

Boudry, C., Alvarez-Muñoz, P., Arencibia-Jorge, R., Ayena, D., Brouwer, N. J., Chaudhuri, Z., Chawner, B., Epee, E., Erraïs, K., Fotouhi, A., Gharaibeh, A. M., Hassanein, D. H., Herwig-Carl, M. C., Howard, K., Kaimbo Wa Kaimbo, D., Laughrea, P.-A., Lopez, F. A., Machin-Mastromatteo, J. D., Malerbi, F. K., … Mouriaux, F. (2019). Worldwide inequality in access to full text scientific articles: The example of ophthalmology. PeerJ, 7, e7850. https://doi.org/10.7717/peerj.7850

Chao, M., Monini, C., Munck, S., Thomas, S., Rochot, J., & Van de Velde, C. (2015). Les expériences de la solitude en doctorat. Fondements et inégalités. Socio-logos. Revue de l'association française de sociologie, 10, Article 10. https://doi.org/10.4000/socio-logos.2929

Cosmo, R. D., & Zacchiroli, S. (2017). Software Heritage : Why and How to Preserve Software Source Code. 11.

Cousijn, H. (2017). Open Data: The researcher perspective – survey and case studies [Data set]. Mendeley. https://doi.org/10.17632/BWRNFB4BVH.1

Dillaerts, H., Paganelli, C., Verlaet, L., & Hugo, C. (s. d.). Usages et pratiques en lien avec les données de recherche. Une enquête menée auprès des chercheurs de l'université Paul-Valéry Montpellier 3. 93.

Donati, C. S. (2019). Données de la recherche : Quelles pratiques ? Quels besoins ? Enquête à Aix-Marseille Université [Report, Aix Marseille Université]. https://hal-amu.archives-ouvertes.fr/hal-02493679

Duca, D., & Metzler, K. (2019). The Ecosystem of Technologies for Social Science Research. SAGE Publishing. https://doi.org/10.4135/wp191101

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). Preprinting the COVID-19 pandemic. BioRxiv, 2020.05.22.111294. https://doi.org/10.1101/2020.05.22.111294

Gregory, K. (2020). A dataset describing data discovery and reuse practices in research. Scientific Data, 7(1), 232. https://doi.org/10.1038/s41597-020-0569-5

Grudniewicz, A., Moher, D., Cobey, K. D., Bryson, G. L., Cukier, S., Allen, K., Ardern, C., Balcom, L., Barros, T., Berger, M., Ciro, J. B., Cugusi, L., Donaldson, M. R., Egger, M., Graham, I. D., Hodgkinson, M., Khan, K. M., Mabizela, M., Manca, A., … Lalu, M. M. (2019). Predatory journals : No definition, no defence. Nature, 576(7786), 210-212. https://doi.org/10.1038/d41586-019-03759-y

Gruson-Daniel, C. (2018). Numérique et régime français des savoirs en~action : L'open en sciences : le cas de la consultation « république numérique » (2015) [Phdthesis, Université Sorbonne Paris Cité]. https://tel.archives-ouvertes.fr/tel-02515981

Gruson-Daniel, C., & Jean, B. (2021). Étude relative à l'ouverture des codes sources au sein de l'Enseignement Supérieur et de la Recherche (ESR) : Considérations en termes d'usage et de valeur. [Research Report]. INNO3 ; Etalab ; Comité pour la Science Ouverte. https://hal.archives-ouvertes.fr/hal-03125456

Khelfaoui, M. (s. d.). L'effet SIGAPS : La recherche médicale française sous l'emprise de l'évaluation comptable. Consulté 18 octobre 2020, à l'adresse https://www.cirst.uqam.ca/publications/leffet-sigaps-la-recherche-medicale-francaise-sous-lemprise-de-levaluation-comptable-2/

Kramer, B., & Bosman, J. (2016). Innovations in scholarly communication – Global survey on research tool usage. F1000Research, 5, 692. https://doi.org/10.12688/f1000research.8414.1

Le Béchec, M. (2021). Jean-Louis Genard, Marta Roca i Escoda, Éthique de la recherche en sociologie. Revue d'anthropologie des connaissances, 15 (1), Article 1. https://doi.org/10.4000/rac.19116

Le Béchec, M. L., & Charrier, Philippe. (2020a). Questionnaire SOSP-FR – fr. https://doi.org/10.5281/zenodo.3935958

Le Béchec, M. L., & Charrier, Philippe. (2020b). Questionnaire SOSP-FR – en. https://doi.org/10.5281/zenodo.3935971

Leonelli, S. (2019). La recherche scientifique à l'ère des Big Data : Cinq façons dont les Big Data nuisent à la science et comment la sauver (F. Galicia, Trad.).

Mancilla, H. A., Teperek, M., Dijck, J. van, Heijer, K. den, Eggermont, R., Plomp, E., Velden, Y. T. der, & Kurapati, S. (2019). On a Quest for Cultural Change – Surveying Research Data Management Practices at Delft University of Technology. LIBER Quarterly, 29(1), 127. https://doi.org/10.18352/lq.10287

MESRI, L. (2020). L'état de l'Emploi scientifique en France. Ministère de l'enseignement supérieur, de la recherche et de l'innovation. http://www.enseignementsup-recherche.gouv.fr/pid35205/etat-de-l-emploi-scientifique.htmlp. 23, MESRI, 2020

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Watkinson, A., Świgon, M., Xu, J., Abrizah, A., & Herman, E. (2018). Early career researchers: Observing how the new wave of researchers is changing the scholarly communications market. Revue Française Des Sciences de l'information et de La Communication, 15, Article 15. https://doi.org/10.4000/rfsic.4635

Nicholas, D., Rodríguez-Bravo, B., Watkinson, A., Boukacem-Zeghmouri, C., Herman, E., Xu, J., Abrizah, A., & Świgoń, M. (2017). Early career researchers and their publishing and authorship practices: ECRs publishing and authorship practices. Learned Publishing, 30 (3), 205217. https://doi.org/10.1002/leap.1102

O'Brien, D. (s. d.). Open data surveys: How comparable are they and their policy development applications. 68.

Pontille, D., & Torny, D. (2013). La manufacture de l'évaluation scientifique. Reseaux, n° 177 (1), 2361.

Prost, H., & Schöpfel, J. (2015). Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. [Research Report]. Lille 3. https://hal.univ-lille.fr/hal-01198379

Puebla, I., Polka, J., & Rieger, O. (2021). Preprints: Their Evolving Role in Science Communication. MetaArXiv. https://doi.org/10.31222/osf.io/ezfsk

Research Data Management Survey 2019: The results are here! (2019, décembre 2). Open Working. https://openworking.wordpress.com/2019/12/02/research-data-management-survey-2019-the-results-are-here/

Rousseau-Hans, F., Ollendorff, C., & Harnais, V. (2020). Pratiques de publications et d'accès ouvert des chercheurs français en 2019 [Other]. Consortium Couperin. https://doi.org/10.5281/zenodo.3948265

Sbeih Lina, Dedet Fanny, Moreau Patrick, Dzale Esther (2020). L'archivage des données de la recherche à l'Inra. Élément de réflexion, démarche et perspectives. *Cahier des Techniques de l'INRA*, INRA, 2020. ffhal-02861909f

Schöpfel, J. (2018). Vers une culture de la donnée en SHS [Research Report]. Université de Lille. https://hal.archives-ouvertes.fr/hal-01846849

Tennant, J. P., Dugan, J. M., Graziotin, D., Jacques, D. C., Waldner, F., Mietchen, D., Elkhatib, Y., B. Collister, L., Pikas, C. K., Crick, T., Masuzzo, P., Caravaggi, A., Berg, D. R., Niemeyer, K. E., Ross-Hellauer, T., Mannheimer, S., Rigling, L., Katz, D. S., Greshake Tzovaras, B., Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research, 6, 1151. https://doi.org/10.12688/f1000research.12037.3

Vignier, S. (2014). Réseaux sociaux de la recherche et Open Access: Perception des chercheurs Étude exploratoire novembre 2014 (p. 61). Couperin. https://www.couperin.org/images/stories/openaire/Couperin_RSDR%20et%20OA_Etude%20exploratoire_2014.pdf

INDEX DES FIGURES

Figure 1: Effectifs des répondants par disciplines (9 catégories) 22
Figure 2: Fonction déclarée dans la recherche des 1089 répondants 24
Figure 3: Répartition statutaire agrégée 26
Figure 4: Âge déclaré réparti par genre 27
Figure 5: Statut professionnel agrégé en fonction de l'année de naissance 28
Figure 6: Contexte de travail par le nombre de personnes participant à une opération de recherche 28
Figure 7: Usage des outils de gestion du travail avant le confinement 32
Figure 8: Usage des outils de gestion du travail après le confinement 33
Figure 9: Canaux d'accès à l'information scientifique 34
Figure 10: Les canaux de découverte des outils numériques 35
Figure 11: Répartition de l'usage de logiciels payants ou libres et gratuits selon la classe d'âge 38
Figure 12: Répartition de l'usage de logiciels payants ou libre et gratuit selon la discipline 38
Figure 13: Logiciels, suite et langages de programmation cités les plus fréquemment pour les 4 fonctions : production, nettoyage, analyses et visualisation des données 41
Figure 14: AFC croisant la variable outils de production avec la discipline du répondant 42
Figure 15: AFC croisant la variable outils de production avec la fonction du répondant 43
Figure 16: AFC croisant la variable outils d'analyse avec la discipline du répondant 45
Figure 17: AFC croisant les 10 premiers outils de visualisation avec la discipline du répondant 47
Figure 18: Fréquence d'usage des types de données travaillées 56
Figure 19: La pratique de réutilisation des données déjà produites ou publiées 57
Figure 20: La pratique de réutilisation des données déjà produites ou publiées selon la discipline 58
Figure 21: La pratique d'archivage (conservation à long terme) des données 59
Figure 22: La pratique d'archivage à long terme selon la discipline 60
Figure 23: La pratique d'archivage à long terme selon le contexte de travail 61
Figure 24: La pratique d'archivage à long terme selon la fonction 61
Figure 25: La pratique d'archivage à long terme selon la classe d'âge 62
Figure 26: Répartition de l'usage des métadonnées 63
Figure 27: Tiers d'archivage utilisé en rapport avec la fonction 65

Figure 28 : Tiers d'archivage utilisé en rapport avec le contexte de travail 66
Figure 29 : Tiers d'archivage utilisé en rapport avec la discipline 67
Figure 30 : Répartition des outils d'enregistrement et de sauvegarde des données 68
Figure 31 : Répartition de la pratique fréquente des outils d'enregistrement et de sauvegarde selon le statut 69
Figure 32 : Les outils de rédaction ou d'aide à la rédaction des publications 75
Figure 33 : Choix des formats de publication des résultats de recherche 75
Figure 34 : Fréquence de la publication dans une revue à comité de lecture selon la classe d'âge 76
Figure 35 : Fréquence de la publication sur les réseaux sociaux numériques selon la classe d'âge 77
Figure 36 : Pratique de diffusion des prépublications sur des plateformes « *non for profit* » selon la discipline 79
Figure 37 : Usages des canaux de diffusion des données de recherches produites 80
Figure 38 : Diffusion des données sur des plateformes d'éditeurs en complément de l'article selon la discipline 81
Figure 39 : Formats de valorisation des résultats de la recherche utilisés 83
Figure 40 : Connaissance des plateformes d'évaluation ouverte par les pairs ? 84
Figure 41 : Dendrogramme des 338 réponses à la question de compatibilité (Iramuteq, Lucie Loubère) 87
Figure 42 : Matrice d'interprétation des questionnements autour de la compatibilité diffusion-valorisation des données de recherche (excalidraw) 88

INDEX DES TABLEAUX

TABLEAU 1 : Les dix enquêtes consultées pour l'état de l'art 17
TABLEAU 2 : Comparaison entre effectifs et part des chercheurs entre l'enquête du MESRI et SOSP-FR 23
TABLEAU 3 : Statut professionnel des répondants 26
TABLEAU 4 : Genre déclaré par le répondant 27
TABLEAU 5 : Tableau de comparaison du contexte de travail selon trois variables (âge, discipline, fonction) 29
TABLEAU 6 : Répartition des différentes configurations en termes d'OS et leur poids parmi la population répondante 31
TABLEAU 7 : Répartition des sources de connaissance des outils selon le contexte de travail 36
TABLEAU 8 : Types de logiciels de production de données utilisés 37
TABLEAU 9 : Liste des logiciels et langages de programmation cités pour la production de données de recherche 40
TABLEAU 10 : Liste des 15 logiciels et langages de programmation cités pour le nettoyage des données de recherche 44
TABLEAU 11 : Liste des 20 logiciels et langages de programmation cités pour l'analyse des données de recherche 44
TABLEAU 12 : Liste des 24 logiciels et langages de programmation cités pour la visualisation des données de recherche 46
TABLEAU 13 : Répartition des utilisateurs du logiciel R en fonction de leur classe d'âge 48
TABLEAU 14 : Répartition des utilisateurs du logiciel R en fonction de leur discipline 49
TABLEAU 15 : Répartition des utilisateurs du logiciel R en fonction de leur contexte de travail 50
TABLEAU 16 : Répartition des utilisateurs du logiciel Excel en fonction de leur discipline 50
TABLEAU 17 : Répartition des utilisateurs de Python selon leur classe d'âge 51
TABLEAU 18 : Répartition des utilisateurs de Python selon leur discipline 52
TABLEAU 19 : Répartition des utilisateurs de Python selon leur contexte de travail 52
TABLEAU 20 : Répartition des utilisateurs de Python selon leur fonction 53
TABLEAU 21 : Répartition des utilisateurs de Python selon leur système d'exploitation 53
TABLEAU 22 : Répartition des utilisateurs de Python selon leur préférence de type de logiciels 54
TABLEAU 23 : Répartition des utilisateurs de Python selon leur genre 54

TABLEAU 24 : Le volume de données estimé par le chercheur 57
TABLEAU 25 : Le tiers de confiance pour l'archivage des données de travail du chercheur 64
TABLEAU 26 : Tiers d'archivage mobilisé en fonction du statut 64
TABLEAU 27 : Souhait de partage des données et résultats de la recherche 70
TABLEAU 28 : Répartition de la sous-population de répondants défavorables au partage des données de la recherche 71
TABLEAU 29 : Répartition par date de naissance de la sous-population de répondants défavorables au partage des données de la recherche 72
TABLEAU 30 : Répartition par statut de la sous-population de répondants défavorables au partage des données de la recherche 72
TABLEAU 31 : Disciplines des répondants (n = 263) défavorables au partage des données de la recherche 73
TABLEAU 32 : Critères de choix pour soumettre une publication ou une communication (Question 26) pour les avis défavorables à l'ouverture des données de recherche 73
TABLEAU 33 : Fréquence de la publication dans une revue à comité de lecture selon la discipline 76
TABLEAU 34 : Critères de choix pour soumettre une publication ou une communication 78
TABLEAU 35 : Canaux de diffusion des prépublications non-évaluées (*preprint*) 79
TABLEAU 36 : Répartition de la diffusion des autres types de production de la recherche 82
TABLEAU 37 : Connaissance des plateformes d'évaluation ouverte par les pairs 85
TABLEAU 38 : Tableau des 7 caractéristiques des adhérents à la SO dans sa dimension humaine 95
TABLEAU 39 : Moyennes pour les types de données travaillées 96
TABLEAU 40 : Les 7 caractéristiques des adhérents à la SO dans sa dimension technique 97
TABLEAU 41 : Répartition disciplinaire des adhérents à la SO dans sa dimension technique 97
TABLEAU 42 : Effectifs et répartitions des chercheurs par disciplines de recherche et statut des établissements employeurs en 2019 110
TABLEAU 43 : Annexe 2. Répartition des chercheurs selon les disciplines et le secteur d'activité 111
TABLEAU 44 : Annexe 3. Nomenclature des disciplines d'activité de recherche des enquêtes R&D et du tableau de bord de l'emploi scientifique (Sies) : secteur public et secteur privé 112

ANNEXES

ANNEXES 1. Effectifs et répartitions des chercheurs par disciplines de recherche et statut des établissements employeurs en 2019

	Effectifs				% de chaque discipline **			
Discipline d'activité de recherche *	EPST ***	8 EPIC-ISBL	EPSCP ****	Ensemble	EPST ***	8 EPIC-ISBL	EPSCP ****	Ensemble
Mathématiques	2916	358	3368	6642	11,1	2,8	11,6	9,8
Sciences physiques	3218	1472	1315	6005	12,3	11,6	4,5	8,8
Chimie	2490	540	1600	4630	9,5	4,3	5,5	6,8
Sciences de l'ingénieur 1	803	4061	1771	6635	3,1	32,0	6,1	9,8
Sciences de l'ingénieur 2	1254	2713	1815	5782	4,8	21,4	6,2	8,5
Sciences de la terre/ Environnement	2792	371	569	3732	10,6	2,9	2,0	5,5
Sciences agricoles	97	175		272	0,4	1,4		0,4
Sciences biologiques	9296	2000	2692	13988	35,4	15,8	9,3	20,6
Sciences médicales	426	176	3932	4534	1,6	1,4	13,5	6,7
Sciences sociales	1377	194	6091	7662	5,2	1,5	21,0	11,3
Sciences humaines	1566	4	5468	7038	6,0		18,8	10,4
STAPS			418	418			1,4	0,6
Total	**26235**	**12064**	**29040**	**67339**	**100**	**100**	**100**	**100**

TABLEAU 42 : Effectifs et répartitions des chercheurs par disciplines de recherche et statut des établissements employeurs en 2019

ANNEXE 2. Répartition des chercheurs selon les disciplines et le secteur d'activité

Discipline d'activité de recherche*	Entreprises, doctorants inclus		Principaux secteurs du public	
	Eff. fin 2015, en PP	%	Eff. fin 2018, en EER	%
Mathématiques	44038	20,1	6644	9,7
Sciences physiques	6404	2,9	6001	8,8
Chimie	9003	4,1	4571	6,7
Sciences de l'ingénieur 1	70469	32,1	6762	9,9
Sciences de l'ingénieur 2	60689	27,7	5791	8,5
Sciences de la terre/ Environnement	2601	1,2	3782	5,5
Sciences agricoles	5111	2,3	301	0,4
Sciences biologiques	8134	3,7	14613	21,3
Sciences médicales	8293	3,8	4465	6,5
Sciences sociales	3445	1,6	7610	11,1
Sciences humaines	1185	0,5	6920	10,1
Sûreté, sécurité			618	0,9
STAPS			423	0,6
Sous-total	219372	100	68501	100
Gestion/Encadrement de la R&D; non renseigné	6364		4625	
Total chercheur	**225736**		**73126**	

TABLEAU 43 : Annexe 2. Répartition des chercheurs selon les disciplines et le secteur d'activité

La nomenclature utilisée a été construite pour s'appliquer à des activités de recherche tant pour le secteur privé que pour le secteur public. Dans notre enquête, les chercheurs du secteur privé sont trop peu nombreux pour que l'on applique la répartition proposée par le tableau ci-dessus. C'est le cas tout particulièrement des deux catégories de sciences de l'ingénieur, catégories qui ont un poids considérable dans le secteur privé alors qu'elles sont nettement plus discrètes dans le secteur public.

ANNEXE 3. Nomenclature des disciplines d'activité de recherche des enquêtes R&D et du tableau de bord de l'emploi scientifique (Sies) : secteur public et secteur privé

01. Mathématiques et informatique (conception de logiciel)

02. Sciences physiques

03. Chimie

04. Sciences de l'ingénieur 1 : informatique, automatique, traitement du signal, électronique, photonique, optronique, génie électrique

05. Sciences de l'ingénieur 2 : mécanique, génie des matériaux, acoustique, génie civil, mécanique des milieux fluides, thermique, énergétique, génie des procédés

06. Sciences des milieux naturels ou de l'univers (terre, océan, atmosphère, espace)

07. Sciences de l'agriculture et alimentation

08. Sciences de la vie et biologie fondamentale

09. Sciences médicales et odontologie

10. Sciences sociales : sociologie, démographie, ethnologie, géographie, aménagement de l'espace, économie et gestion, sciences politiques et juridiques, psychologie

11. Sciences humaines : philosophie, histoire, archéologie, anthropologie, littérature, linguistique, langues, sciences de l'art

12. Gestion de la R&D : fonction de gestion et d'encadrement des activités de R&D exclusivement

TABLEAU 44 : Annexe 3. Nomenclature des disciplines d'activité de recherche des enquêtes R&D et du tableau de bord de l'emploi scientifique (Sies) : secteur public et secteur privé