Données de la recherche : pratiques et besoins dans un laboratoire pluridisciplinaire SHS

Anne Bonneville, Ingrid Tucci, Antoine Vion, Laurent Giglio

Données de la recherche : pratiques et besoins dans un laboratoire pluridisciplinaire SHS

Anne Bonneville (coord.), Ingrid Tucci, Antoine Vion, Laurent Giglio

Mai 2021 – Rapport final

Laboratoire d'Economie et de Sociologie du Travail.
(LEST UMR 7317)
Aix-Marseille-Université, CNRS, LEST
35 avenue J. FERRY - 13626 AIX-EN-PROVENCE CEDEX 01
Tél. : Standard : 04.42.37.85.00 - Fax : 04.42. 26.79.37
Email lest@univ-amu.fr
http://lest.fr/

TABLE DES MATIERES

Les auteur·es remercient l'ensemble des membres du laboratoire pour leur participation à cette enquête.

INTRODUCTION

Le Laboratoire d'Economie et de Sociologie du Travail (LEST) est une unité mixte de recherche en Sciences Humaines et Sociales du CNRS (UMR 7317) et de l'université d'Aix-Marseille (AMU). Ses recherches portent sur les dynamiques du travail et l'emploi. Il est Centre associé du Céreq pour les régions Provence-Alpes-Côte d'azur et Corse. Il rassemblait au moment de l'enquête :

- 57 chercheur·es et enseignant·es-chercheur·es,
- 13 post-doctorant·es,
- 37 doctorant·es lesquels sont rattaché·es à deux écoles doctorales (ED 372 Economie-Gestion ; ED 355 Sociologie),
- 24 membres associé·es.

Le LEST est pluridisciplinaire — la gestion, la sociologie et l'économie étant les disciplines majoritaires. Les données de recherche sont au cœur des activités de recherche et, parce qu'elles s'inscrivent dans des enjeux et problématiques actuelles fortes, il est apparu fondamental de faire l'état des pratiques dans ce domaine et d'engager une réflexion en interne.

Cette enquête a été lancée en accord avec la direction en 2019-2020 au sein au LEST et visait trois objectifs :

1. Appréhender les pratiques, connaissances et difficultés en matière de collecte, d'organisation et de traitement des données de recherche pour mieux appréhender les besoins.

2. Amorcer au sein du LEST une réflexion sur la place des données dans le travail de recherche pour favoriser la mutualisation, voire le renforcement de l'expertise existante au sein du laboratoire dans ce domaine.

3. Réfléchir enfin sur les ressources et sur l'accompagnement qui pourrait être mis en place avec le soutien de la direction et des partenaires institutionnels.

L'enquête a été conduite en deux phases. Une première phase, en 2019-2020, a consisté à adresser et exploiter un questionnaire à destination des membres du laboratoire. Le questionnaire complet est consultable en annexe. La méthodologie et les résultats de cette enquête sont présentés dans la première partie de ce rapport. Une deuxième phase, fin 2020, a été consacrée à l'organisation de *focus groups* avec des chercheur·es volontaires. Les synthèses de ces trois discussions sont présentées dans la deuxième partie de ce rapport. La troisième partie présente un ensemble de conclusions en dressant à la fois les objectifs d'une politique de la donnée pour un laboratoire de SHS comme le LEST et les enjeux d'organisation qu'une telle politique recouvre.

I. ENQUÊTE PAR QUESTIONNAIRE

L'enquête par questionnaire a été conduite en 2019, soit avant la réorganisation du travail imposée par la crise sanitaire. La méthodologie (1) et les résultats (2) en sont présentés dans cette première partie.

1 MÉTHODOLOGIE, ECHANTILLON

1.1 Le questionnaire

A partir de deux enquêtes similaires sur la gestion des données menées à Lille *(Prost 2015*) et Rennes (*Serres 2017*), nous avons pu élaborer notre propre questionnaire en conservant la trame principale. Certaines questions ont été ajoutées ou modifiées pour adapter l'enquête au contexte du terrain. Le Règlement Général sur la Protection des Données (RGPD), et la documentation sur les données, ont fait l'objet d'une plus grande attention que dans les précédentes enquêtes. Les méthodes dans la pratique de recherche ont été également introduites pour permettre de relier les données aux méthodes employées par les équipes du laboratoire.

Il existe diverses définitions des données de la recherche selon les approches des acteurs et des institutions. Pour cette enquête, nous avons choisi de retenir la définition générale suivante : « *Un ensemble d'informations factuelles enregistrées sur des supports, produites ou collectées selon divers procédés au cours d'un processus de recherche* » *(Cartier et al. 2015)*.

Nous avons conservé la distinction entre données sources et données résultats selon la typologie de Chloé Gauquelin et al. (2017) afin de distinguer ce qui est utilisé et ce qui est concrètement produit par les chercheur·es. Un glossaire des termes, posé en amont du questionnaire et diffusé au moment de l'enquête, a permis de fixer le cadre conceptuel afin de limiter les ambiguïtés au moment des réponses. L'enquête s'est centrée uniquement sur les données de recherche. Tout autre document comme les publications scientifiques, les bibliographies, les documents administratifs, les documents pédagogiques en ont été exclus. Le cadre posé a été le suivant :

DONNÉES SOURCES	**Données qui constituent les matériaux de base de la recherche (avant tout traitement et analyse)**
DONNÉES SOURCES EXTERNES	**Données non produites par la ou le chercheur·e (en libre accès, sur demande ou en réutilisation).** *Exemple : Base de données quantitatives externes ; corpus et archives textuelles collectées par d'autres personnes/ institutions, archives sonores/audiovisuelles ; statistiques produites par des acteurs publics/privés, contenus de sites web*
DONNÉES SOURCES COLLECTÉES	**Données collectées et/ou compilées par la ou le chercheur·e en direct ou à distance.** *Exemple : Base de données élaborée à partir de plusieurs sources, base de données d'une enquête par questionnaire/sondage, données d'enquête de terrain (entretiens, transcriptions, observations, photos, note de terrain…), corpus textuels élaborés à partir de données du web*
DONNÉES RESULTATS	**Données produites après exploitation et traitement des données sources et qui constituent les données résultats** *Exemple : données d'analyses statistiques, d'analyses qualitatives, d'analyses réseaux, extraits d'entretiens, cartographies, graphes, modèles et indicateurs, données agrégées chiffrées, productions audiovisuelles, synthèses, notes, annotations*
DOCUMENTATION SUR LES DONNÉES	**Données qui renseignent sur les données elles-mêmes** *(méthodologie, matériaux d'analyse utilisés, traitements appliqués, propriété et protection des données, valorisation des données*) *Exemple : Guide de procédure/méthodologie d'enquête, programme et application, grille d'analyse, dictionnaire de variables, demande d'autorisation à la CNIL, clauses de confidentialité, métadonnées (titre, mots-clés, DOI)*

L'enquête s'adressait à celles et ceux qui produisent des résultats de recherche au laboratoire, et ce quel que soit le statut (permanents, retraité·es émérites, associé·es, contractuel·les ou stagiaires). Il s'est agi pour nous en effet d'entrer dans une logique d'autoanalyse de « la science telle qu'elle se fait » (Callon, Latour 1991), plutôt que dans une logique évaluative préformatée par la « manufacture de l'évaluation scientifique » (Pontille, Torny 2013). Les champs disciplinaires présents sont ceux du laboratoire au moment de l'enquête : l'économie, la sociologie, la gestion, la science politique, le droit, la géographie, l'anthropologie. Notons que ces quatre dernières disciplines sont largement minoritaires.

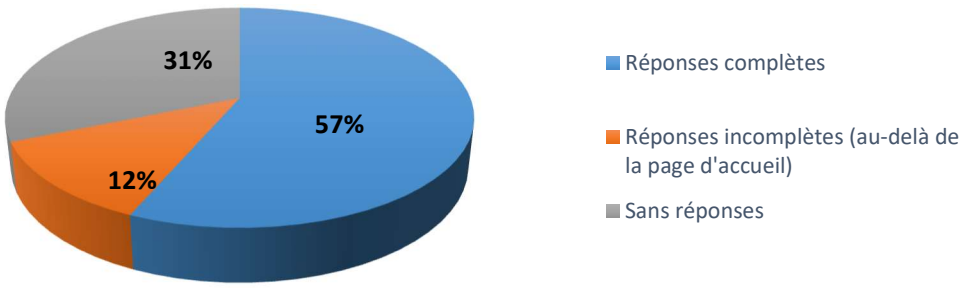
Le questionnaire s'est articulé autour de 4 thèmes principaux : 1) les méthodes de recherche utilisées et les données associées, 2) les usages (pratiques de stockage et d'archivage, de partage et de diffusion), 3) les besoins et attentes et 4) le profil des enquêté·es. Il comportait dans sa structure 41 questions — 27 facultatives et 14 obligatoires. Le temps de réponse était estimé à 20 minutes. La programmation du questionnaire dans LimeSurvey et la passation de l'enquête ont été confiées à Sylvain Bausson (prestataire). Au final, nous disposons d'une base de données de 160 variables.

Lancée le 20 novembre 2018 et close le 31 janvier 2019, elle a duré un peu plus de deux mois avec plusieurs relances. Des informations sur l'enquête ont été passées au sein des axes de recherche du laboratoire et deux interventions ont eu lieu afin de mobiliser les collègues : une intervention en séminaire général du laboratoire et une autre dans l'assemblée générale de l'association des doctorant·es Déclic (avec proposition d'appui pour les aider à remplir le questionnaire).

1.2 Participation à l'enquête, taux de réponses

131 invitations individuelles ont été lancées depuis LimeSurvey. La participation a été plutôt satisfaisante avec un taux de **réponse de 69 %** (91 participant·es). Au total, 16 personnes n'ont pas terminé le questionnaire. Nous disposons ainsi pour cette enquête de **57 % de réponses complètes** (74 répondant·es).

Figure 1 : Participation à l'enquête « Données de la recherche au LEST »



Le taux de participation a varié selon le statut. Il était plus faible parmi les associé·es, les post-doctorant·es ou encore les jeunes doctorant·es, une population se sentant vraisemblablement moins concernée par cette enquête.

1.3 Aperçu du profil des répondant·es

Les tableaux ci-dessous donnent un aperçu du profil des répondant·es à l'enquête : 57% sont des femmes, un taux qui correspond à leur représentation dans le champ de l'enquête (58%). 49% sont enseignant·es-chercheur·es, 19% chercheur·es et associé·es, 26% doctorant·es dont 7% de jeunes en 1ère année de thèse au moment de l'enquête, 7% sont post-doctorant·e. Parmi l'ensemble des doctorant·es répondant·es, un peu moins des trois-quarts sont rattaché·es à l'école doctorale ED355 et un peu plus du quart sont rattaché·es à l'école doctorale ED372. Le taux de participation a été le plus faible parmi les associé·es et les post-doctorant·es.

Tableau 1 : Profil des répondant·es (N=74)

SEXE	**%**
Hommes	43
Femmes	57

STATUT	**%**
Chercheur·es & Associé·es	19
Enseignant·es-chercheur·es	49
Post-doctorant·es	7
Doctorant·es avancé·es	19
Jeunes Doctorant·es	7

DISCIPLINE	**%**
Gestion	27
Economie	19
Socio/Sciences politiques /Anthro	49
Autres disciplines	5

TRANCHE D'AGE	**%**
Moins de 30 ans	16
30-40 ans	38
41-50 ans	19
51 ans et plus	27

DUREE D'ACTIVITE	**%**
Depuis moins de 5 ans	23
5-10 ans	14
11-15 ans	24
16-20 ans	11

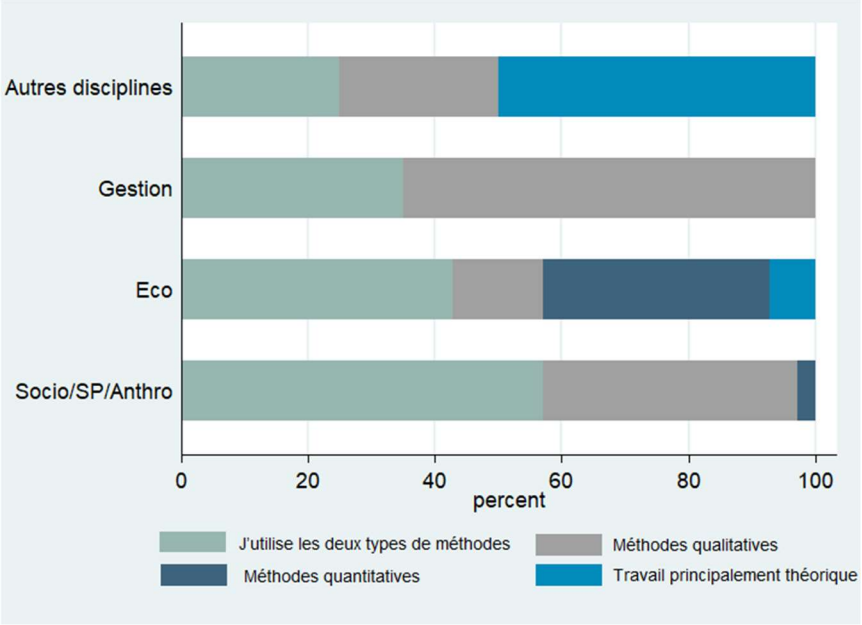
2 PRÉSENTATION DES RESULTATS

2.1 Méthodes de recherche

Sur l'ensemble des personnes qui ont répondu à la question sur les méthodes utilisées dans leurs recherches : **46%** utilisent aussi bien les approches quantitatives que qualitatives (pas nécessairement dans une approche « *mixed methods* ») et **43%** utilisent uniquement les méthodes qualitatives. Les répondant·es pratiquant exclusivement les méthodes quantitatives ou faisant un travail principalement théorique sont largement minoritaires (8% pour les premier·es et 3% pour les second·es).

Un croisement des méthodes avec la discipline montre des différences dans leurs utilisations (Figure 2) : les méthodes qualitatives sont majoritaires chez les gestionnaires tandis qu'en économie, sociologie, sciences politiques et anthropologie, l'utilisation des deux types de méthodes prédomine.

Figure 2 : Types de méthodes par disciplines



Pour celles et ceux qui utilisent les méthodes quantitatives (soit exclusivement ou à côté d'autres approches), les statistiques descriptives sont les plus courantes (Tableau 2), suivies des analyses économétriques et des analyses géométriques de données (ACP, ACM, analyses factorielles). Il est entendu que ces parts varient en fonction de la discipline : 90 % des économistes utilisent des modèles économétriques contre 43% des gestionnaires et 38% des sociologues, politistes et anthropologues. Les économistes sont celles et ceux qui utilisent le plus de méthodes quantitatives différentes.

Pour ce qui concerne les méthodes qualitatives, les entretiens prévalent largement, suivis de l'observation et de l'analyse de documents (textuels, d'archives, films, …). La réalisation de *focus groups* arrive en dernière position au laboratoire. Les méthodes qualitatives utilisées sont les plus variées en gestion et elles le sont un peu plus qu'en sociologie.

Tableau 2 : Types de méthodes quantitatives et qualitatives utilisées

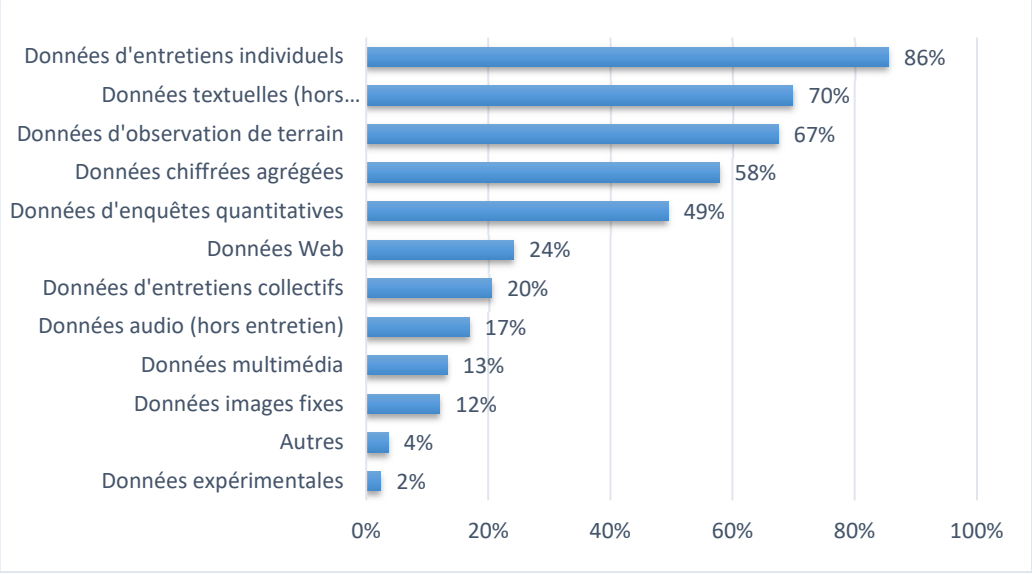
Méthodes quantitatives (N=48)	**%**	**Méthodes qualitatives** (N=79)	*%*
Statistiques descriptives	92	Entretiens	97
Econométrie	56	Observations	71
Analyses géométriques de données	54	Analyse de documents	68
Analyse de séquences et trajectoires	38	Ethnographie	35
Analyse de réseaux	6	Méthode participative	35
Autres	2	*Focus groups*	23
		Autres	1

2.2 Données de recherche

2.2.1 Données sources

Les données sources ont été définies pour cette enquête comme des matériaux de base avant traitement et analyse. Les données qui prédominent se justifient facilement par les pratiques de recherche menées sur le terrain et les disciplines du laboratoire (sciences sociales) : **au 1er rang arrivent les données d'entretien (86 %, Figure 3).** Elles *donnent lieu dans les trois quarts des cas systématiquement à des transcriptions.* Arrivent ensuite **les données textuelles et les données d'observation *(70 % et 67 %).*** Les enquêtes statistiques et données agrégées concernent une population de chercheur·es moins importante. En particulier, les données quantitatives utilisées sont plus souvent des **données agrégées** que des données d'enquête (micro-données). Les données de nature multimédia sont, elles, nettement moins utilisées dans les recherches.

Figure 3 : Catégories des données sources

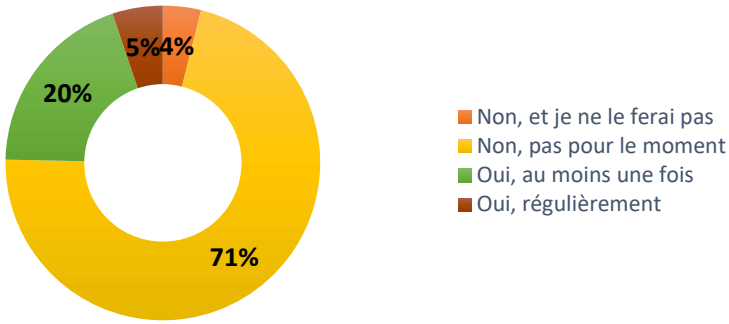


2.2.2 Données réutilisées

Les trois-quarts des chercheur·es ont déclaré ne jamais avoir utilisé des données collectées par d'autres collègues (Figure 4). Parmi ces derniers, une petite partie déclarent qu'ils ne le feraient pas. Un quart a en revanche déjà travaillé à partir de données d'autres collègues. Les doctorant·es sont ici plus concerné·es que les chercheur·es.

La réutilisation de données est donc une pratique peu fréquente au laboratoire, liée semble-t-il plus aux pratiques de recherche qu'à un véritable refus de la démarche.

Figure 4 : Réutilisation de données d'autres collègues



2.2.3 Données résultats

Les données résultats sont les données qui découlent de l'exploitation et du traitement des données sources. Les **données d'analyse textuelle** (76%) et les **statistiques** (65%) prédominent (Figure 5). On retrouve plus rarement les données liées à des supports médias. Celles et ceux qui utilisent des méthodes quantitatives *et* qualitatives produisent une plus grande variété de données (Figure 6).

Figure 5 : Données résultats (produites)

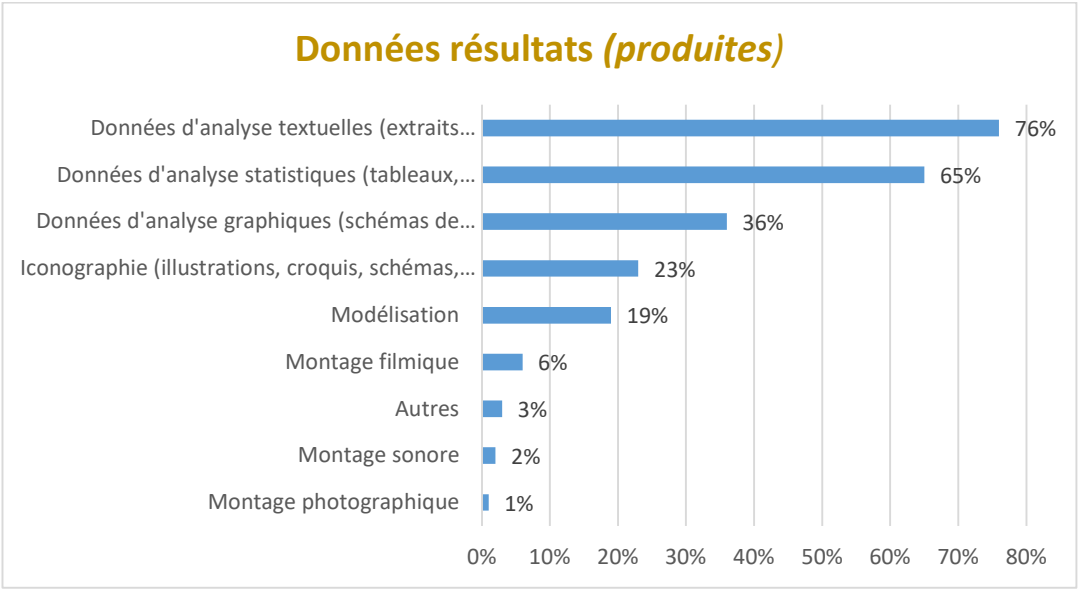
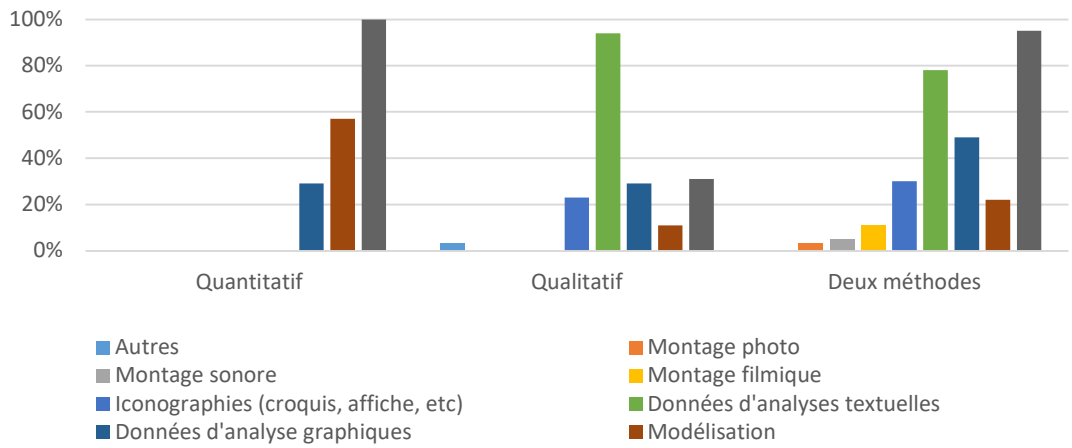


Figure 6 : Données produites selon la méthode (hors travail théorique)



2.2.4 Documentation produite sur les données

Un intérêt particulier doit être porté à la documentation produite au laboratoire sur les données de recherche. Cette documentation fournit une description sur les données produites, en apporte des éléments d'éclairage précis, la manière dont elles ont été collectées, ordonnées, analysées, leurs caractéristiques et les droits qui y sont appliqués. La documentation des données produites permet ainsi la reproduction et la réutilisation des résultats de recherche ; d'exposer par exemple les biais potentiels induits par le mode de collecte ou encore par l'échantillonnage. En cas de partage des données avec d'autres chercheur·es, elle

fournit les connaissances indispensables à leur traitement. Un autre intérêt non négligeable est celui de pouvoir les inscrire dans une démarche de valorisation *via* les infrastructures mises à disposition dans nos communautés scientifiques.

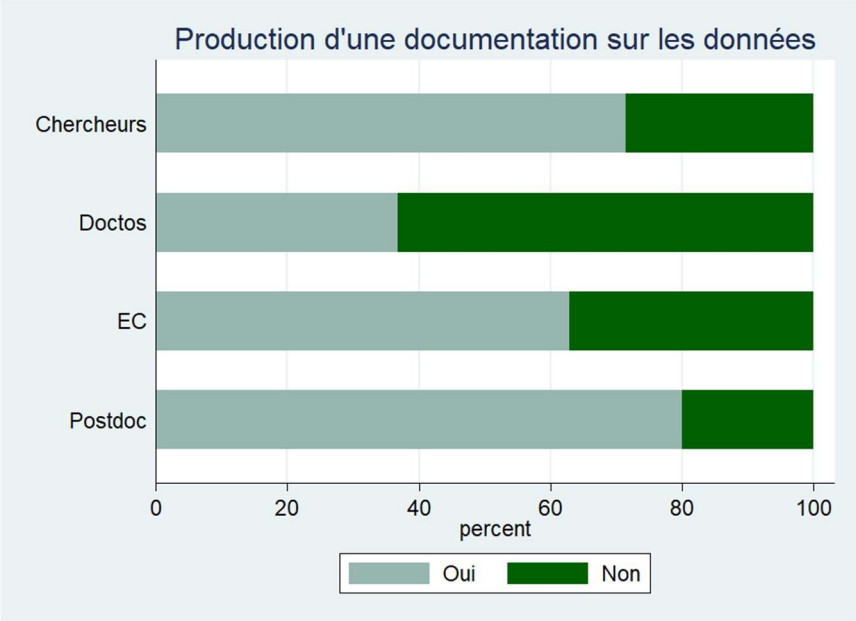
Plus de la moitié des chercheur·es (sans distinction de statut) au laboratoire produisent une documentation sur leurs données (59%). Elle est de nature essentiellement scientifique. L'information technique *(logiciel utilisé, liste des fichiers de données, format, …)* arrive en seconde position (Tableau 3). La part de la documentation standardisée pour la description des jeux de données *(métadonnées, mots-clés, DOI, droits associés sur les accès et licence*) témoigne d'une pratique moins courante au laboratoire.

Tableau 3 : Type de documentation produite

Documentation produite : *nature de l'information*	**%**
Information scientifique	93%
Information technique	54%
Information standardisée	43%
Information standardisée	**%**
Les deux types de méthodes	32%
Méthodes qualitatives	41%
Méthodes quantitatives	75%

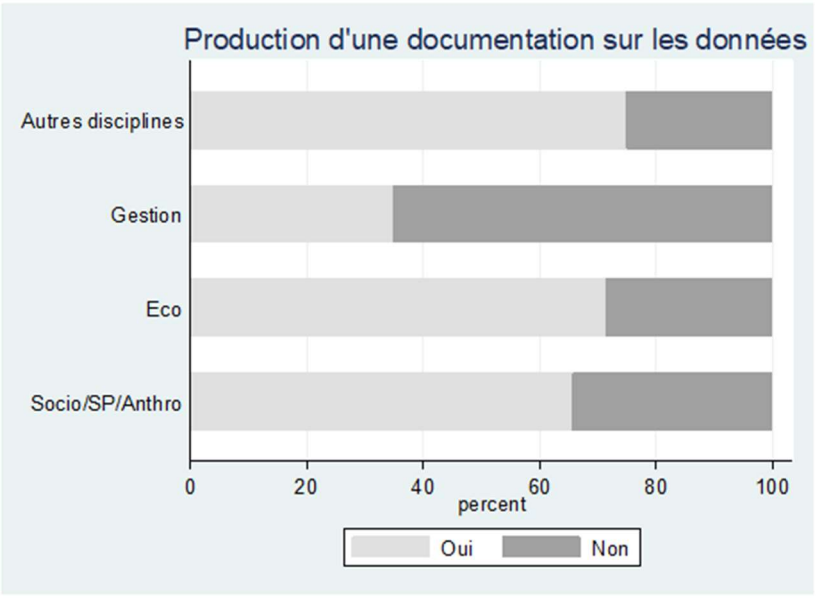
La part varie par ailleurs selon le statut : les post-doctorant·es sont celles et ceux qui produisent le plus de documentation, suivis des chercheur·es, et des enseignant·es-chercheur·es. La faible part parmi les doctorant·es s'explique en partie aussi par le fait que certain·es sont en début de doctorat (Figure 7).

Figure 7 : Part de la documentation sur les données par statut



L'économie et la sociologie/sciences politiques et l'anthropologie sont à un niveau très proche et au-dessus de la moyenne (Figure 8). En sciences de gestion, il y a moins souvent une production de documentation sur les données mais lorsque c'est le cas, cette documentation est plus variée.

Figure 8 : Part de la documentation par disciplines



En moyenne, la moitié de celles et ceux qui documentent leurs données produisent entre 4 et 8 types de documents (guide d'entretien, échantillon, méthode d'enquête, définition des populations, dictionnaire des variables), l'autre moitié produit en particulier une documentation sur la procédure d'enquête et la population enquêtée et, dans une moindre mesure, le guide d'entretien et la méthode d'échantillonnage (Tableau 4). Il s'agit donc dans les deux cas et pour l'essentiel d'une documentation de nature scientifique. **Tout ce qui relève de la gestion et de la sécurité des données est en revanche nettement moins souvent documenté.**

Tableau 4 : Type et diversité de la documentation produite

Type de documentation produite	**% total** (46 personnes)	**Parmi ceux/celles qui produisent de 1 à 4 types de documentation (%)** (20 personnes)	**Parmi ceux/celles qui produisent de 4 à 8 types de documentation (%)** (24 personnes)
Procédure d'enquête	91	90	100
Définition des populations	87	80	100
Guide d'entretien	67	45	92
Méthode d'échantillonnage	65	45	88
Dictionnaire de variables	37	15	58
Grille de codage données d'entretien	37	10	63
Carnet de recherche	20	10	29
Fichier d'exécution des traitements	17	10	25
Description de l'arborescence	17	5	29
Cadre de sécurité des données	9	0	17
Règle de nommage des jeux de données	4	0	8
Autres types de documentation	2	5	0

2.2.5 Sensibilité des données

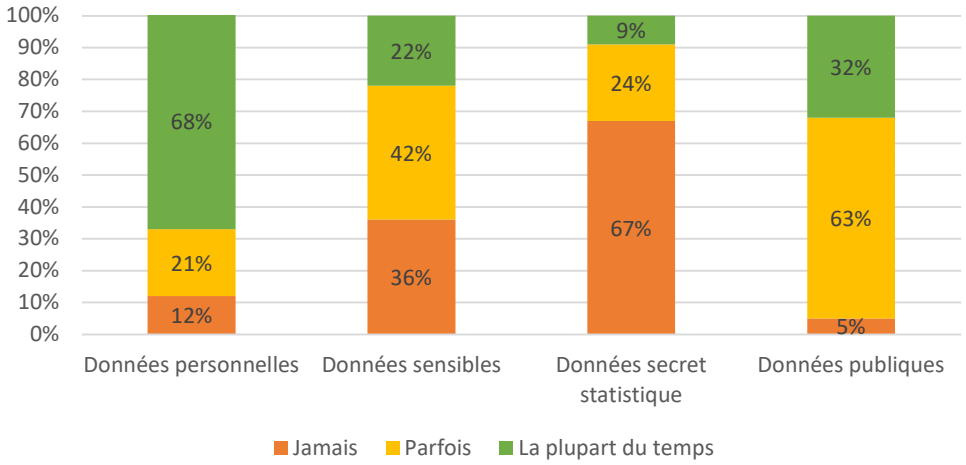
Plus des deux tiers des chercheur·es collectent des données personnelles issues de leur terrain et/ou auprès de fournisseurs[1] (Figure 9). Les données sensibles[2] représentent une part non négligeable des données sources : les deux tiers des enquêté·es en collectent parfois ou la plupart du temps. Sans surprise, c'est en anthropologie, sociologie et sciences politiques que les données personnelles ou sensibles sont collectées le plus fréquemment, de même lorsque les méthodes qualitatives sont utilisées. Les données concernées par le secret statistique[3] sont, elles, rarement utilisées au laboratoire alors que les données publiques le sont très largement.

[1] D'après l'article 2 de la loi informatique et libertés, une donnée à caractère personnelle consiste en « toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres ». Elle ajoute que « pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne ».

[2] Données sensibles est pris ici au sens de la définition de la Commission nationale de l'informatique et des libertés (CNIL) : https://www.cnil.fr/fr/cnil-direct/question/une-donnee-sensible-cest-quoi

[3] Données sous secret statistique : Données de statistiques publiques confidentielles dues à la vie privée, personnelle et familiale ou au secret commercial et des affaires :https://www.insee.fr/fr/statistiques/fichier/1300624/guide-secret.pdf

Figure 9 : Nature des informations collectées



Dans le cadre de leur terrain, les trois-quarts des chercheur·es prennent le plus souvent le soin d'informer les personnes à l'oral. En revanche, la formalisation écrite est beaucoup moins inscrite dans les pratiques : la moitié des chercheur·es n'informent jamais par écrit. De même, un peu plus des deux tiers ne demandent pas de consentement écrit. L'anonymisation permet de supprimer tout caractère identifiant à la différence de la pseudonymisation qui présente le risque d'une identification indirecte et réversible. En matière de sécurité des données, **les données collectées par les membres du laboratoire sont largement anonymisées ou pseudonymisées.** On ignore cependant ici s'il s'agit d'une anonymisation liée à la publication des résultats dans un article (par exemple des extraits d'entretiens ou des informations sur les cas étudiés) ou d'une anonymisation totale de l'ensemble des corpus en vue de ne laisser aucune information personnelle ou sensible qui pourrait permettre d'identifier la personne.

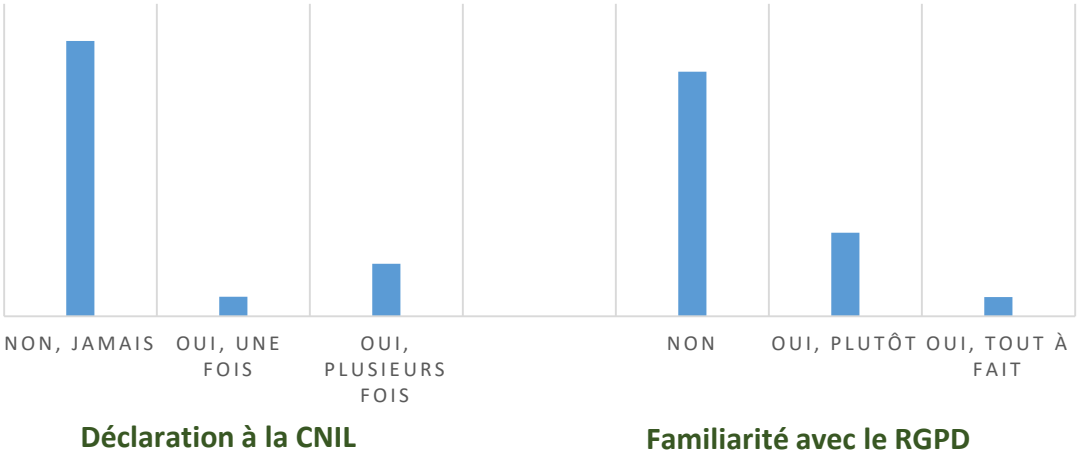
Les données personnelles et sensibles présentent une certaine **vulnérabilité pendant leur phase de traitement** : en effet, seule la moitié des enquêté·es les protègent « la plupart du temps », 28% ne les protègent jamais (Tableau 5). On observe également que **les données personnelles et sensibles ne sont pas systématiquement supprimées après le traitement** : un répondant sur deux indique ne jamais les supprimer. La question de la protection des données doit être également analysée au regard des pratiques de stockage et d'archivage des données de la recherche (partie suivante).

Tableau 5 : Gestion des données personnelles/sensibles collectées

	Jamais	Parfois	La plupart du temps
Suppression des données personnelles ou soumises à un régime administratif de protection	46%	31%	23%
Suppression des données personnelles et/ou sensibles après traitement	46%	31%	23%
L'accès est protégé durant le traitement	28%	24%	48%
Anonymisation des données	7%	17%	76%
Pseudonymisation des données	19%	17%	64%
Demande de consentement écrit	68%	24%	8%
Personnes informées par écrit	48%	27%	25%
Personnes informées par oral	10%	17%	73%

Ce que le graphique ci-dessous révèle, c'est qu'une pratique fréquente de collecte de données personnelles ou sensibles n'est pas en adéquation avec une connaissance de la nouvelle réglementation sur la protection des données, ni avec une pratique de déclaration d'une collecte ou d'un traitement à la CNIL : 79 % de celles et ceux qui ont déclaré collecter la plupart du temps des données personnelles ou sensibles n'ont jamais fait une déclaration à la CNIL (81 % de l'ensemble des répondant·es) et 70% ne sont pas du tout familier·ère avec le RGPD (67 % de l'ensemble des répondant·es).

Figure 10 : Pratique et connaissance de la règlementation



2.3 Pratiques de stockage et d'archivage des données

Plusieurs offres de services sont actuellement disponibles au laboratoire pour le stockage et le partage de leurs données au niveau national ou local : Huma-Num depuis 2013 (dédié aux équipes de recherche) ; MyCORE (CNRS) avec 100 Go dédié et AMUbox avec 60 Go (depuis l'espace numérique de travail de l'Université d'Aix-Marseille).

Depuis septembre 2016, le laboratoire propose à ses membres un espace dédié modulable selon les pratiques et besoins des chercheur·es et des équipes (LESTbox). Depuis 2019, tout nouveau poste d'ordinateur commandé est systématiquement chiffré et ce travail s'opère également sur les anciens postes. Au total 87% d'ordinateurs détenus par les permanent·es (hors doctorant·es, contractuel·les et associé·es) sont actuellement chiffrés.

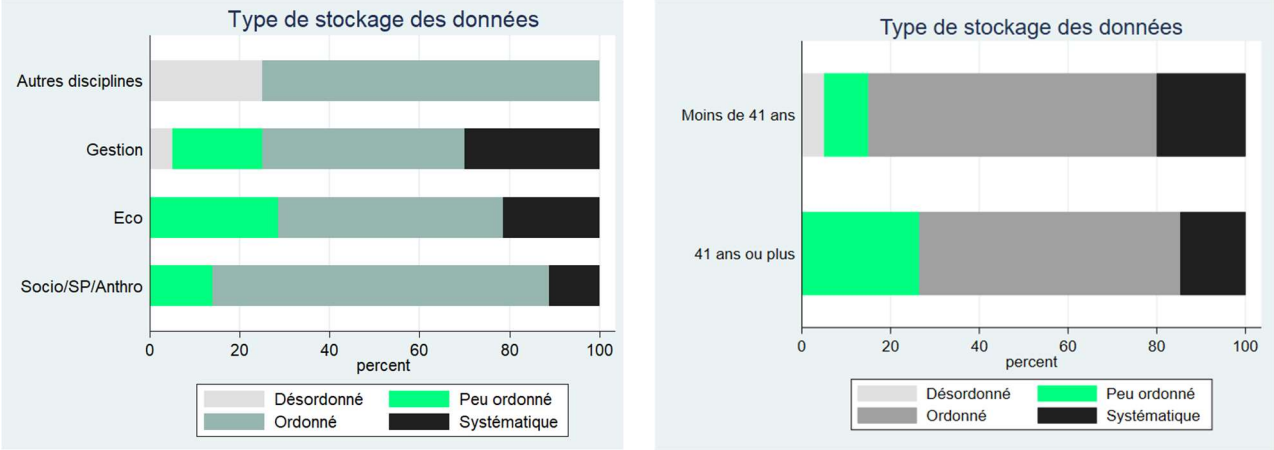
Qu'en est-il alors de la portée de ces offres dans les pratiques de recherche des chercheur·es ? Comment s'emparent-elles·ils de ces dispositifs pour stocker, archiver et sécuriser leurs données ? Ont-elles·ils une démarche volontaire et régulière de sauvegarde (politique personnelle d'archivage avec choix de la fréquence et choix des supports) ? Leurs données sont-elles dispersées sur de multiples supports ou stockées essentiellement sur un seul support ?

2.3.1 Stockage des données

Les données collectées ou créées par les enquêté·es sont majoritairement numériques (75%) avec une faible part (7%) pour les données nativement numériques (DNN)[4]. 25% de répondant·es (18 personnes) travaillent principalement sur des données non numériques.

80% déclarent stocker leurs données de manière ordonnée ou systématique. Ceci permet à 91% de les retrouver facilement (seulement 9 % mentionnent ne pas retrouver facilement leurs données). On remarque dans les graphiques ci-dessous que c'est en gestion que le stockage des données est plus souvent déclaré comme « systématique », et en économie que le stockage est plus souvent estimé comme « peu ordonné » par rapport aux autres disciplines (Figure 11). De même, celles et ceux qui ont 41 ans ou moins affichent plus souvent une pratique ordonnée et systématique que celles et ceux qui ont 41 ans et plus.

Figure 11 : Type de stockage



Une part non négligeable des enquêté·es réalisent des sauvegardes régulières sur des intervalles d'un mois voire moins (44%). La sauvegarde en fonction des besoins concerne un peu moins d'un tiers des enquêté·es (32%) et peut présenter certains risques pour des données intermédiaires. Une part beaucoup plus faible (7%) ne fait jamais de sauvegarde ou seulement une fois par an.

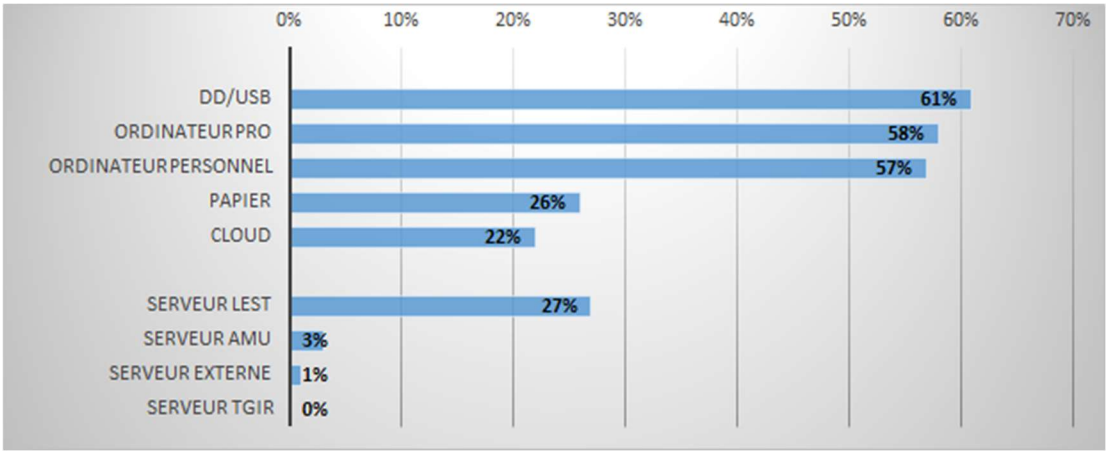
2.3.2 Archivage des données

La **démarche d'archivage est largement individuelle** : dans 85% des cas, la·le chercheur·e archive lui.elle-même ses données ; dans 6% des cas, il s'agit d'une personne dans un projet (*désignée ou non par l'équipe selon son niveau de responsabilité dans le projet*) ; 7 % déclarent ne pas avoir de responsable identifié. Malgré l'existence d'outils d'archivage sur des serveurs, 27 % déclarent avoir des difficultés pour archiver leurs données (29 % des chercheur·es et 21 % des doctorant·es). Dans la majorité des cas, **les données sont stockées sur des supports individuels mobiles ou virtuel** (ordinateur portable, disque dur, USB, cloud privé). L'archivage n'est pas pensé *a priori* dans une stratégie de moyen-long terme au regard des supports à risques utilisés. Les supports de stockage les moins utilisés sont ceux que l'on pourrait qualifier

[4] DNN : Données dématérialisées dès leur création, créées directement sur des supports numériques (*Ex : enregistrements audios numériques, données du web tels les tweets*.)

de supports institutionnels : 99 % des enquêté·es n'ont jamais utilisé le serveur TGIR, 94 % n'ont jamais utilisé un serveur extérieur d'une autre institution (Figure 12). Il est fort probable que ces serveurs ne soient pas identifiés comme supports possibles ou, tout simplement, qu'ils soient méconnus. Les serveurs du laboratoire et d'AMU sont également rarement utilisés bien qu'il existe des outils pour assurer une sauvegarde régulière et automatique : 90 % n'ont jamais utilisé le serveur de la tutelle et 61 % n'ont jamais utilisé le serveur du laboratoire (LESTbox).

Figure 12 : Type d'archivage selon support



Selon le statut, en distinguant les chercheur·es (tous statuts confondus) et les doctorant·es, quelques différences de pratiques d'archivage apparaissent. **Les archivages sécurisés sont moins utilisés par les doctorant·es que par les chercheur·es**. Par exemple, 79 % des doctorant·es n'ont jamais archivé leurs données sur le serveur du LEST. Cela concerne 56 % des chercheur·es. Pour les doctorant·es, les archivages principaux sont l'ordinateur personnel et le disque dur/la clé USB, qui sont tous deux des archivages non sécurisés. Pour les chercheur·es, on retrouve également ce type d'archivage comme archivage principal, auquel s'ajoute l'ordinateur professionnel et en dernière position le serveur du LEST.

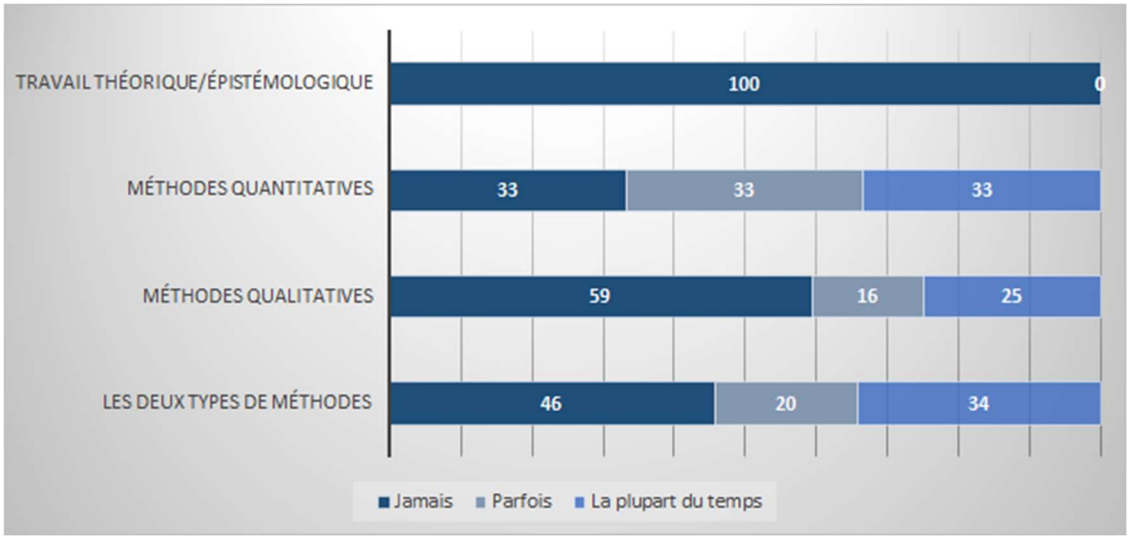
Tableau 6 : Type d'archivage selon statut

	EC et C	Doctos
Papier		
Jamais	27	37
Parfois	45	37
La plupart du temps	27	26
Ordinateur perso		
Jamais	36	5
Parfois	18	5
La plupart du temps	45	89
Ordinateur pro		
Jamais	11	79
Parfois	13	16
La plupart du temps	76	5
DD/USB		
Jamais	11	0
Parfois	25	42
La plupart du temps	64	58
Cloud		
Jamais	55	47
Parfois	25	21
La plupart du temps	20	32

	EC et C	Doctos
Serveur LEST		
Jamais	56	79
Parfois	13	5
La plupart du temps	31	16
Serveur AMU		
Jamais	87	95
Parfois	9	5
La plupart du temps	4	0
Serveur externe (autre instit.)		
Jamais	91	100
Parfois	7	0
La plupart du temps	2	0
Serveur TGIR		
Jamais	98	100
Parfois	2	0
La plupart du temps	0	0

L'archivage sécurisé des données défini ici comme un archivage sur serveur institutionnel (LEST, AMU, TGIR, ou serveur externe) **n'est pas une pratique courante** puisque seulement 29 % des enquêté·es archivent leurs données sur des supports sécurisés. Cette pratique d'archivage sécurisé varie selon le type de méthodes utilisées, et donc selon le type de données. Celles et ceux qui utilisent uniquement des données quantitatives font plus fréquemment un archivage sécurisé de leurs données (Figure 13). Ceci est sans aucun doute lié aux exigences concernant certains types de grandes enquêtes. Toutefois on peut s'étonner que les données qualitatives, dont la collecte est souvent le fruit d'un long travail difficile à répliquer en cas de perte, sont plus rarement archivées de manière sécurisée : près de 60 % de celles et ceux qui travaillent uniquement avec des méthodes qualitatives n'archivent jamais leurs données sur des supports sécurisés, un peu moins de la moitié lorsque les deux types de méthodes sont utilisées et le tiers lorsque seules les méthodes quantitatives entrent en jeu.

Figure 13 : Archivage sécurisé des données selon la méthode privilégiée



2.4 Pratique de partage et de diffusion des données de recherche

2.4.1 Pratique de partage de données

Trois critères viennent éclairer les pratiques d'accès à leurs données (collectées ou traitées) :

- Le partage de leurs données à d'autres chercheur·es
- Le stockage en libre accès
- La réutilisation des données d'autres chercheur·es

Un peu plus de la moitié des répondants déclarent pratiquer le partage des données (47 % ne pratiquent pas de partage). **Celui-ci a lieu majoritairement dans le cadre d'un projet de recherche, le plus souvent avec des collègues du même laboratoire** et dans une moindre proportion avec des collègues extérieur es ou hors projet.

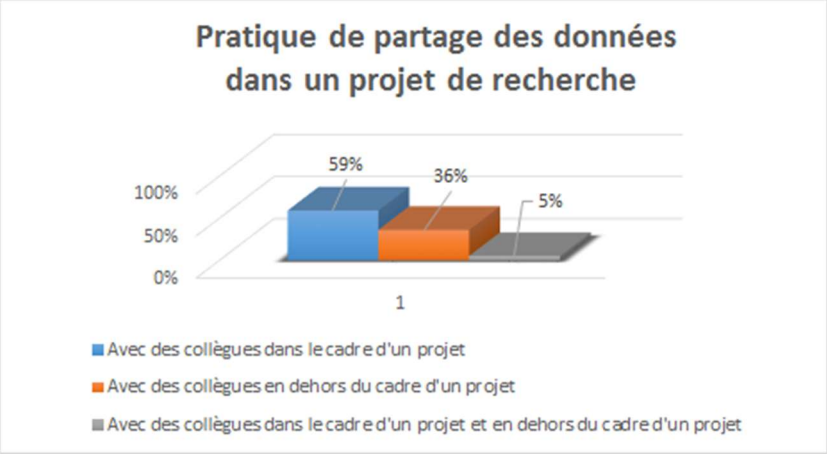
Celles et ceux qui le pratiquent diffusent en **accès réservé.** Les données partagées sont très rarement mises en ligne en libre accès. Sur 40 personnes déclarant partager leurs données : une seule personne déclare les avoir mis en ligne « *plus d'une fois* » et 4 personnes « *une seule fois* ».

Le partage de données est pratiqué plus largement par celles et ceux qui utilisent uniquement les méthodes qualitatives de recherche. Concernant la démarche quantitative, le partage est plus difficile étant donné que les données utilisées sont la plupart du temps des données statistiques soumises à un certain régime de protection et adossées à un projet de recherche déterminé.

Tableau 7 : Pratique de partage

		%	Personnes
OUI		**53%**	**40**
	dont avec des collègues du LEST (*en général*)	55%	22
	dont avec des collègues extérieur.es (*en général)*	45%	18
NON		**47%**	**35**

Figure 14 : Pratique de partage (Projet)



La réutilisation des données collectées par des collègues est peu ancrée dans leurs pratiques de recherche. Néanmoins, celle-ci ne rencontre pas d'opposition réelle. Un pourcentage non négligeable l'a déjà expérimenté « *au moins une fois* » à défaut de l'intégrer plus systématiquement.

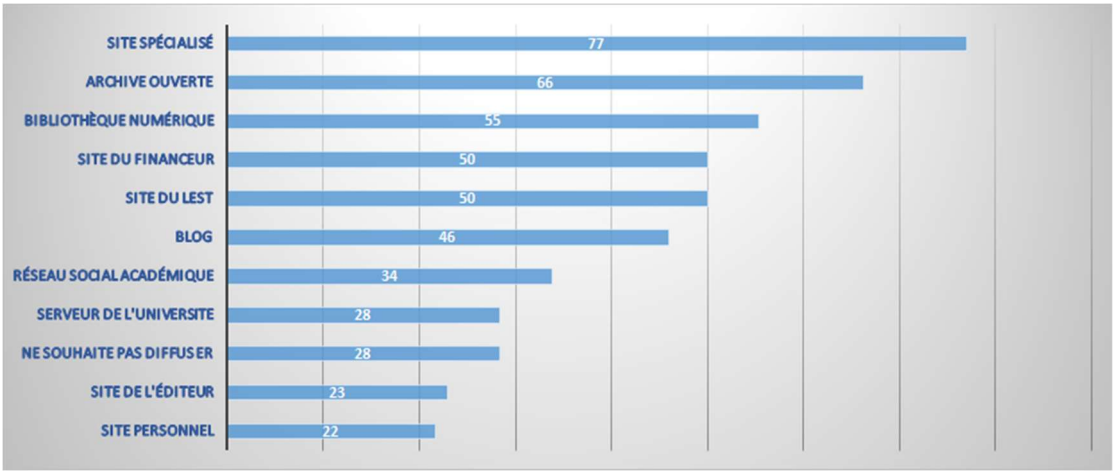
Tableau 8 : Pratique de réutilisation

	%
Non, et je ne le ferai pas	3,9
Non, pas pour le moment	71,43
Oui, au moins une fois	19,48
Oui, régulièrement	5,19

2.4.2 Position sur le libre accès

72% des personnes enquêtées sont favorables à la diffusion de leurs données de recherche. Les « sites spécialisés » pour l'archivage et la diffusion de données de la recherche ainsi que les « archives ouvertes » de type HAL rencontrent le plus d'approbation. Le serveur de l'université arrive bien après le serveur du LEST et le site de l'organisme financeur. Le blog est également assez largement approuvé (Figure 15). Dans l'ensemble des réponses, les lieux de diffusion les moins plébiscités sont les sites des éditeurs et le site personnel.

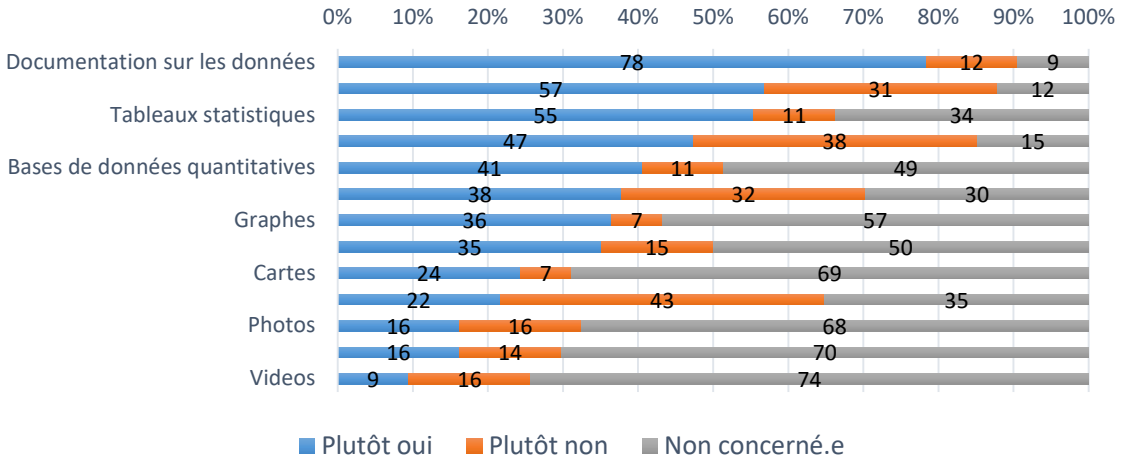
Figure 15 : Lieux de diffusion qui seraient privilégiés



La majorité des enquêté·es privilégierait en priorité leur documentation sur les données, leurs notes et synthèses et leurs tableaux statistiques. Les transcriptions arrivent après. Cela peut s'expliquer aisément par la nature des données *(sensibles*) et des freins précédemment évoqués. Un tiers n'est pas très favorable au dépôt des fichiers d'exécution indispensables pour valider les résultats. Ce constat interpelle au regard de leur position générale plutôt favorable à la valeur de preuve.

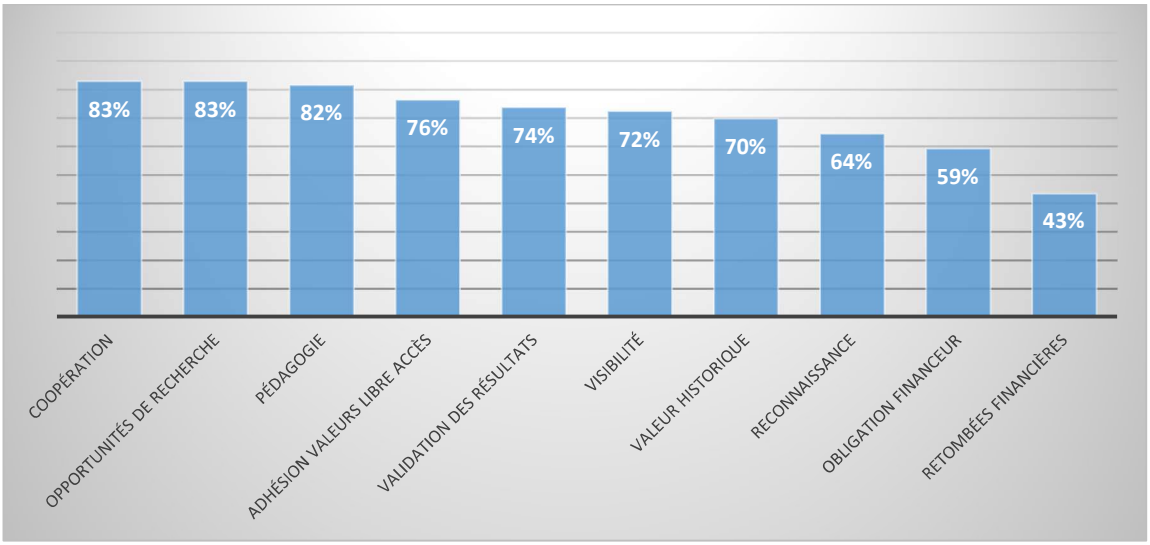
Par ailleurs, une forte proportion de chercheur·es qui pratiquent autant les méthodes quantitatives que les deux méthodes seraient plutôt favorables pour mettre en ligne leurs bases de données quantitatives quel que soit le niveau d'effort de travail fournit sur les données.

Figure 16 : Types de données qui seraient déposées



L'occasion de **nouvelles coopérations scientifiques, le souhait de voir se développer d'autres recherches à partir des données ou encore qu'elles soient utilisées à des fins pédagogiques** sont les **trois incitations plébiscitées par la grande majorité des chercheur·es au LEST**. Elles recouvrent des dimensions sociales et pédagogiques de la recherche. Celles-ci sont suivies de près par une adhésion aux valeurs du libre accès, à la valeur de la preuve associée à la volonté de rendre visible les travaux de recherche réalisés à partir des données, et à la valeur historique. Les retombées financières arrivent sans surprise en dernière position.

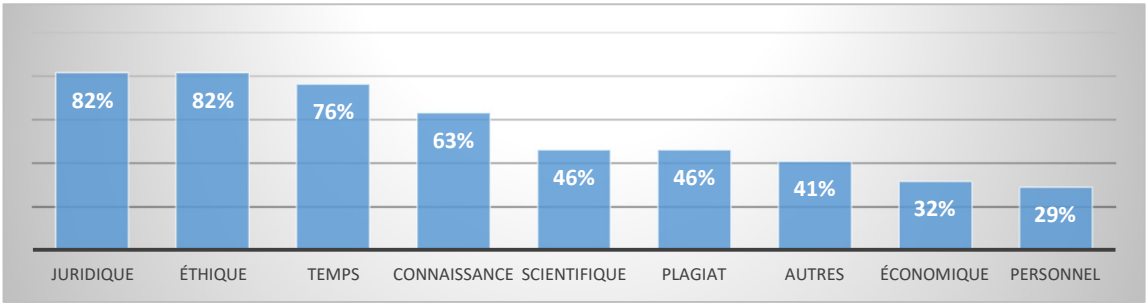
Figure 17 : Les facteurs d'incitations à la mise en libre accès des données



Les trois freins à la mise en libre accès des données qui se dégagent clairement sont les **freins juridiques et éthiques et le manque de temps**. Le manque de connaissance dans la description des jeux de données et sur les modalités de dépôts suit également juste après. Ce qui est intéressant ici c'est que malgré l'adhésion de plus des trois-quarts des chercheur·es aux valeurs du libre accès (cf. ci-dessus), le manque de connaissances dans la description des jeux de données ou dans les modalités de dépôt arrive en quatrième position en ce qui concerne les freins au libre accès des données.

Les freins éthiques sont particulièrement cités par les chercheur.es qui utilisent les méthodes qualitatives (90 %) alors que ce sont les **freins juridiques qui sont cités par l'ensemble de celles et ceux qui utilisent uniquement les méthodes quantitatives**, ce qui s'explique là aussi par les règles particulières qui régissent l'utilisation de ces données la plupart du temps collectées et produites par des institutions externes. La totalité des chercheur·es qui travaillent avec des méthodes quantitatives considèrent le fait de pouvoir faire valider leurs résultats par d'autres chercheur·es comme quelque chose d'incitatif. En revanche, celles et ceux qui travaillent uniquement à partir de méthodes qualitatives représentent en comparaison 25 %.

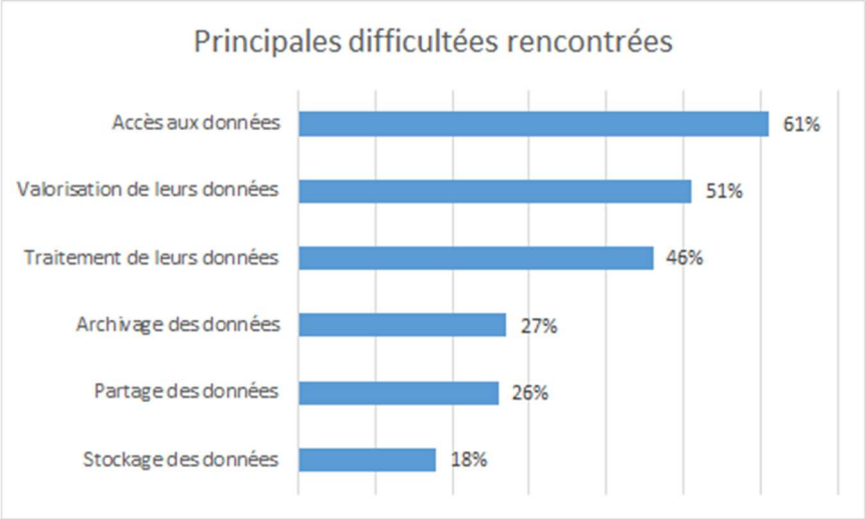
Figure 18 : Les facteurs de freins à la mise en libre accès des données



2.5 Difficultés rencontrées

Les principales difficultés se situent essentiellement au niveau de l'accès aux données, leur valorisation et leur traitement.

Figure 19 : Difficultés rencontrées avec les données



2.6 Besoins et attentes

Dans l'ensemble, c'est avec les méthodes de gestion des données (Plan de gestion des données) qu'il y a le moins de familiarité, viennent ensuite les débats déontologiques sur l'usage de différents types de données, suivis des questions liées au libre accès. On constate ici, que les différences entre les deux populations sont faibles voire inexistantes ce qui peut surprendre au regard de l'expérience des chercheur·es.

Tableau 9 : Etat des connaissances

	EC & C	Doctorant·es
Enjeux du libre accès		
Familier·ère	27	32
Peu familier·ère	73	68
Les méthodes de gestion des données		
Familier·ère	7	5
Peu familier·ère	93	95
Les débats déontologiques		
Familier·ère	20	21
Peu familier·ère	80	79

C'est sur les aspects liés à l'accès et au partage des données ou encore sur des questions juridiques et de documentation des données que les chercheur·es souhaiteraient bénéficier d'aide en priorité. Le besoin dans la gestion des données (PGD/DMP) et l'aide déontologique sont positionnés de façon plus secondaire. Pour les doctorant·es, tous les aspects semblent en revanche avoir de l'importance. Enfin, un quart des répondant·es seulement affirme disposer d'un corpus qui pourrait être valorisé et parmi eux, 8 personnes seraient en capacité de mobiliser du temps dans cet objectif.

Tableau 10 : Etat des besoins

	EC & C	Doctorant·es
Aide en méthodologie de la doc.		
En priorité	22	21
Eventuellement	58	68
Non	20	11
Aide pour l'accès et le partage des données		
En priorité	27	32
Eventuellement	55	58
Non	18	11
Aide juridique		
En priorité	22	32
Eventuellement	58	68
Non	20	0
Aide sur la déontologie		
En priorité	11	32
Eventuellement	58	68
Non	31	0
Aide sur le plan de gestion des données (PGD/DMP)		
En priorité	18	32
Eventuellement	47	58
Non	35	11

3 ENSEIGNEMENTS DE LA PHASE 1

Avec une bonne représentativité de la population au laboratoire qui produisent et traitent des données de recherche, cette enquête rend compte des grandes tendances sur les pratiques et besoins dans ce domaine. Au-delà de leur diversité, les groupes de données qui se distinguent sont propres aux sciences sociales, à leurs disciplines et spécificités. Les recherches en cours du laboratoire, aux croisements de disciplines SHS, ne se traduisent pas par une volumétrie importante de données (big data). Les données lourdes (films) sont plutôt assez rares et repos(ai)ent sur quelques individus par le passé. En revanche, on peut présager naturellement, d'un accroissement du volume des données dans le temps lié aux effectifs croissants du laboratoire, à l'arrivée de nouvelles disciplines, aux usages numériques, à l'essor de la communication événementielle produite par le laboratoire (films et audios des séminaires et journées scientifiques).

Dans l'ensemble, les pratiques courantes de recherche sont plutôt respectueuses de leurs enquêté·es (éthique scientifique portée par les chercheur·es) qui se traduisent en pratique par l'information aux usagers, ou encore par l'utilisation récente de la salle dédiée à la protection des données soumises au secret statistiques (données plus rares au laboratoire) et mise en place en 2019. L'enquête montre un intérêt certain pour les enjeux actuels relatifs aux données de recherche mais souligne une réserve dans les aides attendues pour améliorer les pratiques dans ce domaine. On note que les répondant·es s'inscrivent de façon insuffisante dans le cadre posé par la nouvelle réglementation en matière de protection des données et des personnes ou entités sur lesquelles elles portent.

Il n'y a pas de réelle opposition vis-à-vis de la réutilisation des données de recherche ou du partage et de leur diffusion même si les pratiques ne sont pas courantes. On peut faire un ensemble de premiers constats qui permettent de pointer sur des zones de fragilités, des zones blanches *(à explorer notamment par le biais de focus groups*), ou des zones à risques pour le laboratoire et ses membres (*en lien avec le RGPD ou la gestion d'un « patrimoine institutionnel de la données ») :*

Une vulnérabilité des données dans la phase de leurs traitements et de leur stockage/archivage *(ex : stockage sur des supports individuels mobiles, fréquences des sauvegardes non immédiates et automatisées, le chiffrement des postes limités aux permanents hors doctorants, associés et contractuels).*

Une conformité limitée avec la réglementation en vigueur (déclarations CNIL rares, pas de suppression des données personnelles après traitement, …)

Une documentation sur les données insuffisante à un niveau technique et standardisé *(ex : structuration d'organisation des fichiers de données, gestion des versions, droits appliqués à leurs données sources et produites, formats des logiciels (propriétaire, libre) et volumétrie…)* pour permettre le montage d'un plan de gestion de données ou encore la valorisation et la réutilisation de leurs données pour d'autres recherches.

Un manque de connaissances sur les débats déontologiques et les enjeux du libre accès sur ces données même si ces questions n'apparaissent pas comme étant prioritaires dans les besoins exprimés. Les besoins sont exprimés en particulier sur l'accès aux données et sur leur partage, sur les méthodologies de gestion des données *(moins sur le montage des plans de gestion de données demandés par l'ANR notamment*), et sur les aspects juridiques complexes des données*.*

Un manque de connaissances et d'utilisation des infrastructures et services à la donnée proposés en externe ou en interne[5], et ce malgré quelques actions d'information et de communication scientifique déjà mises en place (*atelier de présentation de la PUD AMU PROGEDO, ateliers proposés par la MMSH, communication et services INF et IST du LEST, etc.*).

[5] LEST-Box, salle quanti fermée, coffre-fort, 10 ordinateurs portables chiffrés mis à disposition du personnel non permanent.

II. FOCUS GROUPS

L'enquête par *focus groups* a été conduite en novembre-décembre 2020. Les groupes ont été constitués sur la base du volontariat et de la disponibilité à partir d'une relative homogénéité de pratiques, afin de simplifier les discussions. Tous·tes étaient volontaires pour se prêter à ces échanges avec la limite qu'impose cet exercice : les propos recueillis ne peuvent être totalement représentatifs de l'ensemble des scientifiques de ce laboratoire et nécessitera une discussion en interne plus étendue. Trois groupes ont ainsi été réunis : l'un regroupant des chercheur·es et enseignant·es-chercheur·es pratiquant l'enquête quantitative, l'autre — dans la même catégorie — pratiquant davantage des enquêtes par observations, entretiens, films, etc. ; le dernier regroupant les doctorant·es, dont l'enquête par questionnaire avait manifesté des enjeux d'apprentissage différents. L'objectif était, à partir de leur pratique, de mieux cerner les questions qu'elles·ils se posent ou les difficultés qu'elles·ils rencontrent dans la gestion des données sur les plans organisationnels, techniques, éthiques, épistémologiques, réglementaires afin de mettre en place à terme une politique de gestion au niveau du laboratoire qui puisse mieux les accompagner. La conduite de ces *focus groups* a permis de constater, entre la phase 1 et la phase 2, une appropriation en cours du nouveau cadre réglementaire par les chercheur·es et enseignant·es chercheur·es et pour les jeunes doctorant·es *via* les formations rendues obligatoires par les ED.[6]

4 FOCUS GROUP RECHERCHE QUANTITATIVE

Six enseignant.es-chercheur.es appartenant aux disciplines de l'économie, de la sociologie et de la gestion ont participé au *focus group* le 30 novembre 2020 (en distanciel).

4.1 Sur les conditions d'accès et d'exploitation des données

La plupart des participant·es travaillent sur des données d'enquêtes de statistiques publiques ou privées de type plutôt micro-données (données individuelles, d'entreprises, de salariés, …). L'accès aux données se fait auprès de diverses sources. Leurs principales sources publiques sont issues des grandes enquêtes nationales accessibles depuis la plateforme PROGEDO-QUETELET, ou d'autres enquêtes d'organismes publics (Ministères, OCDE, Eurostats, CEREQ, …). Les données privées proviennent soit de statistiques issues de bases de données commercialisées (*ex : ORBIS annuaire d'entreprise, IFS – statistiques financières accessibles via Aix-Marseille Université*), soit d'enquêtes par questionnaires menées à plus ou moins grande échelle. Cette population reconnait avoir à sa disposition une offre extrêmement riche, mais malgré cela, elle rencontre des difficultés. Qu'elles en sont les raisons ?

- **Les données d'entreprises commercialisées par des sociétés privées**

Les licences d'exploitation de ces bases sont accordées souvent moyennant l'achat d'une licence d'utilisation. En l'absence d'accords négociés entre les tutelles et les fournisseurs, ces chercheur·es doivent trouver d'autres solutions pour démarrer leurs recherches ou finaliser une publication. Les tests gratuits offerts par certains fournisseurs ne sont pas systématiques et adaptés à leurs besoins immédiats. Les négociations engagées, parfois non prévues dans le montage initial du projet, peuvent s'avérer longs et décevants au regard des données réelles contenues dans la base.

[6]Les doctorant·es ont l'obligation de suivre deux formations : une sur l'éthique de la recherche, l'autre sur l'intégrité scientifique dans leur cursus de thèse depuis 2017, et dès la 1ère année depuis 2020 (prérequis pour l'autorisation de soutenir)

- **Les données issues d'enquêtes quantitatives menées dans les entreprises**

Ces données collectées engagent un investissement non négligeable car ces scientifiques doivent composer avec/sur leurs terrains (entreprises, individus), élaborer le questionnaire (en tenant compte des freins et résistances), expliciter, relancer jusqu'à l'obtention d'un taux de réponses satisfaisant pour leurs enquêtes.

- **Un coût d'entrée pour exploiter les grandes enquêtes publiques**

Pour pouvoir exploiter les grandes bases de données, ces chercheur·es ont besoin d'une documentation très détaillée pour s'approprier la structure et les variables intégrées, en comprendre les logiques et limites ou à défaut pourvoir disposer d'un interlocuteur pour leur garantir une bonne lecture et usage des données. Plus la base est volumineuse, complexe et la documentation insuffisante, plus l'investissement en temps sera important pour eux. Le besoin d'accéder aux programmes informatiques est réel, puisqu'ils permettraient une exploitation directe et plus rapide des données.

- **Un coût élevé pour accéder aux données du CASD**

La plupart des chercheur·es qui travaillent avec des méthodes quantitatives utilisent des données d'enquêtes assez facilement disponibles et exploitables sur le plan statistique directement depuis leur ordinateur portable. Ce n'est pas le cas de quelques-un.es travaillant sur des données publiques sensibles soumises au secret statistique. L'accès direct à ces données comme ce fut le cas par le passé n'est plus possible. Il s'opère désormais par le biais d'un terminal (la SD-box) avec une contrainte biométrique. Le traitement des données ne peut se faire qu'au laboratoire où se trouve la box. Néanmoins, le point le plus préoccupant est celui du coût très élevé de ce dispositif lié à la mise à disposition et à l'exploitation des données sous licence (*coût de la technologie d'accès*). Il faut donc trouver des financements, au sein du LEST ou à l'extérieur. L'accès aux données est délimité dans le temps — un temps qui n'est pas compatible avec celui de la recherche. En effet, un·e chercheur·e qui a soumis une publication sur la base des données traitées avec la box, n'aura plus accès à ces données pendant les révisions de son article (*clôture entretemps de l'exercice budgétaire du projet*). La plateforme universitaire des données de l'université d'Aix-Marseille (PUD AMU) propose à l'ensemble de la communauté scientifique (chercheur·es et doctorant·es) une SD-box dans les locaux de la MMSH mais les coûts restent majoritairement à leur charge (*seul le point d'accès est proposé dans le cadre de cette expérimentation*). Même si beaucoup d'enquêtes ne tombent pas encore dans ce dispositif du CASD, certain·es pourraient se sentir exclu·es de l'exploitation de ce type de données du fait d'un manque de financement. On constate paradoxalement une asymétrie forte entre ces producteurs de données qui reçoivent des injonctions fortes à les partager en *open science*, et doivent trouver en parallèle des financements pour accéder aux données brutes.

4.2 Quelques questionnements autour du partage et de la valorisation

- **Entre données publiques et données du chercheur·e : *Quels matériaux partager ?***

Sur le principe, les participant·es au *focus group* sont plutôt ouvert·es à cette démarche. Mais tous les matériaux ne sont pas concernés ou partageables. Certain·es travaillent sur des données issues de sources publiques qui sont déjà librement accessibles à la communauté scientifique. Pour celles et ceux, qui produisent leurs propres enquêtes, il apparait par ailleurs contreproductif de partager sur des plateformes ouvertes ce matériau central pour leurs recherches. Les matériaux qui semblent plus facilement partageables en externe sont : les données produites par la·le scientifique à partir de sa base des données, le guide des variables

et le programme logiciel associé aux traitements de leurs données ayant permis de créer ces variables.

- **Autour de l'aspect juridique et de la mise en conformité avec le RGPD : *Comment faire pour être en règle et à moindre coût ?***

On constate un réel désir de se mettre en règle avec le RGPD. Des démarches d'autoformation ont déjà été entreprises chez certain·es mais cela a nécessité beaucoup de temps et d'énergie. Pour l'ensemble, pouvoir mettre en application le RGPD dans leur quotidien de travail à partir de situations concrètes (recherche ou enseignement) n'est pas évident par manque de connaissances et de temps pour acquérir ces connaissances.

- **Autour de la publication des données : *Un dilemme - comment répondre aux exigences de publication de leurs tutelles et des éditeurs tout en protégeant les données ?***

L'injonction de publier dans des revues anglosaxonnes de qualité est très forte dans les disciplines des participant·es. Leur priorité se centre donc sur la rédaction et la valorisation de leurs articles et non sur la mise à disposition de leurs matériaux. Cette mise à disposition nécessiterait un travail de préparation et de dépôt impossible actuellement dans la gestion de leur temps à flux tendu. Depuis peu, certains éditeurs demandent les données qui ont conduit à la production de leurs résultats, une question qui est en train de devenir centrale pour publier. Certain·es ont conscience de ces enjeux liés à la reproductibilité de la science. En revanche, tous·tes méconnaissent ce à quoi ils s'engagent avec ces éditeurs en matière de droit d'auteurs.

4.3 Outils de stockage et partage

Au quotidien, ces scientifiques ont l'habitude de travailler sur leurs données depuis leur ordinateur portable qui les suit dans tous leurs déplacements (sauf cas des données *via* le CASD). Ils conservent leurs données sources à part et travaillent de façon journalière sur la base qu'ils ont élaborée et enrichie. Ils ont besoin d'y avoir accès facilement et ont donc tendance à multiplier les sauvegardes à plusieurs endroits, à chaque nouveau traitement, au risque de saturer leur disque dur, d'autant plus si leur base est volumineuse. Le problème de saturation se pose également pour le travail des données qualitatives, qui imposent un stockage de photographies de terrains et de vidéos qui ne sont pas compressées.

Plusieurs participant·es travaillent avec une Dropbox personnelle — seul outil qui leur permet de tout faire ; avoir accès facilement à leurs données pour stocker leurs petites et grosses bases de données alimentées par leur soin au quotidien, leurs sources secondaires, etc. L'outil est gratuit avec une limite de stockage de 2 Go (2 000 Mo). Il nécessite un abonnement payant en cas de besoin de stockage plus important. Une partie de ces chercheur·es peut partager ses données dans une box institutionnelle manuellement ou par synchronisation automatique. L'outil Owncloud est utilisé par certain·es pour synchroniser leurs données avec l'AMUbox ou la LESTbox. Il y a une certaine réticence à utiliser la box de l'Université en raison de pertes de données constatées par le passé ou de difficultés actuelles à organiser son contenu à des fins de partage. Les nouveaux·elles arrivant·es ont du mal à se retrouver dans l'environnement local face à la diversité des offres institutionnelles pour le partage ou la sauvegarde de leurs données : des clouds différents dans leurs prestations et leurs ergonomies. La·le chercheur·e peut se retrouver seul.e à tester cette offre quand il arrive. La LESTbox est perçue essentiellement comme un espace de partage collaboratif. Mieux connaitre les fonctionnalités de chacune avec leurs avantages, leurs restrictions ou limites les aideraient dans cette appropriation. Néanmoins, une limite à cet usage s'observe, qui n'est pas du fait de ces scientifiques eux-mêmes : elles·ils ne se retrouvent pas forcément sur ces outils institutionnels

quand elles·ils travaillent sur des projets avec des partenaires différents qui ont eux-mêmes leurs propres outils institutionnels ou leurs propres méthodes de travail.

5 FOCUS GROUP ENQUÊTES QUALITATIVES

Ce groupe a été mis en place le 1er décembre 2020 en distanciel. Sept enseignant·es-chercheur·es appartenant majoritairement à la sociologie et à la gestion appliquant des méthodes qualitatives se sont retrouvé·es à cette occasion. *(1 CNRS, 5 AMU, 1 postdoctorant, 1 contractuel-associé ; 3 en sociologie, 3 en gestion, 1 dans une autre discipline).*

Beaucoup travaillent avec des données personnelles et sensibles collectées au cours d'entretiens ou d'observations qu'elles ils doivent sécuriser eux-mêmes. Les points saillants de ce groupe sont liés à des questionnements forts autour de la mise en conformité avec le RGPD et des difficultés épistémologiques et de faisabilité d'enquêtes qu'elle induit. Pour l'ensemble de ces participant·es, tous·tes se sentent concerné·es par les problématiques de gestion des données. Elles ils ont conscience de ne pas forcément être toujours en règle avec la réglementation, reconnaissent manquer en effet de connaissances et de temps, ceci malgré les efforts de certain·es pour tenter de clarifier les règles à appliquer dans leur quotidien. Elles ils essaient de faire au mieux avec leurs connaissances et leurs contraintes techniques et sont enfin très volontaires pour faire évoluer leurs pratiques et mieux l'inscrire dans leur nouvel environnement de recherche.

5.1 Premiers constats : stockage, sauvegarde et partage/réutilisation

- **Accès aux données**

La difficulté d'accéder aux données qualitatives pour certains d'entre eux peut s'expliquer par la technique d'enquête sur le terrain, le type de terrain et l'expérience du chercheur·e. Les doctorant·es ou les jeunes chercheur·es ont souvent plus de difficultés à accéder à un terrain nouveau que les chercheur·es plus familier·es des méthodes d'approches et en parallèle bien inséré·es dans leurs réseaux professionnels. Certains d'entre eux déclarent même manquer de temps pour exploiter l'ensemble des matériaux extrêmement riches qu'elles·ils ont à disposition notamment dans des espaces partagés réservés.

- **Stockage et sauvegarde individuelle : des supports à risques (mobiles et/ou privés…)**

Au quotidien, les données sont recueillies, stockées et conservées en amont sur le disque dur de leur portable professionnel ou personnel. Le portable personnel plus à risque concerne en particulier les non titulaires (post-doctorant·es, …), les contractuel·les indépendant·es, les nouveaux·elles arrivant·es (membres permanent·es) peu familier·ères des usages/outils ou certain·es chercheur·es qui ont gardé une méfiance en l'égard des serveurs institutionnels à la suite d'une expérience antérieure de perte de données. Des sauvegardes à leur initiative sont assurées sur des supports privés (disques durs, disques externes, clés USB) à des fréquences variables en mode manuel, ou sur un cloud privé en mode manuel (*ex : grosse sauvegarde manuelle tous les 15 jours*) ou synchronisé. Certaines données de projet sont partagées sur des clouds collectifs distincts.

Les données sont ainsi disséminées avec une cohérence d'ensemble de stockage très relative. La perte de données a déjà été constatée par certains membres ; des difficultés d'accès et de lisibilité en cas de réutilisation (lecture d'anciens formats, oubli de mot de passe, …) pourrait également poser problème même s'il n'y a pas eu de cas évoqués dans ce *focus group*.

- **Partage des données : deux méthodes utilisées**

La gestion par projet ne permet de systématiser un mode d'organisation applicable à tous les projets car un projet réunit plusieurs partenaires qui peuvent être très différents. Le partage des données et des supports de stockage va varier selon les membres, leurs pratiques/habitudes, sensibilité et leurs appartenances institutionnelles (secteur public/privé; outils à disposition). L'appui à des projets de l'équipe « soutien à la recherche » du LEST se fait en fonction des besoins et des moyens techniques et humains dont disposent les collaborateurs trices du projet. Tous les documents ne sont pas partagés. Pour les données qui sont partagées, on observe :

- 1ère méthode : Certaines données du projet sont recueillies, stockées sur une infrastructure de recherche mais ne sont pas conservées en local. L'infrastructure sert de dépôt central en amont du traitement et garantit la conservation sécurisée à long terme. Cela peut concerner des projets qui travaillent sur des matériaux primaires volumineux très gourmands en espaces de stockage (ex : vidéos). Ceux-ci sont déposés en amont du traitement/analyse sur une plateforme numérique de type Huma-Num générique ou spécialisé. Chaque chercheur·e au cas par cas (dans le temps, plusieurs fois) télécharge le matériel dont elle·il a besoin pour travailler en local. Ces matériaux n'ont pas vocation à être conservés longtemps sur leur ordinateur pour éviter les risques de saturation en local. Le reste des données produites après traitement peut être partagé ou non sur la plateforme selon leurs collaborations.

- 2ème méthode : Toutes les données du projet ou du moins une partie, prédéfinie de façon consensuelle, sont recueillies et stockées au fil du projet : *soit sur un serveur institutionnel (en parallèle des données propres aux membres sur leurs ordinateurs) : deux personnes dans ce *focus group* ont utilisé la LESTbox pour des projets et en sont très satisfaites ; une autre personne a travaillé sur un projet à titre collaboratif avec un ingénieur recruté à cette occasion sur financement projet (*LEST non porteur du projet*) *soit/ou en complément, sur des clouds *(documents finalisés partagés)* et des drives privés (*pour les documents corédigés)* pour des raisons de facilité de pratique ou de gestion d'urgence (culture Dropbox et Google), et selon les cas, en deuxième sauvegarde sur un cloud hébergé par le partenaire (ex : Nextcloud en open source du partenaire privé).

Dans certains projets, des modes d'organisation et des règles de gestion se mettent en place au démarrage en bonne collégialité (ex : *choix du cloud, du drive, type de documents à partager, etc.*). L'alimentation des espaces communs peut revenir à une seule personne ou à l'ensemble de l'équipe. Les accès peuvent être ouverts à tous·tes ou selon certains principes qui varient selon les projets (à titre d'exemple : un partage exclusivement soit des matériaux bruts selon qui a participé à la collecte soit des analyses monographiques). Le choix du cloud dépend des outils mis à leur disposition, de la connaissance qu'elles·ils en ont et des modalités d'accès.
D'autres projets sont moins bien structurés en raison des pratiques et des appartenances institutionnelles plus hétérogènes de leurs membres. Leur choix de support de stockage correspond alors davantage à un besoin immédiat de centralisation de données pour le partage au sein de l'équipe pendant la durée de vie du projet. Dans ces cas-là, rares sont les règles posées en amont sur le mode d'organisation ou de gouvernance de ces espaces et le devenir de ces données — dispersées au fil des recherches dans les clouds. La diversité des partenaires, les regards différents qu'elles·ils portent sur la gestion de leurs données et leurs niveaux de perception de ses enjeux rends particulièrement difficile une démarche de bonne pratique d'un·e chercheur·e plus aguerri·e.

Quel que soit le type d'organisation, les données qui ont été stockées ou partagées restent sur le cloud comme sauvegarde mais il n'y a pas de réelle politique d'archivage à long terme de leurs données. Elles ils ont conscience de ne pas être véritablement en conformité avec le RGPD malgré le souhait ou des tentatives pour se former en ligne et stocker dans des clouds plus sécurisés *(ex : hors du système Microsoft, Google.)*. En revanche, cette population n'a pas véritablement d'inquiétude sur les risques contentieux avec ses données (à l'exception du chercheur·e contractuel·le en exercice libéral). Ses appréhensions s'expriment davantage à un niveau éthique, par exemple quand elle fait ses retours auprès des enquêté·es : Comment les protéger en raison des risques d'identification ? Comment éviter toute violence symbolique liée à l'objectivation de son étude ? Comment les préserver en amont d'interprétations erronées de leurs discours ?. En fonction de la connaissance de leur terrain et de leurs enquêté.es, elle est ainsi très attentive sur tout ce qu'elle communique.

- **Réutilisation de données – Partage hors projet**

Leurs matériaux d'enquête ne sont pas utilisés que pour leur enquête cible. Elles ils les conservent et les réécoutent selon leurs besoins pour d'autres enquêtes. Certain·es peuvent très occasionnellement partager/recevoir ponctuellement des transcriptions (voire le corpus complet de retranscriptions d'une enquête) à/d'un·e collègue dans le cadre d'un projet de recherche ou de publication. La dimension de confiance joue alors un rôle fondamental dans cet échange, où l'information est rarement anonymisée. La recontextualisation en direct auprès de l'enquêteur·trice pour mieux appréhender les matériaux apparait comme une condition essentielle. Réutiliser d'autres données qui n'ont pas été collectées par les chercheur·es en question pose en effet le risque d'un appauvrissement de ces données en dépit d'une documentation et de publications existantes. Partager hors du groupe projet son corpus de données référent (ex : entretien et retranscriptions sur Sonal) soulève également un réel problème éthique (éléments identifiants des enquêté·es). La propriété intellectuelle sur les bases de données ne se pose pas de façon aussi claire que pour les chercheur·es en quanti.

5.2 Données personnelles/sensibles

Ces chercheur·es collectent et centralisent sur leur espace de stockage et de partage tout un ensemble de données personnelles/sensibles sur leurs enquêté·es (entretiens audios, retranscriptions ou annotations non anonymisées). En plus des renseignements classiques sur leur état civil (*nom, prénom, etc),* viennent s'ajouter à ces données régies par le RGPD des commentaires pris à chaud au moment de l'entretien *(descriptions ethnographiques des enquêté.es comme par exemple leur vêtement, leur affiliation politique, etc)* et autres notes personnelles de l'enquêteur. Qu'en est-il des dispositifs pris pour protéger leurs enquêté·es ? On constate que le RGPD soulève des questions éthiques et de faisabilité pour leurs enquêtes. Pour autant elle ne doit pas être un obstacle à leurs pratiques de recherche.

- **Pratiques d'accords de consentement ou de confidentialité**

Ce groupe pratique en grande partie des consentements oraux en début d'entretien afin de pouvoir enregistrer les individus ou salarié·es. Les accords spécifiquement écrits concernent davantage les terrains institutionnels ou d'entreprises qui nécessitent une formalisation par accord de consentement ou accord de confidentialité. Ils sont dans les deux cas, une condition obligatoire pour accéder au terrain, fixer le cadre de diffusion et de valorisation de leurs matériaux/écrits, et garantir la confidentialité de certaines informations.

La formalisation du consentement les questionne sur différents aspects :

- Faut-il faire signer ou pas selon le terrain, ses méthodes ou selon l'interviewé·e ?

- Quelle forme écrite faut-il mettre en place, qui puisse être le moins contraignant possible tout en protégeant au mieux les personnes ? Faut-il opter pour un consentement large et imprécis ou un consentement à tiroirs c'est-à-dire avec des consentements adaptés selon leur interlocuteur et leurs usages potentiels ?

- Quel contenu adopter ? Au-delà de leur manque d'expérience sur cet exercice très formel imposé par le nouveau cadre réglementaire, la difficulté à formaliser ce document à l'écrit s'explique en partie par la méconnaissance de l'usage réelle qu'elles·ils vont faire de leurs entretiens. Un certain nombre d'entretiens archivés dans leur cloud personnel n'ont jamais été en effet utilisés. Mais chaque entretien possède le germe d'une réutilisation potentielle future au-delà du contrat financé. Or, cette réutilisation pour d'autres recherches (recyclage de ses matériaux pour d'autres enquêtes) est à présent encadré par le RGPD qu'ils maitrisent encore mal.

- Y a-t-il un « bon moment » pour faire signer la personne ? Si le formulaire est présenté en amont de l'entretien, il risque d'avoir un effet déductif qui peut biaiser les réponses. *A contrario*, le présenter en aval c'est prendre le risque de devoir supprimer l'enregistrement si la personne refuse en définitif cette formalisation en fin d'entretien.

- Quelle que soit leur pratique de consentement (écrit ou oral en amont de l'entretien ou à la fin) ces scientifiques constatent qu'ils n'ont jamais eu de problèmes post-enquêtes avec leurs enquêté·es. Ils veillent toujours à les protéger tout au long de leur pratique professionnelle, jusqu'à même s'autocensurer dans la valorisation d'une enquête quand l'identification est trop à risque. Ils restent très vigilants dans les réunions de restitutions, les communications et publications.

- Ils prennent garde à ce qu'aucune information ne puisse nuire à leurs enquêté.es pendant leurs enquêtes mais aussi au-delà. Ce cadre éthique s'inscrit depuis toujours dans leur pratique professionnelle. Cela commence par un travail de mise en confiance fondamental et qui permettra au chercheur·e de recueillir le maximum de données de qualité sur sa problématique d'autant plus sur des terrains difficiles. Pouvoir créer, rendre/conserver la confiance à/ leurs enquêtes qui leur ont eux-mêmes fait confiance est essentielle. Elle est une garantie pour des échanges fructueux et potentiellement de futures enquêtes. Ainsi, la demande de consentement, quand elle n'est pas une condition obligatoire d'accès au terrain, apparait comme une contrainte extérieure imposée pour éviter de potentiels risques de contestations voire mesures judiciaires qu'elles·ils n'ont jamais connues. Et ce procédé peut mettre plus mal à l'aise l'interviewé·e voire la·le faire se rétracter par crainte de la tournure juridique que revêt l'entretien formalisé avec cette signature. Cette demande de consentement écrit ne s'adapte pas à certains terrains ou à certaines méthodes d'enquête en SHS (ex : observations de terrain, terrains sensibles).

- **Pratiques d'anonymisation/pseudonymisation**

Dans ces méthodes quantitatives — en particulier pour les approches ethnographiques, ces chercheur·es préfèrent travailler, le temps du traitement et de l'analyse, sur leurs matériaux bruts non pseudonymisés. C'est sans doute lié à la relation de confiance et de proximité forte qu'elles·ils établissent et entretiennent avec leurs enquêté·es ou sur le terrain, et la crainte d'analyser des données décontextualisées, plus difficiles du coup à appréhender et à bien interpréter. Elles ils pratiquent en majorité la pseudonymisation au moment de la restitution auprès de leurs enquêté·es — un gage pour eux de respect et de maintien de cette relation de confiance qui ont réussi à établir avec eux — et pour leurs publications. La pseudonymisation se fait souvent de manière contextuelle selon l'extrait retenu pour un article avec une table de correspondance. Certains terrains ne permettent pas une véritable anonymisation car l'entreprise ou les acteurs clés d'un domaine ou secteur peuvent être très facilement identifiés malgré le changement de noms et lieux.

- **Questionnements techniques**

- Le problème de l'altération de la qualité des vidéos à chaque transfert entre les 2 supports de travail — plateforme de stockage et disque dur pour les analyses — aggravé d'autant par la fréquence de ces transferts ;

- La difficulté à bien maitriser l'usage et les possibilités de la LESTbox pour communiquer des matériaux entre dossiers de chercheurs.

- **RH & Budget**

- Le besoin exprimé d'une aide au niveau local pour synchroniser des vidéos, faire des montages documentaires pour des publications, etc.

- Une question soulevée : « *Comment valoriser des matériaux plus anciens d'enquêtes qui nécessitent une numérisation ? Quels moyens le laboratoire pourrait-il mettre en place pour sauvegarder et valoriser ces matériaux* ? »

- **Questionnements RGPD & Ethique**

- Comment connaitre ce qu'il est possible de faire ou de ne pas faire avec la nouvelle réglementation ?

- Comment homogénéiser leurs pratiques en matière de gestion des données (conventions communes) sans que les règles ne deviennent trop contraignantes (voire handicapantes) pour nos recherches ?

- Comment faire en l'absence de consentements écrits et de difficultés pour retrouver les personnes interviewé·es pour valoriser d'anciens matériaux de recherche ?

6 FOCUS GROUP DOCTORANT.ES

Ce groupe a été mis en place le 21 décembre 2020 en distanciel. Il y a eu une très forte mobilisation de ces doctorant·es dans une période qui s'y prêtait peu (COVID et fêtes de fin d'année). Ce fut un échange riche et dense en raison de la diversité des profils avec la contrepartie d'une discussion un peu moins approfondie afin que tou·tes s'expriment. Ces participant·es étaient ravi·es de pouvoir échanger sur des questions qu'elles·ils se posent ou se sont posées sans oser véritablement les formuler. On les remercie vivement pour ces échanges très fructueux.

Portrait rapide de cette population d'enquêté.es :

Onze 11 doctorant·es appartenant majoritairement à la sociologie et à la gestion (*6 en sociologie et 4 en gestion*) appliquant des méthodes principalement qualitatives ou dans une moindre mesure des données mixtes (*4 étudiant·es selon leur déclaratif*). Les 2ème et 3ème années prédominaient. Seul·es celles et ceux de 4ème année n'étaient pas représenté·es car en situation de rédaction ou en fin de terrain. La répartition était la suivante : 2 de 1ère année, 4 de 2ème année et 3ème année respectivement, 2 de 5ème année. Trois au moins de ces doctorant·es étaient sous contrat CIFRE. Un seul étudiant était inséré dans un projet de recherche.

Beaucoup travaillent avec des données qualitatives : entretiens, observations de terrain, comptes rendus de terrain, données produites par les entreprises, etc. Leurs problématiques sont davantage orientées sur des questions de collecte et d'exploitation des données. Les points saillants portaient en particulier sur les demandes de consentement (RGPD), le rôle du comité éthique, l'utilisation des données d'entreprises/d'organisations, leur positionnement de chercheur·e sur le terrain (CIFRE).

Compte tenu de leur long apprentissage du métier de chercheur·e, il est assez logique de constater qu'elles·ils ne se projettent pas au-delà de leur thèse et de l'exploitation potentielle de leurs données. Les échanges au cours de ce *focus group* les ont amenés à prendre un peu de hauteur sur leurs matériaux, à voir plus loin en prenant conscience de la diversité des enjeux autour de la donnée, de leurs données. Il y a un vrai besoin de leur part de discuter sur ces problématiques et de bénéficier de conseils et d'expériences. Tous·tes ne les abordent pas vraiment au sein de leur comité de suivi de thèse qui est plus centré sur la méthode scientifique, et eux-mêmes n'y pensent pas ou ne s'y autorisent pas. En outre, ce comité peut arriver tardivement pour ces questions de données qu'il faudrait aborder très tôt pour mettre en place de bonnes pratiques et les rassurer (la 3ème année pour certain·es).

6.1 Des difficultés diverses à différents niveaux du travail doctoral

- **Collecte des matériaux**

Les doctorant·es de ce laboratoire ne pratiquent pas **la réutilisation de données** de chercheur·es du LEST. En revanche, elles·ils sont plus facilement associé es à des projets de recherche où elles·ils peuvent recueillir leurs propres données. Elles·ils sont très attentif·ves. Certain·es ont recours à une convention labo-institution pour accéder à des données plus fines « données publiques » produites par l'institution dans laquelle le terrain se déroule.

- **Type de matériaux**

Avec le COVID, un plus grand nombre ont collecté davantage de **matériaux de terrain avec des données personnelles ou institutionnelles prenant la forme de captations** d'images/écrans, de vidéos, de réunions ou de messages intra-entreprises, chats de groupes de discussion réservés comme dans WhatsApp (*matériaux peu utilisés au niveau du laboratoire)*. Ces matériaux très volumineux posent toutes sortes de problèmes juridiques (RGPD) et de posture/relation avec les institutions ou groupements qui les accueillent (droit de stockage, sauvegarde et traitement). Ont-ils le droit de stocker et sauvegarder pour leur propre usage ces données ? Les conventions CIFRE leur paraissent difficiles à comprendre. Elles·ils ont du mal à faire la traduction juridique sur leurs cas concrets. Quelques exemples de questions que se posent ces doctorant·es : *De quelle liberté peuvent-ils disposer en marge des demandes d'accords formels pour recueillir et sauvegarder, exploiter leurs matériaux de recherche ? Quels sont les risques ? A qui appartient finalement les données internes d'entreprises collectées par eux ?*

- **Formulaire de consentement dans le cadre des entretiens**

On a pu observer dans ce groupe deux générations ou deux profils de doctorant·es.

- Le 1er profil concerne celles et ceux qui ont déjà suivi des cursus à l'étranger en particulier dans les pays anglosaxons ou héritiers et sont plus au courant des démarches en matière d'intégrité et d'éthique scientifique : Elles·ils ont suivi des formations intégrées à leur cursus et dès le projet de thèse ont appliqué la procédure de leur université (cotutelle principale). La soumission de leur projet est passée devant le Comité éthique de l'université qui a validé la faisabilité de la thèse, après consultation de leur guide d'entretien et de leur formulaire de consentement. Il s'agit d'une condition obligatoire pour démarrer la thèse, de même celle de faire signer le formulaire aux enquêté·es au moment de l'entretien pour garantir sa soutenance. Tout changement en cours de thèse doit faire l'objet d'un examen par le comité. Cette population est ainsi beaucoup plus rigoureuse et prudente sur la collecte de données personnelles/sensibles en raison des procédures réglementaires plus contraignantes dans ces pays. Les formations obligatoires qu'elles·ils ont suivies sont reconnues par eux comme très utiles mais aussi très insuffisantes pour répondre à leurs questionnements relatifs à la mise en pratique sur leur terrain.

- En l'absence de cadre pédagogique (notamment pour les générations de thésard·es non formé·es) ou de cadre réglementaire plus stricts, les autres doctorant·es du laboratoire ne savent souvent pas comment procéder et quel formulaire de consentement proposer pendant leur entretien. D'autres qui pratiquent des méthodes d'enquêtes ethnographiques ne se voient pas utiliser de formulaire de consentement même après plusieurs années de thèse derrière eux. Selon eux, en formalisant trop les choses, cette démarche pourrait leur fermer certains terrains ou l'accès à certaines personnes, voire rendre plus difficile la construction de leur place de chercheur·e sur le terrain. Les étudiant·es en CIFRE ont également du mal à se projeter avec un formulaire de consentement à faire signer à leurs collègues de travail alors qu'elles·ils ont bien réussi leur intégration dans l'entreprise. Cela les questionne sur la manière de se positionner en tant que chercheur·e : Comment faire passer des procédures dans le cadre d'un travail académique hors contexte, au risque de remettre en question leur image, leur positionnement ou intégration ?

Elles·ils se questionnent également sur la nécessité de se protéger et des risques réels qu'encoure la·le chercheur·e : Se protéger contre quoi ? Y a -t-il des cas judiciaires observés en SHS qui puissent les éclairer et les conforter ? Certain·es ont formulé leurs inquiétudes sur les pratiques et procédures plus standardisées pratiquées à l'étranger, en particulier sur les comités éthiques. Leur légitimité (rôle et composition) est soulevée car elles·ils craignent que le processus de validation pour les SHS nuisent à la réalisation de certains terrains. Le comité éthique de l'université d'Aix Marseille ne s'inscrit pas dans une procédure obligatoire. Ce comité doit être saisi par le chercheur ou le doctorant qui en fait la demande explicitement. Par cette souplesse, ces derniers conservent leurs choix épistémologiques et la responsabilité pleine de leur recherche.

- **Stockage et sauvegarde des données**

Compte tenu de l'enjeu que représente leur thèse, de sa place centrale dans leur quotidien de travail, d'ordinateurs personnels — non sécurisés par un service informatique (*risques d'obsolescence ou de failles techniques*) — **tous·tes ont conscience de l'importance de garantir la sauvegarde de leurs données sur des supports qui offrent une pérennité au-delà de leur soutenance de thèse.** Comme leur trajectoire institutionnelle est incertaine, beaucoup ont recours à deux types d'outils principaux pour sauvegarder leurs données : les outils gratuits des GAFAM (Google Drive, Microsoft OneDrive, Apple iCloud, etc.) ou des services payants comme Dropbox par abonnement privé. Plus rarement et plus récemment certain.es utilisent la LESTbox et son outil de synchronisation avec l'appui de l'informaticien du laboratoire avec la garantie de pouvoir conserver leurs données même après le doctorat.

- **Traitement des données,** *en mode bricolage pour certain.es*

Les principales difficultés exprimées portent sur :

- Le choix des logiciels de traitement de données et la connaissance du rapport coût d'entrée/rendu, et leur appropriation (absence de formation). Pour les méthodes quantitatives, elles·ils optent pour le logiciel disponible dans leur université et pour le qualitatif, elles·ils optent pour une analyse apparemment plus facilement manuelle des données.

- Les techniques d'anonymisation et pseudonymisation : leurs méthodes, selon certain·es, sont très perfectibles et peuvent être améliorées.

- **Restitution des résultats de l'enquête**

Elles·ils ne savent pas toujours comment procéder en matière de restitution auprès de leurs enquêté·es. Elles·ils sont demandeurs d'échanges et d'expériences. Certain·es par exemple ont de simples accords informels pour observer et collecter leurs données sur le terrain. Le cadre n'est pas forcément propice pour des restitutions plus officielles. Se pose également la question de l'anonymisation. *Comment conserver par exemple l'anonymat d'un terrain facilement identifiable même sans le nommer ?*

- **Partage des données**

L'idée de partager leur donnée n'est pas envisageable à leur niveau d'apprentissage du métier.

6.2 Une expression de besoins particulière

- **Un·e référent·e pour les données et intégrité** bien identifié·e au laboratoire qui pourrait les écouter et les conseiller, mieux les renseigner, les outiller et avoir des échanges sur leurs questionnements ponctuelles ;
- **Un comité de suivi de thèse qui aborde plus tôt ces questions dans leur formation ;**
- **Une page ressources** présentant un panorama des logiciels de traitement de données (quali et quanti) avec leur disponibilité à l'université, dans les laboratoires ;
- **Des formations complémentaires** à celles prochaines proposées par les écoles doctorales (PGD/DMP en 2021) comme celles sur les logiciels de traitement des données.

III. PISTES ET PROPOSITIONS

Les résultats de cette enquête invitent à élaborer au laboratoire une « politique de la donnée » à la fois plus globale et davantage définie. Une telle politique permettrait de fournir aux équipes un ensemble d'outils, de formations et de ressources documentaires pour monter en compétences sans alourdissement excessif de la charge de travail. De ce point de vue, elle ne peut que s'inscrire dans un temps long, afin que personne ne se décourage. Le RGPD prévoyant un régime dérogatoire pour la recherche scientifique[7], il s'agit moins d'envisager une mise en conformité sous stress organisationnel qu'une montée graduelle en compétence collective.

7 OBJECTIFS D'UNE POLITIQUE DE LA DONNEE : 7 ORIENTATIONS

- ✓ **Formation de l'ensemble des chercheur·es et des doctorant·es** pour permettre une montée en compétences RH dans la préparation des enquêtes, l'élaboration des plans de gestion de données, la documentation sur les données (formation, recrutement, etc.).
- ✓ **Mutualisation des pratiques de recherche sur les données par des échanges d'expériences** inscrits dans la programmation scientifique du laboratoire afin de gagner en efficacité et en temps dédié à la recherche.
- ✓ **Appui aux équipes de recherche pendant les phases stratégiques des étapes du projet** (période de soumission, démarrage, etc.)
- ✓ **Sécurisation des données** (en particulier les données personnelles et sensibles) par la mise en place d'un espace sécurisé dédié aux « données » au LEST (dissocier le stockage des données en cours d'utilisation de l'archivage qui feront ou pourraient faire l'objet d'une valorisation future ou d'une réutilisation à des fins pédagogiques ou de recherche) ; chiffrement systématique des disques durs pour l'ensemble des chercheur·es, etc.
- ✓ **Elaboration d'une politique d'archivage des données et des plans de gestion de données**. Une conservation institutionnelle des futurs plans de gestion de données élaborés au LEST dans le cadre de projets de recherche, permettra de mutualiser les expériences et d'identifier à la fois les forces et les faiblesses dans ce domaine en vue d'un appui du laboratoire.
- ✓ **Diffusion et valorisation des données de recherche des équipes sur des plateformes institutionnelles** : création de compte institutionnel sur certaines plateformes de données (par exemple Zenodo) et auprès de services comme l'INIST pour la mise à disposition de DOI pour leurs jeux de données, etc.
- ✓ **Appui à la publication** pendant la phase d'écriture en facilitant l'accès aux données et pour l'écriture d'articles de données (*data journals)*.

[7] https://scinfolex.com/2018/07/18/donnees-personnelles-et-recherche-scientifique-quelle-articulation-dans-le-rgpd/

8 ENJEUX D'ORGANISATION POUR LE LABORATOIRE : 18 PROPOSITIONS

Les suggestions proposées ici ont vocation à être discutées, enrichies et modifiées dans le cadre d'une discussion ouverte avec les membres du laboratoire et la Direction avant la mise en place d'un plan d'action plus opérationnel.

- **PROPOSITION 1.** Organiser, en lien avec AMU et le CNRS, un **dialogue resserré avec les acteurs des services et plateformes nationales sur l'accès et le dépôt aux données** (PROGEDO, Huma-Num). Une possibilité serait de prendre appui sur le plan de formation de l'unité pour mettre en place des formations ou des journées d'étude pour aider les chercheur.es à se les approprier.

- **PROPOSITION 2. Lancer une tribune nationale de chercheur.es et d'institutions** *« pour avertir les producteurs que le CASD est un système extrêmement coûteux pour les sciences sociales et que ces données publiques pourraient ne pas être exploitées à l'optimal ou du moins par tous·tes au sein des UMR et des communautés de recherche ».*

- **PROPOSITION 3.** Organiser en lien avec AMU et les autres laboratoires, sur une base de volontariat, un **dispositif simple de recensement des licences individuelles des chercheur·es afin de pouvoir ouvrir des négociations plus avantageuses pour des licences commerciales** de bases de données spécialisées.

- **PROPOSITION 4.** Construire avec nos partenaires **une page de ressources à destination des doctorant·es et arrivant·es présentant un panorama des logiciels de traitement de données (quali et quanti)** avec leur disponibilité à l'université et dans les laboratoires (*via* par exemple l'offre en ligne des tutoriels sous libguides du SCD ou le site ressources des PUD AMU).

- **PROPOSITION 5.** Initier une **discussion CEREQ/LEST sur une politique de la donnée du centre associé** : voir, par exemple, comment mieux intégrer les chargés d'études du Céreq dans les recherches du LEST pour avoir des interlocuteurs capables d'aiguiller sur la prise en main des données, voire faciliter potentiellement l'accès à des programmes qui permettent d'utiliser plus facilement les données.

- **PROPOSITION 6.** Privilégier des **accords de confidentialité sur les données entre les institutions (niveau le plus pertinent)** plutôt qu'au niveau des chercheur·es et de l'enquêté·e quand cela est possible.

- **PROPOSITION 7.** Mettre en place **un·e ou des référent·es** au sein du laboratoire : - une personne de proximité spécialisée sciences sociales qui connaisse la richesse et diversité de ces disciplines, des enjeux du laboratoire et qui puisse répondre de façon pointue aux problématiques éthiques *(ex : un·e chercheur·e senior·e)* ; - une personne— avec une position « hybride » entre les activités de la documentation, l'informatique et la communication qui puisse répondre aux problématiques techniques et réglementaires, être référent·e données & intégrité. Cette personne pourrait écouter les juniors, les conseiller, mieux les renseigner, les outiller et avoir des échanges sur leurs questionnements ponctuels.

- **PROPOSITION 8.** Dégager une **enveloppe financière pour des accès à des données/bases de données dans le cadre d'un projet de publication après clôture du financement sur contrat ou hors contrat** de la même façon qu'il y a des financements pérennisés pour des relectures en anglais en appui à la publication. Etudier également les possibilités de monter certains projets *ad hoc* pour mutualiser des ressources, si c'est possible.

- **PROPOSITION 9.** Recourir à une **aide financière pour les projets de valorisation des anciennes données** *(retranscriptions d'entretien, numérisation, etc.).*

- **Proposition 10.** Inscrire et aborder très tôt dans le suivi des doctorant.es ces questions de gestion, d'éthique et d'intégrité des données **au sein** du **comité de suivi de thèse.** Celui-ci pourrait les aiguiller au démarrage et selon leur terrain sur une méthodologie de recueil et d'organisation/classement de leurs données, et sur la rédaction de leur plan de gestion pour les préparer au dépôt de projets de recherche.

- **PROPOSITION 11.** Inscrire dans la **politique d'accompagnement des doctorants un suivi relatif à la restitution aux enquêté·es** et au tissage des relations de partage de données avec les organisations enquêtées.

- **PROPOSITION 12**. Relancer **la commission annuelle de suivi sur la sécurité des données du laboratoire** (Direction, SG, Administrateur réseau et support informatique, correspondant IST, chercheur·e expert·e volontaire) qui permettait de mettre en discussion les enjeux de sécurité et de partager la responsabilité et les risques autour des données.

- **Proposition 13.** Inscrire systématiquement **tout projet de recherche porté par le LEST dans une démarche d'anticipation de sa gestion des données** en lien avec l'équipe soutien à la recherche.

- **PROPOSITION 14.** Organiser des **temps d'échanges, d'expériences** permettant de mieux appréhender les pratiques de terrain selon les types de données/méthodes ou population (*focus spécial doctorant·es*) lors des discussions du Conseil des axes du laboratoire, ou lors de la présentation des principaux résultats au cours d'un séminaire général avec tous les membres du laboratoire (*point à inscrire à l'ordre du jour*).

- **PROPOSITION 15. Proposer un partage annuel de bonnes pratiques de gestion de données de recherche** autour de questions ciblées (éthiques, RGPD, méthodologiques) pour améliorer collectivement les pratiques, faciliter par la suite les conditions de partage ou de valorisation dans le respect du RGPD. Par exemple : Comment chacun protège-il ses enquêté·es ? Comment met-on en commun correctement des données qualitatives ? Quelles sont les bonnes pratiques en matière d'anonymisation, de restitution ? Le cadre proposé : séminaires, approche individuelle confidentielle sur la pratique de terrain.

- **PROPOSITION 16.** Créer une **boîte à outils organisée par grandes thématiques** (RGPD, plan de gestion de données, services de partage et stockage, etc.) qui fournirait les tutoriels, les formulaires et les outils nécessaires pour accompagner les équipes sur la gestion de leurs données de recherche en tenant compte de ce qui existe déjà dans l'environnement recherche (ex : DORANum, Libguides AMU « données de la recherche, etc.) et qui puisse proposer des solutions étape après étape, au fur et à mesure de l'avancée d'une enquête et si possible de façon graduée (obligatoire et facultatif) pour être au plus près de leurs besoins et disciplines**.**

- **PROPOSITION 17.** Tenir un **registre d'inventaire des données d'enquêtes du LEST** pour rendre possible d'éventuelles collaborations *(nom de l'enquête, données utilisées/produites, personne à contacter, etc.).*

- **PROPOSITION 18.** Organiser **l'accompagnement au montage des vidéos** prises sur le terrain sous des angles différents (encodage, synchronisation de vidéos, sous-titres, etc.).

IV. RÉFÉRENCES, ANNEXES MÉTHODOLOGIQUES

9 BIBLIOGRAPHIE

Bizien L., Dom V. (2019). *Les données de la recherche AAU-CRENAU : résultats de l'enquête sur les usages des chercheurs, doctorants et ingénieurs en matière de gestion de données*. Research Report. Centre de recherche nantais Architectures Urbanités. https://hal.archives-ouvertes.fr/hal-02420916.

Callon M., Latour B. (1991). *La science telle qu'elle se fait. Anthologie de la sociologie des sciences de langue anglaise*, Paris, La Découverte, 340 p.

Cartier A., **Moysan M., Reymonet N.** (2015). Construire des outils pour la gestion des données de la recherche dans une communauté d'universités**,** ***Journée sur les données de la recherche***, ADBS, Jan 2015, Paris, France. https://hal-descartes.archives-ouvertes.fr/hal-01138663v1

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

Gauquelin C., Lutoff C., La Branche S. (2017). « Quelle place pour les données issues des sciences humaines et sociales dans le développement de services climatiques régionaux ? », *VertigO - la revue électronique en sciences de l'environnement* [En ligne], Volume 17 numéro 3, décembre 2017. http://journals.openedition.org/vertigo/18891 ; DOI : 10.4000/vertigo.18891(1)

INSHS. (2019). *Guide Pour La Recherche : Les Sciences Humaines et Sociales et La Protection Des Données à Caractère Personnel Dans Le Contexte de La Science Ouverte.* https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf

Pontille D., Torny D. (2013). La manufacture de l'évaluation scientifique. *Réseaux*, (1), p. 23-61.

Prost H., Schöpfel J. (2015). L*es données de la recherche en SHS. Une enquête à l'Université de Lille 3. : Rapport final*. https://hal.univ-lille3.fr/hal-01198379/document

Rebouillat V., Chartron G. (2019). Services de gestion et de partage des données de recherche : ce qu'en pensent les chercheurs ?. *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, Oct 2019, Montpellier, France. https://hal.archives-ouvertes.fr/hal-02307085

Schöpfel J. (2018). *Vers une culture de la donnée en SHS : Une étude à l'Université de Lille*. Rapport de recherche de l'Université de Lille. https://hal.archives-ouvertes.fr/hal-01846849.

Serres A., Malingre M.L, Mignon M., Pierre C., Collet D. (2017). *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2* : Rapport de l'Université Rennes 2. https://hal.archives-ouvertes.fr/hal-01635186

10 ANNEXE 1 : QUESTIONNAIRE

Structure du questionnaire et nombre de questions

	Questions obligatoires	Questions facultatives	TOTAL
1-Vos méthodes et données de recherche ?			
1.1-Méthodes	0	3	3 (2 à choix multiples)
1.2-Données sources et produites	1	4	5 (1 à choix multiples)
1.3-Documentation sur vos données	2	1	3
1.4-Traitement des données personnelles	2	2	4
2-Pratiques			
2.1-Stockage et archivage	4	5	9
23.2-Partage et diffusion	2	4	6
3-Besoins et attentes	2	2	4
4-A propos de vous...	0	6	6
TOTAL			40

QUESTIONNAIRE

1.Vos méthodes et données de recherche

1.1-Méthodes

Q1. Quelles sont vos méthodes de recherche ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Méthodes qualitatives
2. Méthodes quantitatives
3. J'utilise les deux types de méthodes
4. Travail principalement théorique, épistémologique

Q2. Si méthodes qualitatives ou les deux : merci de préciser
Choisissez TOUTES les réponses qui conviennent

1. Enquête par entretiens
2. Observations
3. Ethnographie
4. Analyse de documents textuels, d'archives ou de photos, vidéos
5. Focus Groups ou débats
6. Méthode participative/Recherche action
7. Autres

Q3. Si méthodes quantitatives ou les deux : merci de préciser
Choisissez TOUTES les réponses qui conviennent

1. Analyses statistiques descriptives (univariées, bivariées)
2. Analyse factorielle ou géométrique des données (ACM, ACP,…)
3. Analyses statistiques multivariées et économétrie (régressions, modèles logistiques, probit, modèles multiniveaux, analyse de panel, etc.)
4. Analyses de réseaux
5. Analyses de trajectoires, de séquences
6. Autres

1.2-Données sources et produites

Q4. D'où proviennent les « données sources » *(bulle - définition)* que vous utilisez dans vos recherches ?

Choisissez la réponse appropriée pour chaque élément

	SYSTEMATIQUEMENT	FREQUEMMENT	RAREMENT	JAMAIS
Je collecte moi-même mes données sur le terrain				
De mon réseau de recherche (données mise à disposition directement par d'autres chercheurs)				
De sources publiques (bases de données de grandes enquêtes disponibles sur PROGEDO ADISP (Quetelet) ou auprès d'autres organismes publics…)				
Depuis Internet (web scraping)				
De sources privées (données d'entreprises, associatives, etc.)				
D'un environnement contrôlé (laboratoire artificiel)				
Autres sources				

Q5. Dans quelle(s) catégorie(s) placeriez-vous les données sources que vous utilisez ?

Choisissez la réponse appropriée pour chaque élément

	SYSTEMATIQUEMENT	FREQUEMMENT	RAREMENT	JAMAIS
Données textuelles - hors entretiens (corpus de textes, archives papiers ou numériques)				
Données d'entretiens individuels				
Données d'entretiens collectifs (focus groups par exemple)				
Données d'observations du terrain				
Données d'enquêtes quantitatives pour un traitement statistique				
Données chiffrées agrégées (sous forme de tableaux, etc.)				
Données web (données d'usages, cartographies de sites, données de réseaux sociaux, etc.)				
Données multimédias (documentaires, films, etc.)				
Données audio hors entretiens (archives sonores de discours, prises de sons diverses…)				
Images fixes d'objets, de paysages, d'architectures, scans de textes (cartes, plans, photographies, affiches…)				
Données expérimentales				
Autres données				

Q6. A propos des entretiens individuels ou collectifs, les transcrivez-vous par la suite ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, systématiquement
2. Oui, parfois
3. Jamais

Q7. Concernant leur support matériel, vos données sources sont-elles majoritairement :
Veuillez sélectionner SEULEMENT UNE réponse

1. Non numériques (par exemple : des entretiens non enregistrés en numérique initialement, etc.)
2. Numérisées (par ex : des corpus numériques, des données statistiques , etc.)
3. Nativement numériques (par exemple : les données web)

Q8. Comment utilisez-vous les données que vous obtenez de sources extérieures ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Comme elles sont, sans aucun problème
2. Avec un peu d'effort (nettoyage et petites modifications)
3. Après beaucoup de travail et d'effort avant de les exploiter
4. Je n'utilise pas de données de sources extérieures, car je collecte toujours moi-même mes données

Q9. Dans quelle catégorie classeriez-vous vos données produites après exploitation (vos « données résultats ») ?
Choisissez TOUTES les réponses qui conviennent

1. Données d'analyses graphiques (schémas de réseaux, graphes, frises chronologiques, etc.)
2. Données d'analyses statistiques (tableaux, etc.)
3. Données d'analyses textuelles (extraits d'entretiens analysés, annotations de corpus de texte, etc.)
4. Iconographies (illustration, croquis, schémas, affiches, cartes, plan)
5. Modélisation
6. Montage sonore
7. Montage filmique
8. Montage photographique
9. Autres

1.3-Documentation sur vos données

Q10. Produisez-vous une information sur vos « données sources » ou vos « données résultats » qui permettent de comprendre vos données ?

1. Oui
2. Non

Q11. Si oui pourriez-vous préciser leur forme :

Choisissez la réponse appropriée pour chaque élément

	Plutôt OUI	Plutôt NON
Information scientifique (méthodologie d'enquête, plan d'échantillonnage…)		
Information technique (logiciel utilisé, liste des fichiers de données, format, …)		
Information standardisée (titre, auteur, mots-clés, date de création, droits associés sur les accès et les licences, DOI)		

Q12. Pourriez-vous nous indiquer plus précisément les types de documents que vous produisez pour permettre une meilleure compréhension dans leur lecture et dans leur traitement ?

Choisissez la réponse appropriée pour chaque élément

	Systématiquement	Fréquemment	Rarement	Jamais
Procédure d'enquête et méthodologie				
Méthode d'échantillonnage				
Définition des populations/entités inclus dans l'enquête				
Guide d'entretien				
Dictionnaire des variables et distribution des variables				
Grille de codage des données d'entretien				
Description de l'arborescence de classement des fichiers				
Fichier d'exécution de traitements de données				
Carnet de recherche papier ou numérique (blog)				
Règle de nommage des jeux de données				
Cadre de sécurité des données				
Autres				

Q13. Les données que vous collectez pour vos travaux contiennent différents types d'informations, pouvez-vous nous en dire plus à ce sujet :

Choisissez la réponse appropriée pour chaque élément

	La plupart du temps	Parfois	Jamais
Les données collectées contiennent des données personnelles			
Les données collectées contiennent des données sensibles soumises au régime de protection des données (ex : données d'opinions politiques collectées avec le consentement des enquêtés)			
Les données collectées contiennent des données couvertes par le secret statistique ou fiscal (ex : données fournies par le CASD)			
Les données collectées sont déjà largement partagées dans l'espace public (articles de presse, documents publics en ligne, etc.)			

Q14. Comment gérez-vous les données personnelles ?

Choisissez la réponse appropriée pour chaque élément

	La plupart du temps	Parfois	Jamais
Les données sont anonymisées une fois le traitement terminé			
Les données sont pseudonymisées une fois le traitement terminé			
J'informe les personnes concernées oralement			
J'informe les personnes concernées de la collecte et du traitement de leurs données personnelles en leur fournissant un document écrit			
Je demande le consentement des personnes par écrit			
Les données personnelles ou sensibles sont supprimées après traitement			
L'accès aux données personnelles et/ou sensibles est protégé le temps du traitement			
Je supprime les données qui contiennent des données personnelles ou soumise à un régime administratif de protection			

Q15. Avez-vous déjà fait une déclaration à la CNIL pour enregistrer la collecte et le traitement de données ?

Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, plusieurs fois
2. Oui, une fois
3. Non, jamais

Q16. Etes-vous familier·ère avec la nouvelle règlementation en vigueur sur la protection des données (RGPD), en application depuis le 25 mai 2018 ?

Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, tout à fait
2. Oui, plutôt
3. Non

2-Pratiques

2.1-Pratique de stockage et archivage des données :

Q17. Retrouvez- vous facilement vos données ?

1. Oui
2. Non

Q.18 Diriez-vous que vous faites un stockage de vos données plutôt…
Veuillez sélectionner SEULEMENT UNE réponse

1. Systématique
2. Ordonné
3. Peu ordonné
4. Désordonné

Q19. A quelle fréquence faites-vous les copies de sauvegarde de vos données ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Une fois par mois ou plus souvent
2. Tous les trimestres
3. Tous les ans
4. Irrégulièrement, en fonction des besoins
5. En fonction du prestataire de service ou du gestionnaire de site
6. Je ne fais jamais de copie de sauvegarde

Q20. Qui se charge en règle générale de l'archivage des données du projet ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Moi-même
2. Un membre l'équipe de recherche dans une cadre d'un projet collaboratif
3. Un.e doctorant.e
4. Un membre d'un service de soutien à la recherche
5. Un.e prestataire de service externe
6. Il n'y a pas de responsable identifié.e

Q21. Où archivez-vous les données de recherche de vos projets achevés ?
Choisissez la réponse appropriée pour chaque élément

	La plupart du temps	Parfois	Jamais
En local, dans des dossiers papiers			
En local, sur mon ordinateur personnel			
En local sur mon ordinateur professionnel			
Sur un disque dur externe ou autre support (clé USB, CD, etc.)			
En réseau, sur le serveur du laboratoire (serveur du LEST, LEST-box)			
En réseau sur le serveur de nos tutelles (préciser : université ou CNRS)			
Sur un serveur externe hébergé par une autre institution			
Sur le serveur d'une TGIR (Très Grande Infrastructure de Recherche)			
Sur une plate-forme privée dans le "cloud" (Google Drive, Dropbox, etc.)			

2.2-Pratiques de partage et de diffusion des données

Nous aimerions maintenant en savoir plus sur vos pratiques de partage et de diffusion des données

Q22. Avez-vous déjà réutilisé des données collectées par d'autres chercheur·es (mises à part les données dites de « Grandes enquêtes »)
Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, au moins une fois
2. Oui, régulièrement
3. Non, pas pour le moment
4. Non, et je ne le ferai pas

Q23. Parmi les raisons suivantes, qu'est-ce qui vous inciterait à rendre vos données de recherche accessibles en libre accès ? *Choisissez TOUTES les réponses qui conviennent le mieux à votre expérience personnelle : Choisissez la réponse appropriée pour chaque élément*

	Très incitatif	Assez incitatif	Pas du tout incitatif
Une meilleure visibilité de mes travaux de recherche			
La reconnaissance de la communauté scientifique			
L'occasion de nouveaux contacts, de nouvelles coopérations scientifiques			
Le souci de pouvoir faire valider mes résultats à partir des jeux de données *(valeur de preuve*)			
Le souhait de voir se développer d'autres recherches à partir de mes jeux de données			
Leurs valeurs historiques			
A titre pédagogique pour la formation des jeunes chercheur·es			
L'adhésion aux valeurs du libre accès aux données et aux résultats de la recherche scientifique			
Une obligation faite par le financeur de ma recherche *(ex. dans le cadre d'Horizon 2020)*			
De possibles retombées financières (crédits de recherche, etc.)			

Q24. Quelles seraient les principales raisons qui vous empêcheraient ou vous freineraient pour rendre certaines de vos données de recherche accessibles en libre accès ?
Choisissez la réponse appropriée pour chaque élément

	D'accord	Pas d'accord
Des freins juridiques liés à la nature des données		
Des raisons économiques *(protection par rapport à la concurrence*)		
Des raisons *scientifiques (peur de perdre l'avantage, manque de reconnaissance dans le processus d'évaluation*)		
Des raisons éthiques (*la crainte d'une utilisation des données qui ne correspond pas aux objectifs premiers de la recherche (confiance des personnes enquêtées)*		
La crainte du plagiat		
Des raisons personnelles *(je ne veux pas montrer mon "arrière-cuisine", mes données sont à moi)*		
Le manque de temps *(trier les données*), la lourdeur de ce type de travail		
Le manque de connaissances dans la description des jeux de données, les modalités de dépôt		
Le manque d'infrastructures		

Q25. Dans vos recherches, à quels niveaux se sont situées les principales difficultés que vous avez rencontrées en termes de données ?

	OUI	NON
Accès aux données		
Traitement de vos données		
Partage de vos données		
Stockage de vos données		
Archivage de vos données		
Valorisation de vos données		

Q26. Pratiques de partage et de diffusion des données. Nous aimerions maintenant en savoir plus sur vos pratiques de partage et de diffusion des données. Avez-vous déjà partagé les données que vous avez-vous-mêmes collectées (entretiens, bases de données…) ou que vous avez traitées (transformation et traitement des données d'enquêtes) ?

1. Oui
2. Non

Q27. Avec qui avez-vous principalement partagé vos données ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Avec des collègues dans le cadre d'un projet
2. Avec des collègues en dehors du cadre d'un projet
3. Avec des collègues dans le cadre d'un projet et en dehors du cadre d'un projet

Q28. Les collègues avec lesquels vous avez partagé vos données appartenaient-elles·ils au même laboratoire que vous ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, pour la plupart
2. Non, en général, il s'agissait plutôt de collègues extérieurs au laboratoire

Q29. Avez-vous déjà mis des données que vous avez collectées en libre accès sur le web ?
Veuillez sélectionner SEULEMENT UNE réponse

1. Oui, plus d'une fois
2. Oui, une fois
3. Non, jamais

Q30. Si vous souhaitiez diffuser vos données en libre accès, quels sont (seraient) vos supports principaux ? *Choisissez TOUTES les réponses qui conviennent*

	Plutôt OUI	Plutôt NON
Une plateforme spécialisée dans l'archivage et la diffusion des données de recherche		
La plateforme préconisée par mon financeur		
Un carnet de recherche, blog de recherche		
Le site du laboratoire		
Une bibliothèque numérique (collections numériques)		
Un réseau social académique (ResearchGate, Academia, etc.)		
Votre blog ou votre site personnel		
Un serveur de l'université		
Une archive ouverte (de type HAL, etc.)		
Le site de l'éditeur		
Je ne souhaite pas diffuser mes données en libre accès		

Q31. Quels types de données de recherche déposeriez-vous ?
Choisissez la réponse appropriée pour chaque élément

	Plutôt OUI	Plutôt NON	Je ne suis pas concerné·e
Bases de données quantitatives (brutes ou transformées)			
Cartes			
Corpus d'archives			
Fichiers audios			
Photos			
Vidéos			
Graphes			
Tableaux statistiques			
Transcriptions			
Notes, synthèses, comptes-rendus			
Documentation sur la méthodologie ou le traitement			
Fichier d'exécution de traitements de données			
Autres types de données			

3- Besoins et attentes

Q32. Dans quelle mesure êtes-vous familier.ère avec … ?

Choisissez la réponse appropriée pour chaque élément

	Familier·ère	Peu familier·ère
Les enjeux du libre accès *(Open access)*		
Les méthodes de gestion des données (plan de gestion de données)		
Les débats déontologiques sur l'usage de différents types de données		

Q33. Sur quels points souhaiteriez-vous bénéficier d'une aide (de votre unité de recherche, de la DSI, du SCD, de l'URFIST, de la MMSH, de Bequali, de la PUD AMU, etc.) ?

Choisissez la réponse appropriée pour chaque élément :

	En priorité	Eventuellement	Non
Des **conseils en ingénierie d'accès et de partage** (trouver des données pour mes recherches, être informé·e sur le libre accès, négocier ma demande d'accès à des données, savoir comment diffuser ou partager mes données, etc.)			
Des **conseils méthodologiques en documentation de données** (citation des données de la recherche, description des données sous une forme normée, normes, archivage à long terme, nommage cohérent de fichiers, gestion des versions d'un ensemble de données, etc.)			
Des **conseils juridiques sur la protection des données** personnelles et des personnes enquêtées (Règlement Européen sur la Protection des données, etc.)			
Des **conseils déontologiques quant à l'usage des données** (par exemple disposer d'une charte d'utilisation des données personnelles au laboratoire)			
Des conseils pour rédiger **un plan de gestion des données**			

Q34. Auriez-vous des corpus de données, des enquêtes etc. que vous souhaiteriez diffuser/valoriser ?

1. Oui
2. Non

Q35. Si oui, Etes-vous en capacité de mobiliser du temps pour réfléchir à la manière de procéder ?

1. Oui
2. Non

4 - A propos de vous ….

Q36. A quelle tranche d'âge appartenez-vous ?

1. Moins de 30 ans
2. 30-40 ans
3. 41-50
4. 51 ans et plus

Q37. Quel est votre statut ?

1. Chercheur·e CNRS (permanent·e ou associé·e)
2. Enseignant·e-chercheur·e
3. Ingénieur·e
4. Post-doctorant·e
5. Doctorant·e

Q38. Si doctorant·e : à quelle école doctorale êtes-vous rattaché·e ?

1. ED355, Sociologie
2. ED372, Eco-Gestion

Q39. Vous êtes …

1. Une femme
2. Un homme

Q40. Depuis combien de temps exercez-vous une activité de recherche ?

1. Depuis moins de 5 ans
2. 5-10 ans
3. 11-15 ans
4. 16-20 ans
5. Depuis plus de 20 ans

Q41. Quelle est votre discipline ?

1. Sociologie
2. Gestion
3. Sciences économiques
4. Sciences politiques
5. Droit
6. Géographie
7. Anthropologie

Merci beaucoup pour le temps que vous avez consacré à compléter ce questionnaire !

11 ANNEXE 2 : CADRE METHODOLOGIQUE DE LA CONDUITE DES *FOCUS GROUPS*

Cadre

Ressources :	1 animateur·trice, 1 modérateur·trice, et 2 observateur·trices
Nombre de participants :	8-10 maximum
Lieu :	neutre, agréable, calme (visio)
Durée :	2h

OBJECTIFS

1. Identifier les représentations que les participant·es ont des problématiques de la gestion de données dans leur activité

2. Identifier les besoins en formation

3. Pour chaque besoin identifié, lister des pistes d'actions concrètes à réaliser (quoi ? quand ? où ? comment ? quels acteurs ? quel(s) public(s) ?)

COMPOSITION DES GROUPES

- Construire sans l'expliciter des groupes à
  - Dominante quanti ;
  - Dominante quali ;
  - Plutôt mixte.
- Mixer âges, sexes, disciplines.
- Faire un groupe doctorant·es sur la même base.
- Faire un groupe IT dans un deuxième temps, pour réagir sur la synthèse de l'enquête et la synthèse des groupes.
- Prérequis à la participation : lecture préalable du document de synthèse de 8 pages.

GUIDE D'ENTRETIEN

Introduction : Projet interne au laboratoire mené en deux temps.

Premier temps : enquête sur les pratiques. Rappel des conditions et des garanties.

1. Résumé rapide des principaux résultats de l'enquête : aucun jugement de valeur, pas de rappel *à priori* des pistes données en conclusion.
2. Objectif principal = comprendre plus en profondeur et ajuster l'organisation.
3. Focus group : Faire appel à leur expertise. Discussion-débat autour de 3 grandes questions : 3 X 30 mn (+ 20 mn d'introduction + 10 mn de conclusion).

Échanges d'idées, il n'y a pas de bonnes ou mauvaises réponses, débat anonyme pour connaitre leur avis, leur expérience. La séance est enregistrée avec leurs accords et sera conservée uniquement le temps de la durée de l'enquête. Les notes prises également ne seront pas nominatives.

Temps 1 : Question initiale sous forme de tour de table (question globale, création climat de confiance) :

« Comment avez-vous vécu cette enquête ? Quels ont été les points de difficulté pour répondre ? Comment recevez-vous les premiers résultats présentés ? Quelles sont vos premières réactions ? Est-ce que ça vous aide à prendre du recul sur votre rapport aux données dans la conduite des activités de recherche ? ».

Temps 2 : Question à propos des risques et des besoins : « Qu'est-ce qui vous pose particulièrement problème dans la gestion des données ? ». L'animateur·trice reprend les différents éléments qui sont cités. Lorsque tout le monde s'est exprimé et que l'animateur·trice estime avoir assez de contenu, il demande aux participant·es quels sont les éléments les plus critiques pour le laboratoire, pour eux individuellement, et dans le rapport de l'un à l'autre.

Temps 3 : À partir de chaque problème/piste d'amélioration identifiée, brainstorming des actions concrètes à mettre en place. L'animateur·trice énonce un problème/risque et au fur et à mesure du brainstorming, les participant·es peuvent réagir sur les pistes d'actions tout autour. Importance de creuser chaque piste d'actions (quoi ? comment ? qui ? quand ? où ?).

Conclusion : Synthèse du débat : l'animateur·trice demande à chacun·e ce qui le marque le plus après toute cette discussion.