



HAL
open science

Enquête sur les données de la recherche à l'université Paris-Saclay : les entretiens qualitatifs

Maximilien Petit

► **To cite this version:**

Maximilien Petit. Enquête sur les données de la recherche à l'université Paris-Saclay : les entretiens qualitatifs. Université Paris-Saclay. 2023. hal-04224366

HAL Id: hal-04224366

<https://hal.science/hal-04224366v1>

Submitted on 2 Oct 2023

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

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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Enquête sur les données de la recherche à l'université Paris-Saclay : les entretiens qualitatifs

Maximilien Petit (IGE, laboratoire CHCSC)

2023-09-25

Résumé

Retour d'expérience sur la campagne d'entretiens qualitatifs de l'enquête intitulée « Les données de la recherche à l'université Paris-Saclay, panorama et perspectives » (<https://universite-paris-saclay.hal.science/hal-03857804>).

Table des matières

| | | |
|-----|--|----|
| 1 | La méthodologie de l'enquête..... | 1 |
| 1.1 | Les avantages d'une campagne d'entretiens..... | 1 |
| 1.2 | Les obstacles d'une campagne d'entretiens..... | 3 |
| 1.3 | Les choix de notre campagne d'entretiens..... | 4 |
| 2 | Les verbatims..... | 7 |
| 2.1 | Le profil des personnes interviewées..... | 7 |
| 2.2 | Généralités sur les données de la recherche..... | 8 |
| 2.3 | Stockage des données de la recherche..... | 12 |
| 2.4 | Archivage des données de la recherche..... | 14 |
| 2.5 | Le partage et la diffusion des données de la recherche..... | 16 |
| 2.6 | Attentes et besoins concernant la gestion des données de la recherche..... | 18 |

1 La méthodologie de l'enquête

1.1 Les avantages d'une campagne d'entretiens

L'enquête organisée par l'Université Paris-Saclay ayant pour but de dresser un panorama des pratiques concernant la gestion des données de la recherche, la question s'est rapidement posée de savoir s'il était pertinent, en plus de l'enquête quantitative, d'organiser une campagne d'entretiens qualitatifs. Ce travail de terrain étant avant tout une commande institutionnelle, il s'agissait de faire remonter les besoins avant la mise en place d'une offre ciblée sur la question des données au sein des unités de recherche. Le groupe de travail de l'enquête, composé de personnels qui relèvent de l'Information Scientifique et

Technique (IST), a mis en évidence plusieurs avantages que l'on peut tirer de la pratique des entretiens :

- Il y a un intérêt à croiser les données d'une enquête quantitative avec celles qui proviennent d'une campagne d'entretiens qualitatifs. Cette triangulation a déjà été utilisée pour d'autres enquêtes sur le sujet au sein des universités (Lille 3 en 2015, Rennes 2 en 2017, Bordeaux Montaigne en 2018, Paul-Valéry Montpellier 3 en 2019, Aix-Marseille en 2019) et les résultats étaient pertinents à consulter ;
- Les entretiens permettent notamment de faire remonter efficacement les problématiques des communautés scientifiques qui relèvent des sciences humaines et sociales (SHS). Les enquêtes sur la gestion des données de la recherche mettent très souvent en avant le fait qu'il y a d'abord un problème de vocabulaire, de terminologie, et que les définitions institutionnelles des données de la recherche ne coïncident pas nécessairement avec les pratiques que l'on observe en détail chez les enseignants-chercheurs¹. Rencontrer des chercheurs au cours d'entretiens, et aborder la question des matériaux de recherche, des données, c'est mettre en lumière des terminologies, des pratiques, des discours et des attentes qui sont riches d'informations ;
- La question des données de la recherche au sein des universités n'est pas un sujet nouveau, mais il reste difficile à appréhender : il y a une actualité juridique, scientifique et institutionnelle mouvante. Les pratiques sur la question sont diverses au sein des disciplines, et il est intéressant pour un établissement de mieux connaître ce qui existe en amont avant de mettre en place un dispositif support et une feuille de route institutionnelle ;
- La mise en place d'entretiens qui sont menés par des personnels IST est souvent un bon moyen pour créer un premier contact avec les communautés scientifiques. La question des données de la recherche fait appel à des compétences que l'on retrouve dans de nombreux métiers au sein des universités, il y a toute une dynamique de collaborations à construire : chercheurs, IST, service juridique, service informatique, etc. Ces entretiens peuvent être une porte d'entrée vers d'autres projets précieux ;
- Organiser une collecte de données issues d'entretiens qualitatifs, c'est également mettre à l'épreuve une organisation de travail et expérimenter le sujet que l'on ambitionne d'interroger : le cycle de vie des données de la recherche. C'est se confronter à la planification, la collecte, l'analyse, la curation, le dépôt, la publication et la réutilisation de données particulières². Pour les entretiens, notamment, c'est

¹ Marie-Laure Malingre et al., "Construction(s) et contradictions des données de recherche en SHS," *Recherche d'information, document et web sémantique* 2, no. 1 (2019), <https://doi.org/10.21494/ISTE.OP.2019.0336>.

² Le cycle de vie de la donnée selon la plateforme Recherche Data Gouv : <https://recherche.data.gouv.fr/fr/page/quelles-donnees-de-recherche>

rationaliser la collecte, la transcription, l'analyse et la documentation d'une méthode qui nous semble pertinente.

1.2 Les obstacles d'une campagne d'entretiens

Et il y a bien évidemment des obstacles que nous avons identifié en amont de la campagne d'entretiens qualitatifs :

- Pour tirer pleinement parti de la triangulation entre les dispositifs quantitatifs et qualitatifs, il s'agissait également de se questionner sur la stratégie d'échantillonnage : quelles sont les personnes à interroger qui seraient susceptibles d'apporter des informations que nous ne pouvons pas tirer d'emblée de l'enquête quantitative ?
- La démarche de l'entretien qualitatif ne s'improvise pas, et le groupe de travail étant composé de personnels IST, il s'agit également de se soucier de la formation des personnes mobilisées sur le projet. Il y a effectivement des pratiques attendues qui relèvent de la démarche de l'entretien qualitatif : la posture de l'intervieweur, l'importance des questions ouvertes, de la « mise en train », la co-construction de sens, le respect de la parole de l'interviewé, etc. Toutes ces notions, explicitées dans les travaux sur les méthodes³, sont importantes à connaître pour garantir la bonne mise en œuvre du dispositif ;
- L'objectif de départ concernant les entretiens était de formuler, via les résultats et l'explication méthodologique, un objet suffisamment proche de la recherche, un outil d'aide à la décision, et un objet de valorisation des pratiques pour les unités de recherche qui souhaiteraient s'emparer du dispositif. Mais avoir une approche triple nécessite de mobiliser des personnels avec des profils divers (IST, services, unités de recherche), et implique une charge de travail conséquente dans un contexte sanitaire qui ne facilite en rien les choses ;
- En effet le covid et les différents protocoles de travail au sein des établissements ont eu un impact sur le déroulement des entretiens qualitatifs. De fait, les entretiens en visioconférence étaient privilégiés pour des raisons de praticité. Cette modalité de travail à distance était envisagée en amont et fut prise en compte concernant la prise de contact avec les personnes interviewées, la question du dispositif technique d'enregistrement et la rédaction d'un guide d'entretien. La relation interpersonnelle n'était certes pas la même que pour des entretiens en présence, mais il ne nous semblait pas pertinent d'exclure ce type d'entretiens.

³ Gérald Boutin, *L'entretien de recherche qualitatif. Théorie et pratique* (Presses de l'Université du Québec, 2018).

1.3 Les choix de notre campagne d'entretiens

Nous indiquons ici les différents choix effectués durant la campagne. Nous souhaitons faciliter l'appropriation des résultats des entretiens qualitatifs et permettre aux établissements qui souhaitent également effectuer une enquête sur les données de la recherche de bénéficier d'un retour d'expérience. Nous abordons la stratégie d'échantillonnage, la méthodologie d'entretien, le stockage et le traitement des données.

Concernant la stratégie d'échantillonnage, il s'agissait de partir d'un choix simple : un item de l'enquête en ligne était un appel aux volontaires souhaitant poursuivre l'échange au cours d'un entretien. Sur les cinq cent douze réponses de l'enquête en ligne, quarante et une personnes étaient volontaires pour participer à un échange : un personnel IST, douze chercheurs, dix doctorants, quatorze enseignants-chercheurs, trois ingénieurs et un technicien de recherche. Sur ces volontaires, onze personnes travaillent dans le champ des sciences du vivant et de l'environnement, dix en sciences humaines et sociales, vingt en sciences et technologies. Enfin, ces personnes proviennent de trente et une unités de recherche au sein de dix établissements différents de l'Université Paris-Saclay. Cette base de volontaire nous semblait être suffisamment diversifiée en termes de profils et de disciplines.

Le groupe de travail devait néanmoins se fixer un objectif chiffré à atteindre concernant les entretiens pour plusieurs raisons : la campagne devant se dérouler entre septembre et décembre 2021, les sollicitations sont très nombreuses pour les communautés scientifiques, et également pour les personnels IST qui composent le projet. Il est difficile, pour ce qui est de la charge de travail, de se fixer un objectif trop ambitieux. L'idée première restait de couvrir des entretiens avec des profils diversifiés en termes d'approche disciplinaire, mais de ne pas couvrir trop d'entretiens au risque d'éprouver ensuite des difficultés pour traiter les données d'enquête, mais également d'avoir une saturation d'informations. Un objectif initial de trente entretiens, ne devant pas dépasser une durée de quarante-cinq minutes, a donc été posé.

En tout, finalement, il y a eu vingt-quatre entretiens, et donc, seize heures et six minutes d'enregistrement audio à traiter. Sur ces vingt-quatre entretiens, cinq personnes travaillent dans le champ des sciences du vivant et de l'environnement, onze personnes en sciences et technologie, huit personnes en sciences humaines et sociales. Ces vingt-quatre personnes proviennent de sept établissements différents de l'Université Paris-Saclay : le Commissariat à l'énergie atomique et aux énergies alternatives (CEA), CentraleSupélec, le Centre national de la recherche scientifique (CNRS), l'Office National d'Études et de Recherches Aérospatiales (ONERA), l'Université de Versailles Saint-Quentin-en-Yvelines, l'Université d'Évry et l'Université Paris-Saclay. Enfin, ces vingt-quatre personnes exercent dans une vingtaine d'unités de recherche.

Le choix a été fait d'intégrer quelques personnes qui ne se trouvaient pas dans l'échantillon de départ pour plusieurs raisons : pallier aux désistements de volontaires ne pouvant plus se libérer, et s'entretenir avec des profils ayant une approche particulière concernant la gestion des données de la recherche. C'est notamment le cas d'ingénieurs d'études et

d'ingénieurs de recherche, en laboratoire, qui ne sont pas suffisamment interrogés sur cette question.

D'un point de vue méthodologique, il nous semblait judicieux d'effectuer des entretiens de type semi-structuré. Le but étant de permettre aux personnes interviewées de se sentir libres d'exprimer leur ressenti, leurs attentes et leurs difficultés concernant la gestion des données de la recherche. Un guide d'entretien a été conçu en amont par le groupe de travail, et devant permettre de nouer des échanges concernant les différentes catégories utilisées dans l'enquête en ligne : le profil du répondant et l'enquête en ligne, les généralités sur les données de la recherche, le stockage des données, l'archivage des données, le partage et la diffusion des données, les attentes et besoins concernant la gestion des données. Cette « trame » de questions de départ et de questions de relance devait permettre d'effectuer des entretiens n'excédant pas les quarante-cinq minutes. Cet objectif a été atteint puisqu'en moyenne les entretiens se déroulaient sur une quarantaine de minutes. Ce guide d'entretien a été utilisé collectivement par les personnes du groupe de travail impliquées dans le déroulement des entretiens. Des indications méthodologiques (la méthode de l'entretien semi-structuré) et techniques (dispositif d'enregistrement) apparaissent sur le document pour harmoniser les usages des intervieweurs qui opéraient en binôme : une personne pour l'animation de l'entretien, une personne pour la prise de notes. Il n'est pas aisé de revêtir la posture de l'intervieweur, notamment lorsque les personnes impliquées n'ont pas nécessairement une expérience en la matière. Les manuels de méthodes, conseils de chercheurs des sciences sociales et les différentes formations, notamment celles de l'*Open Science Week* du réseau des URFIST (Unité Régionale de Formation à l'Information Scientifique et Technique) permettent tout de même de se préparer correctement.

Il n'a pas été question de réaliser des transcriptions intégrales d'entretiens : la charge de travail ne pouvant être supportée par les membres du groupe. Néanmoins, sur la base des enregistrements et des prises de note, une personne impliquée était chargée de transcrire les extraits significatifs et de les catégoriser en utilisant le logiciel *MAXQDA Analytics Pro* accessible grâce au soutien du dispositif DLab SHS de l'UVSQ⁴. Pour bénéficier néanmoins d'un fichier de départ correctement horodaté, différents tests furent effectués en amont concernant les solutions de transcription automatique, notamment les outils abordés dans une publication du réseau Mate-SHS⁵. Aucun outil n'a été retenu, mais seulement pour des raisons pratiques : coût d'utilisation trop important, qualité des transcriptions pas suffisante au regard du calendrier de travail contraint. Afin de compenser l'absence de transcription intégrale, et pour faciliter le traitement des seize heures de données par une seule personne, une grande partie des prises de notes par les binômes furent effectuées en adoptant le modèle de la fiche chrono-thématique. L'entretien est découpé en séquences dans un tableur et regroupe de brèves analyses. Ce modèle permet à la personne chargée du

⁴ DLab SHS porté par les laboratoires CHCSC et DYPAC : projet soutenu par la Région Ile-de-France dans le cadre du Domaine d'intérêt majeur « Matériaux anciens et patrimoniaux ».

⁵ Elise Tancoigne et al., "La transcription automatique : un rêve enfin accessible ?" Rapport de recherche (MATE-SHS, 2020), <https://halshs.archives-ouvertes.fr/halshs-02917916>.

traitement de se repérer rapidement dans le fichier audio et de distinguer plus efficacement les extraits significatifs à catégoriser ensuite dans un logiciel d'aide à l'analyse qualitative CAQDAS (*Computer-Assisted Qualitative Data Analysis Software*).

Certaines tâches liées au traitement des données d'enquête ne sont pas visibles en amont, nous les consignons ici pour le retour d'expérience. Il convient par exemple de réfléchir précisément au dispositif d'enregistrement, que ce soit pour des entretiens en présence ou à distance. Deux modalités retenues en amont pour notre campagne : un dispositif « léger » comportant un enregistreur Olympus LS-P4, et un dispositif « lourd » avec un enregistreur Zoom et deux micros XLR. Finalement, le dispositif « léger » a été le plus souvent utilisé, car simple à mettre en place, peu déstabilisant pour les intervieweurs et les interviewés. En l'absence de matériel, les fonctionnalités d'enregistrement de téléphones portables furent également utilisées. Pour ce qui est des entretiens à distance, toutes les solutions utilisées (principalement Zoom) permettent d'enregistrer l'audio des entretiens. On se retrouve donc avec différents formats de fichier audio : du MP3, MP4, du WMA. Dans le cadre d'un traitement de données, il y a un intérêt à harmoniser la qualité d'enregistrement et il y a des tâches que l'on n'envisage pas en amont, ne serait-ce que le « nettoyage » des fichiers audio : éliminer les bruits parasites, gérer l'amplification. Ces tâches sont d'autant plus importantes si l'on souhaite utiliser à bon escient les solutions de transcription automatique. Pour les entretiens à distance, les solutions de visioconférence comme Zoom traitent au préalable le fichier audio, mais les résultats sont très médiocres pour l'horodatage des fichiers, ce qui complique beaucoup le travail par la suite dans un logiciel CAQDAS. Ce type de logiciels, dans le cadre de notre campagne d'entretiens, permet de travailler aisément les fichiers, de signaler et catégoriser les extraits, d'extraire le texte et l'audio et de produire automatiquement des rapports. Le gain de temps pour le traitement et l'analyse n'est pas négligeable, et le logiciel nous permet de produire un jeu de données pouvant être réutilisé par la suite par les communautés scientifiques intéressées.

Nous produisons dans ce document une sélection seulement des verbatims des entretiens et proposons, sur Recherche Data Gouv, la totalité des passages : trois cent soixante-six segments codifiés. Les codes reprennent à la fois les catégories du guide d'entretien, et donc, de l'enquête en ligne. Les résultats font apparaître un déséquilibre sur certaines catégories. La thématique de l'archivage, notamment, n'est pas réellement fructueuse. Les questionnements sur ces sujets étaient déstabilisants pour une grande partie des personnes interviewées de toutes les disciplines. La thématique « Généralités » rassemble le plus d'extraits car elle permet aux personnes interviewées d'explicitier aisément leurs matériaux de recherche, et d'aborder des points importants à propos de la gestion des données : définitions, terminologies, données sensibles, plans de gestion des données. La catégorie consacrée au « stockage des données » laisse apparaître à la fois des stratégies personnelles et institutionnelles diversifiées concernant cet aspect. Là où, par exemple, les échanges de la partie dédiée au « partage des données » génèrent des réactions très marquées. Enfin, pour ce qui est de la catégorie des « attentes », elle nous renseigne de façon pertinente, mais on regrette le placement en fin d'entretien. Le guide d'entretien que nous proposons également sur Recherche Data Gouv est bien équilibré, mais il gagnerait à être retravaillé si d'autres projets d'enquête souhaitent s'en emparer. Disposer d'un tel document nous semble

essentiel si le travail d'enquête est collectif, il est également une aide non négligeable pour celles et ceux qui ne sont pas familiarisés avec les techniques d'entretien.

2 Les verbatims

2.1 Le profil des personnes interviewées

On interroge ici les volontaires sur leur terrain de recherche, sur le niveau de connaissance qu'ils pensent avoir sur la question des données et sur la compréhension des items de l'enquête en ligne. Ce dernier aspect est difficilement exploitable étant donné la période assez longue entre l'enquête en ligne et la campagne d'entretiens. Les profils et les champs disciplinaires sont diversifiés et rares sont les interlocuteurs qui pensent disposer d'un niveau de connaissance élevé par rapport à la question des données de la recherche.

« Donc, je travaille sur Zola dans une perspective transnationale. Je m'intéresse surtout à la réception de ses œuvres en ce moment dans l'espace soviétique, puisque ma thèse portait sur la réception de Zola des origines en RDA. » ;

« Très faible, parce que je ne sais pas ce que c'est vraiment les données de la recherche. Après, j'utilise quand même beaucoup internet » ;

« Je dirais intermédiaire. Un niveau intermédiaire. On va dire 3 sur 5, quelque chose comme ça. C'est vrai que je ne suis pas très au fait, mais étant en lien beaucoup avec les collègues de la BU, menant un cours d'initiation à la recherche depuis plusieurs années avec les étudiants de master et ayant suivi quelques formations à l'URFIST aussi sur la science ouverte, etc. C'est vrai que je suis quand même assez averti, mais c'est vrai que les sciences humaines ne sont pas aussi développées je pense que d'autres champs. » ;

« Je travaille sur le domaine de chasse de Rambouillet. Plus largement, les domaines des princes, dont d'autres exemples sont Chambord et Marly, je travaille en histoire environnementale, en humanités environnementales, sur ce qu'on appelle le gouvernement de la nature. Donc comment on gère la nature depuis trois siècles dans ces lieux-là en lien et puis autour des activités de chasse, d'élevage, donc d'agriculture et puis de foresterie. » ;

« Mon métier principalement, c'est un métier associé à la réalité virtuelle. » ;

« Nous faisons principalement des essais de caractérisation de matériaux, ce que certains appellent de la RDM (Résistance des matériaux), mais de manière on va dire, soit sur des matériaux nouveaux, soit avec des techniques nouvelles d'investigation pour optimiser les essais et les numériser et les simuler de manière à réduire les coûts de réalisation d'essai et d'avoir des calculs qui nous permettent de simuler la tenue des différents matériaux dans différentes conditions d'utilisation. » ;

« Je suis ingénieur de recherche à l'ONERA. Et du coup, je travaille essentiellement sur des problématiques pour les satellites, l'interaction entre du plasma. Soit plasma environnement, soit du plasma lié à des propulseurs de satellites » ;

« J'enseigne dans les systèmes, les réseaux et l'informatique et mes recherches portent principalement sur les interfaces cérébrales pour l'assistance aux personnes en situation de handicap. Et donc du coup on utilise des enregistrements de données cérébrales. Et moi, je travaille particulièrement sur des approches mathématiques pour traiter ces signaux cérébraux. » ;

« Donc je fais de la recherche en physique expérimentale sur la stabilité d'interfaces liquides donc je travaille sur en fait des objets du quotidien que sont les bulles de savon, les films de savon, les mousses, voilà. Donc une grosse partie de ma recherche, pour ne pas dire l'intégralité, est faite en collaboration avec des thésards et donc c'est essentiellement eux qui prennent les points de données et qui accumulent les données et donc pour moi la question des données de recherche c'est beaucoup la question de la sauvegarde des données de thèse. » ;

« Je suis ingénieur d'études en instrumentation, donc je suis spécialisé dans les instruments de recherche. Donc j'essaie d'adapter l'instrument et la façon de l'utiliser aux problématiques des chercheurs » ;

« Je suis chimiste, biochimiste de formation, mais je me suis très vite tourné au cours de ma thèse vers la physique, pour la biologie, les méthodologies physiques pour l'analyse des macromolécules. Et en particulier, je fais énormément de spectrométrie de masse et j'ai toujours développé les logiciels moi-même pour l'analyse de masse, donc je suis un informaticien chevronné pour l'informatique scientifique. Et donc je suis très sensible aux données et au traitement des données ».

2.2 Généralités sur les données de la recherche

Cette catégorie génère le plus d'extraits significatifs. On demande aux personnes interviewées de situer leur terrain de recherche par rapport à la problématique des données et de leur gestion. Des sujets importants s'y agrègent : définitions, terminologies, données sensibles, standards de métadonnées, rédaction de plans de gestion des données. On remarque, pour les sciences humaines et sociales, des objets de recherche transdisciplinaires, une pratique de la recherche solitaire, et des questionnements récurrents sur la « valeur » des données qu'ils collectent, réutilisent ou génèrent dans leurs disciplines. Concernant la rédaction de plan de gestion de données, la pratique n'est pas encore suffisamment répandue et, lorsqu'elle est connue, elle est perçue comme une tâche administrative en plus. Ou alors, le PGD vient confirmer une pratique méthodologique déjà ancrée dans la discipline en question. Plus généralement, il apparaît difficile d'appréhender une gestion globale des données de la recherche, tant il y a des particularités disciplinaires très marquées en termes d'outils et de protocoles. Et ces particularités vont également dépendre de l'objectif final que l'on associe aux données : publication, livrable pour un projet, commande d'un client, etc. Enfin, sans surprise, le « temps » est souvent abordé par les chercheurs lorsqu'il s'agit d'évoquer la gestion des données de recherche. Il manque toujours, et il semble être en décalage avec celui des institutions et financeurs. Difficile de savoir où situer un temps de gestion des données, et même un temps de formation sur ces questions : des tâches qui vont s'agréger dans le temps administratif, pédagogique ou de recherche ?

« Pas trop, parce que nous, parce que ce truc, c'est un truc d'historiens ça. Parce que nous les sources, ce n'est vraiment pas un terme qu'on utilise, pas du tout, du tout, du tout. On va parler de documents, d'archives, et pour nous effectivement c'est des données sur lesquelles on peut s'appuyer. On s'appuie aussi sur des œuvres théoriques. Et puis après, on va produire nos propres réflexions et si je comprends bien, les deux sont des données de la recherche, à la fois les choses sur lesquelles on s'appuie, que ce soit des archives, des enregistrements, des œuvres théoriques. » ;

« Je n'ai jamais vraiment eu l'impression de trouver quelqu'un qui mesurait la valeur de ce que j'avais et la fragilité de la chose. » ;

« On peut faire de l'analyse effectivement directement sur la partition de tout ce qu'on appelle l'analyse paramétrique, donc l'analyse de mélodies, l'analyse harmonique, l'analyse rythmique. Donc, soit sur une partition qui est déjà existante, que l'on questionne. C'est à dire, est ce que c'est une production qui a lieu avant l'œuvre ou une production qui a eu lieu après l'œuvre, donc une partition d'écoute. Soit c'est une partition que l'on fabrique nous-même, donc des relevés d'oreille, que l'on peut établir à partir des écoutes que l'on peut faire d'une œuvre » ;

« *Finale*, c'est vraiment l'éditeur de partitions historique qui s'est développé. Maintenant, on a des logiciels *open source* type *MuseScore* qui fonctionnent très, très bien, sur lequel on forme les étudiants » ;

« Donc, on a peu de données. En fait, on peut s'échanger des articles, mais on ne travaille pas souvent sur les mêmes objets. On a assez peu de choses à partager. L'esprit collaboratif se trouve sur des co-organisations de colloques, sur des moments, sur des gestions de cours plus que sur des objets de recherche en eux-mêmes » ;

« Chacun travaille un peu dans son couloir quand même, sauf exception, donc on n'est pas amené en fait à avoir une réflexion globale sur la manière dont on traite les données » ;

« Et puis le deuxième pan, c'est le terrain vivant actuel donc à la fois la forêt en tant qu'écosystème, et puis les humains qui y habitent. Donc là, il y a des données qui sont mouvantes, voilà la présence d'animaux ou d'espèces. Et puis des données avec lesquelles on peut plus dialoguer, donc des humains qui sont là à fournir leurs données de vie, donc là, il y a tout un enjeu aussi d'éthique, de savoir respecter leurs données puisqu'ils sont plus que ça. » ;

« Alors, qu'est-ce qu'on manipule comme données ? On manipule des données numériques, on manipule aussi des données verbatims et conceptuelles, à savoir soit des observations chiffrées ou des observations littérales. Et puis il y a aussi des données qui sont un peu plus abstraites, qui sont des données conceptuelles. Alors je ne sais pas si on peut parler réellement de données, mais en tout cas il y a des notions conceptuelles qui sont importantes à formaliser. Aussi à déposer quelque part et à transmettre et ça, c'est quelque chose qui est plus compliqué à partager. » ;

« Plan de gestion de données. Alors ça me parle, mais pas sous ce terme-là, mais c'est plus comment est-ce qu'on pratique ses sauvegardes, comment est-ce qu'on manage ses données, c'est à dire comment est-ce qu'on les intitule, où est-ce qu'on les range, c'est ça que vous voulez dire ? » ;

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

« On produit aussi en interne nos propres données. De toute façon l'ONERA est un établissement, c'est un EPIC, donc on a différents types de clients. Le but est, contrairement par exemple au CNRS, ce n'est pas uniquement de la production scientifique, ça peut être aussi de répondre d'un côté à des industriels, de l'autre côté à des établissements ou encore à des agences spatiales. À ce titre-là, donc on a différents types de données. Certaines sont publiées, d'autres ne sont pas publiées. » ;

« Pour moi les données de la recherche seraient les données qui sont collectées lors d'expérimentations et puis qu'on peut du coup mettre à disposition. Et pour moi, ça irait même au-delà des mesures expérimentales, et ça inclut aussi, puisque j'y suis sensibilisé, toutes les chaînes de traitement, tous les outils logiciels, toutes les bibliothèques. Enfin voilà, c'est très bien d'avoir les données, c'est très bien d'avoir les données traitées, mais si je n'ai pas la capacité de refaire le cheminement pour arriver aux données traitées, ce n'est pas vérifiable et pour moi du coup c'est moins intéressant quoi. » ;

« Généralement on n'a pas le genre de la personne, on n'a pas ses informations personnelles, tout est anonymisé. Mais du coup, ça pose d'autres biais aussi parce qu'on se retrouve à étudier finalement des interactions homme-machine et le fait que ça soit un expérimentateur ou une expérimentatrice qui s'adresse à un sujet, homme ou femme, ça a des conséquences sur la passation des expériences. » ;

« Oui, alors du coup, pour les données cérébrales, il y a un standard de données qui s'appelle BIDS et qui est vraiment le standard qui permet à la fois de récupérer les données, mais aussi toutes les métadonnées et toutes les informations parce qu'on a fait passer des expériences à des personnes en les installant dans un dispositif avec un certain protocole et du coup l'intérêt du format, c'est que ça permet d'avoir une façon de structurer les données qui soit interopérable, explicite, faite par les chercheurs. » ;

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

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

« Mais ce n'est pas qu'une question de connaissances, je pense qu'il y a des questions de manque matériel. Donc ça, éventuellement, je sais ce qu'il me faudrait, mais je l'ai pas quoi. Et il y a des questions de manque de soutien technique, informatique, etc. À l'intérieur du

labo, avec des collègues qui sont complètement surchargés par le support informatique et voilà. Et des problèmes de fond qui sont... enfin la question des données ouvertes, qu'est-ce qu'on ouvre comme données, et avec quel intérêt ? [...] Et si oui, quelles informations au minimum il faut ? [...] Donc j'ai l'impression que c'est quand même un gros travail, même de le faire. Enfin, si pour chaque donnée, il faut que je mette ensemble toute cette question-là ailleurs que dans un cahier de manip, c'est un travail qui est énorme avec une utilité qui me paraît proche de 0 parce que je ne vois pas ce qu'un autre chercheur pourrait en faire et quelqu'un qui pourrait en faire quelque chose, je pense qu'il faut qu'il me contacte. » ;

« Donc moi mon problème, c'est le rapport entre le temps à y passer et l'efficacité que ça a. Je pense que l'efficacité est très faible par rapport au temps qui est très grand. » ;

« On est toujours en tant que chercheur, producteur de données et utilisateur de données, à la recherche d'espace de stockage, que ce soit bien conçu, qu'on puisse trouver les données et aussi je dirais un aspect assez crucial de mon point de vue une forge intelligente performante de type *GitLab* pour gérer les logiciels quoi. Et moi je suis un friand utilisateur de Renater et en particulier de SourceSup. » ;

« Les données de la recherche, qu'elles soient confidentielles ou pas, parce qu'il faut bien voir que les données ont une vie qui est polaire. La polarité de cette vie est très simple, elles naissent confidentielles et peu à peu elles deviennent non confidentielles puisque les données, une fois qu'elles sont publiées, elles ne sont plus secrètes et on peut les ouvrir à tout le monde. » ;

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

« Mais pour ce qui concerne le cycle de vie que j'évoquais avant, oui, je pense qu'il faudrait pouvoir assurer un mécanisme d'ouverture progressive. » ;

« Nous on appelle ça des ontologies. Donc, dans notre discipline, il y a ça, absolument. Je m'en sers tout le temps parce que dans mon logiciel j'ai besoin de m'en servir parce que c'est un vocabulaire spécifique, précisément, parce que... ce que je disais tout à l'heure pour les métadonnées : définir quel type de fragmentation, est ce que c'est de l'impact électronique, est-ce que c'est de la collision, est ce que c'est de la capture d'électrons, etc. Les énergies... tout ça ce sont des paramètres qui sont sous contrôle d'ontologies. » ;

« On a un projet ANR, on va avoir beaucoup d'argent pour faire des manips qui coûtent très cher, qui passent par des plateformes, etc. En général, le temps du projet ANR, c'est le temps de production des données. » ;

« Finalement ce n'était pas si compliqué. On a utilisé le truc, Opidor là. Je pense que si on avait eu un service d'appui qui aurait pu relire par exemple, on aurait apprécié. » ;

« En tout cas, ça ne fait pas partie de notre formation, on ne nous a pas dit que c'était à nous de gérer ou de réfléchir à la manière dont c'était géré derrière. Après, qu'on participe à ce

travail, je pense que, quelque part, c'est une évolution assez logique puisqu'il faut qu'on soit au courant de ce qui se passe et de ce qu'on fait avec les choses qu'on produit. » ;

« J'ai vécu toute cette période où en fait il y avait une vision plus ouverte, plus accessible de la recherche sans forcément qu'il y ait cette notion de paiement systématique pour accéder au fruit du travail des chercheurs. » ;

« Dans les labos académiques, pour caricaturer, les notions de protection sont beaucoup moins drastiques que dans les contextes du privé. » ;

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

2.3 Stockage des données de la recherche

On questionne ici les pratiques de stockage, leur évolution dans le temps, les pertes de données et la responsabilité de cet aspect bien particulier de gestion. La peur de perdre de données favorise des stratégies de stockage : on multiplie les supports. La sauvegarde, elle, est rarement abordée (par manque de connaissances ?). Lorsque l'établissement est en mesure de fournir une solution, cet aspect de gestion passe au second plan et n'est plus questionné. Lorsqu'il s'agit de l'unité de recherche, le stockage et la sauvegarde sont des tâches qui ont leur importance et qui sont confiées à des ingénieurs. Par contre il y a des problématiques qui émergent concernant la souveraineté des données et leur stockage, l'usage de solutions libres et la consommation des ressources. Les récits de perte de données, plus qu'un dysfonctionnement du matériel de stockage, sont liés à des jeux de données que l'on ne documente pas et à une perte de compétences (personnel en charge du stockage et de la sauvegarde ne travaillant plus dans l'unité).

« Moi, j'ai tous mes entretiens vidéo, audio, dans mon ordinateur et je sauvegarde en utilisant une... enfin, je suis un peu folle donc j'ai une *DiskStation* chez moi et j'ai un jeu de clés USB que j'utilise en rotation et j'en ai une dans mon sous-sol qui ne peut pas brûler. Et j'ai aussi un abonnement à Google One. » ;

« Complètement, pour des raisons financières, parce que ça bouge beaucoup. À chaque fois, il faut perdre beaucoup de temps. C'est une question de se mettre... de se garder à jour sur ces questions. C'est un temps que j'ai de moins en moins. Donc, je pense qu'il y a toujours mieux que ce que je fais, que j'ai toujours un temps de retard par rapport aux méthodes de stockage et je pense que si je passais par une personne à mon université responsable de ça, cette personne serait vraiment *up to date* sur toutes ces questions. » ;

« Alors si vous voulez un chiffre sur la quantité de stockage dont j'aurais besoin là, rien qu'avec ce que j'avais pris en photo dans le cadre de mes recherches doctorales, sans les autres, je n'ai pas regardé, j'en suis à 203 Go » ;

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

« Je cherche à utiliser au maximum le libre précisément dans le cas où l'un des serveurs que j'utilise deviendrait obsolète. » ;

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

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

« Il y a un serveur imagerie, des serveurs cytométrie, des serveurs pour tout ce qui est bureautique. Mais tout ça, c'est transparent pour nous, pour les utilisateurs. » ;

« Oui, tout ce qui concerne le terrain que j'ai eu à faire depuis 2019-2020, j'ai une triple copie, c'est-à-dire dans la clé USB, dans la machine et dans le *Google Drive*. » ;

« On nous demande que toutes les données soient enregistrées soit sur un PC de manip qui est sur place et après on reverse toutes ces données sur le réseau interne où là il y a des sauvegardes régulières. Pour nous c'est une boîte noire, on a accès aux données, mais on ne sait pas où elles se trouvent physiquement. Ces données-là, en aucun cas, on ne doit les diffuser à qui que ce soit. Hormis en interne du laboratoire puisque, étant technicien, moi je rends mes données aux ingénieurs qui eux les traite, les font évoluer et en tirent des calculs différents. » ;

« Je travaille sur dix ordinateurs différents [...] donc il y a tout un art de ne pas se tromper quand on passe d'un ordinateur à l'autre, d'un endroit à l'autre. » ;

« Non, il ne faut pas que l'université aille donner plein d'argent à un truc privé qui va encore nous entuber. Je suis complètement contre. » ;

« On a l'obligation d'expliquer à notre employeur quelle méthode de stockage on utilise pour les données. Mais après, on n'a pas d'homogénéité derrière, on a un certain nombre d'options, disons. On n'a pas une très forte homogénéité dans le stockage des données, mais on a différentes options de stockage. » ;

« Après on a plus des problèmes, et je pense comme dans beaucoup d'endroits, de perte de compétences. C'est-à-dire que les données peuvent être là, mais en l'occurrence si on n'a plus forcément la personne pour expliquer les données, derrière, dans ce cas-là, on a de la perte de compétences, et on a finalement un temps qui est nécessaire pour se réappropriier les données. » ;

« Et après nous, notre difficulté, c'est d'avoir autre chose que des petits disques durs qui se triment partout. C'est-à-dire qu'en fait, il n'y a pas grand-chose qui est centralisé au laboratoire. C'est vraiment le serpent de mer depuis dix ans de dire "comment est-ce qu'on fait avec des services informatiques qui sont objectivement débordés ?". Les projets de fond, quand on est débordé par le quotidien, c'est super compliqué pour eux. » ;

« La gestion du stockage, des moyens de calcul, elle est vraiment faite collégalement et on prend de l'argent un peu au labo je dirais, plus qu'à des projets. » ;

« Cela dit, si j'avais une forge ou même un dépôt de données bien conçu, moi je ne serais pas du tout contre pour mettre ça plutôt ici. Parce que moi je suis assez pour le fait de rendre l'Europe assez autonome par rapport aux autres. » ;

« Je fais tout en local à part pour les collaborations, donc souvent on s'écrit des notes, des trucs comme ça, qu'on s'envoie où on met des papiers très importants dans un coin. Et ça, la plupart du temps, depuis que je fais de la recherche, on a toujours utilisé *Dropbox*. » ;

« Je trouve qu'en local, ça a un intérêt. Disons que les machines de stockage ne coûtent pas trop cher en général, en tout cas maintenant. Si quelqu'un sait le faire, et cetera, je pense que ça ne serait pas une mauvaise idée que ce soit géré au niveau local du labo par exemple. Par contre je ne pense pas que ça serait bien que ce soit fait par des services... parce qu'il y a des gens qui mettent tout sur le cloud Google machin. Juste par rapport à la consommation de ressources, je trouve ça un peu bête. Alors qu'en local... évidemment ça consomme des ressources, de l'énergie, mais ça me semble bien au moins pour du court terme. » ;

« On a dans notre unité de recherche... on a la chance, enfin on a eu cette volonté... on a une équipe système informatique. C'est-à-dire, il y a un technicien, un assistant-ingénieur, qui s'occupent du développement, du maintien de tout notre système informatique. En particulier, ils gèrent des serveurs de stockage. » ;

2.4 Archivage des données de la recherche

On questionne ici les personnes interviewées sur l'archivage de leurs données et sur leur connaissance des infrastructures et des métiers qui interviennent sur ce point. C'est la catégorie du guide d'entretien qui génère le moins d'extraits. Beaucoup de personnes interviewées, de toutes les disciplines, perçoivent assez peu la différence entre le stockage et l'archivage des données, et n'ont pas d'échanges avec des archivistes ou de connaissances concernant les infrastructures telles que le CINES. La pérennité des données de la recherche semble être un sujet rarement abordé en amont de la recherche, c'est avant tout le court terme du projet qui prédomine. S'il devait y avoir un interlocuteur sur ces questions, l'échelle du laboratoire/de l'unité de recherche serait à considérer.

« Mais voilà, c'est vrai que je n'ai pas encore réfléchi à ces problématiques d'archivage par rapport en tout cas à ma politique de stockage adoptée. C'est quelque chose que je n'ai pas encore envisagé. » ;

« Alors, il n’y a pas d’archiviste au sens strict du terme, c’est plus des informaticiens qui sont là pour faire la sauvegarde de ce qu’il y a sur le serveur. » ;

« Un projet, il a un temps de vie. Et quand le projet est terminé, on a encore beaucoup de mal à récupérer ce que l’on a fait dans le projet pour le réutiliser après, voilà. On n’est pas bon là-dessus. On ne le fait pas bien, c’est-à-dire quand le projet est fini, bah le truc, après, on le jette un peu à la poubelle hein. Donc le projet, ça peut durer cinq ans, ça peut durer dix ans. Voilà, mais je n’appelle pas ça du long terme. » ;

« C’est un problème d’indexation parce qu’après c’est de la métadonnée. On va très vite vers des fichiers qui sont conséquents et un nombre de fichiers assez astronomique. Alors, tout vouloir conserver, c’est bien. Après, est-ce que ce sera exploitable un jour et de quelle manière, j’en sais rien. » ;

« En fait, je pense que de manière générale, il y a énormément d’éléments qui doivent être archivés et qui ne le sont pas actuellement. Il nous arrive parfois d’avoir des redondances sur certaines études. Je pense qu’à ce niveau-là, ça devrait, disons, être finalement un réflexe, et quasiment une obligation, à la fin de chaque étude, de réfléchir à la manière d’archiver ses données, juste pour ne pas perdre de temps et de l’argent, et des moyens humains sur simplement le fait de refaire plusieurs fois les mêmes travaux. Je pense que, en tout cas, l’expérience nous montre que nous n’archivons pas assez nos données. » ;

« Alors dans l’idéal, moi je pense que c’est au laboratoire. Parce que c’est là, enfin moi j’aime avoir des interlocuteurs auxquels je peux parler, quoi, c’est ça mon critère. » ;

« Ben moi je pense que ce qui serait le mieux c’est qu’il y ait un référent au niveau du laboratoire. Parce que voilà chaque laboratoire, chaque discipline a ses particularités aussi et il va y avoir des petites différences j’imagine entre certains labos. Donc, du coup qu’il y ait vraiment quelqu’un au laboratoire, c’est beaucoup plus en proximité, ça pourrait aussi aider plutôt qu’à l’échelle d’une composante, ou de toute l’université. » ;

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

« La seule personne qui stocke ce que je produis à long terme, bah c’est moi. Donc le jour où, je ne sais pas, je quitte la recherche ou quelque chose comme ça, a priori, pas sûr qu’on revoit ces données. Donc ça c’est peut-être un problème, moi je sais que mes données d’il y a plusieurs années, elles me servent de temps en temps. Donc il y a un intérêt à le faire sur du très long terme, pas sur tout évidemment, mais sur certaines choses. Je pense qu’on peut faire le tri entre ce qui peut être réutilisé ou pas à mon avis. » ;

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

2.5 Le partage et la diffusion des données de la recherche

On questionne ici les opinions des personnes interviewées sur la politique nationale d'ouverture des données de la recherche, de l'impact que cela peut avoir sur leurs pratiques, de leur utilisation des entrepôts de données. Personne ne semble être foncièrement contre, mais beaucoup d'interrogations subsistent, notamment sur l'utilité d'ouvrir des données brutes. La lisibilité de ces données est questionnée, nombreux sont ceux qui n'y voient pas d'intérêt, qui ne souhaitent pas y consacrer plus de temps ou, même, qui se méfient de cet aspect. Les valeurs d'accessibilité et d'ouverture sont comprises, mais ne semblent pas suffire face aux problématiques compétitives dans certaines disciplines. L'ouverture et la « rentabilisation » des données de la recherche sont pour certains deux injonctions qui se contredisent et qui freinent les initiatives sur la question. Pour celles et ceux qui par contre adhèrent aux discours d'ouverture, les problématiques des logiciels et codes sources, du développement du libre et de la science participative sont également des voies à explorer. Nous notons enfin, généralement, une connaissance perfectible des personnes interviewées par rapport aux entrepôts de données qui existent.

« C'est à dire que si j'apprenais que la mise en ligne de certaines données de la recherche pourrait former une sorte de section à part dans mon CV ou être valorisée au moment d'un recrutement. Puisqu'en fait, je suis encore dans, comment dire... je suis encore quelqu'un qui sera un jour peut être évaluée pour obtenir un poste de professeur. » ;

« Je pense que les sciences humaines et la musicologie en particulier ont pâti de leurs difficultés de diffusion et de leur difficulté d'ouverture et de partage. Je pense que c'est l'une des problématiques à prendre à bras le corps pour la musicologie française et internationale. » ;

« Le partage des données de la recherche oui, je pense que c'est important, bien sûr, mais pas à n'importe quel prix et vraiment avec des conditions. Il faut absolument que ce mouvement en avant soit suivi par les sections CNU, par toutes les institutions, qui on va dire mettent des règles quant au type de recherche qu'on doit faire, au type de publication qu'on doit produire. » ;

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

« Ouais, je suis obligé de me fermer. Pour conserver un petit peu de primeur dans ce que je fais. » ;

« Clairement, j'ai l'impression, peut-être que je me trompe, peut-être que dans d'autres équipes c'est plus concret, mais nous dans notre équipe, c'est encore au niveau de la discussion et c'est peut-être pas encore au niveau concret de "qu'est-ce qu'on fait pour partager nos données ?" » ;

« Parce que la mise en application de cette stratégie, elle n'est pas si simple en fait, elle est très contraignante pour nous. De nous dire par exemple, il faut que vous mettiez en ligne toutes vos données, vos *pre-print* par exemple. C'est du temps passé en plus à le faire. C'est sous une forme qui n'est pas forcément très sexy à lire. Et donc le partager, ça veut dire qu'on va partager quelque chose qui n'est pas sexy, qui n'est pas très intuitif. On a envie d'être fier de ce qu'on partage, de partager de belles choses dans les articles. Les choses sont mises en forme correctement, et cetera. Et ça, c'est quelque chose auquel on est très attaché » ;

« C'est une contrainte, un temps supplémentaire, et le temps qu'on consacre à ça, on ne le consacre pas à écrire nos articles, à rechercher d'autres fonds, à encadrer nos étudiants, à faire des congrès, à faire plein d'autres choses quoi » ;

« Par contre, pour que ça soit accepté, je le vois pour HAL par exemple, moi je le fais pour les enseignants-chercheurs. Ils sont très contents, c'est-à-dire que ceux qui le faisaient avant ne le font plus. Voilà, parce qu'ils ont d'autres choses à faire. C'est ça qu'il faut voir. Donc comme ils ont un milliard de choses à faire, ce genre de chose, s'ils peuvent s'en passer, bah ils vont s'en passer, voilà. Eux, ce qu'ils veulent, c'est être actif et pouvoir toujours fonctionner et jamais être arrêté par un certain nombre de choses. Voilà donc ils sont comme ça. Donc en fait, il faut les accompagner. Je dirais que mon expérience [...], c'est qu'il vaut mieux en fait accompagner les gens dans ce qu'ils sont en train de faire plutôt que de les forcer à faire quelque chose parce que ça, ça marche pas bien. » ;

« Alors évidemment, actuellement, on est entre deux eaux, c'est-à-dire qu'on nous demande de plus en plus de valoriser d'un point de vue financier ce que l'on a produit, ce qui signifie fermer. Fermer ce que l'on a fait pour pouvoir soit faire des brevets ou faire des choses comme ça. Moi je n'aime pas ce genre de chose, ce n'est pas ma manière de faire, mais il faut bien voir qu'effectivement, de plus en plus, les gens, on va leur demander effectivement de gagner de l'argent sur leur recherche ou tout ce qui est associé et donc ils devront effectivement vendre leurs données de la recherche en quelque sorte. » ;

« Comment les données qu'on a recueillies pour des expériences tellement spécifiques, même si on les partageait, est-ce qu'elles pourraient vraiment être comprises par d'autres ? [...] Comment trouver un langage commun, finalement, qui soit compréhensible à un plus grand nombre ? » ;

« [Ouvrir les données de la recherche] ça signifie que personne ne regardera. Ce n'est pas provocateur ce que je dis. C'est ce qui se passe dans la réalité. » ;

« En fait, je ne sais pas si je suis d'accord ou pas d'accord. Sauf que c'est ce que je vous disais au début, je suis sensible à la question, je comprends l'idée, je trouve ça important de... ce n'est pas intelligent de refaire dix fois les mêmes choses, c'est important de récupérer les choses qui ont déjà été faites. Par ailleurs, on est financé par le public, c'est normal que nos données soient publiques, et cetera. Donc ça sur le fond, j'ai une sensibilité à cette question et je comprends le souci. Par contre, j'ai l'impression que c'est dépendant dans une mesure ultra fine de quelles données on parle, de quelle science on parle, etc. Et alors en particulier, nous, notre terreur dans notre domaine, c'est qu'on fasse un truc qui soit super adapté aux

grands instruments, qui ont, pour caricaturer, toujours les mêmes colonnes, toujours les mêmes courbes quasiment, ils peuvent se faire un truc automatique. » ;

« Alors j'ai une bonne confiance que ça va pas beaucoup être consulté et téléchargé, je vous le dis tout de suite. Je pense que j'ai les noms de tous les gens que mes données intéressent si vous voulez. » ;

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

« Après c'est des choses qui se développent, je pense, qui sont vraiment très intéressantes. Dans science ouverte, il y a aussi tout ce qui est science participative. Et du coup, le fait que les citoyens peuvent avoir... Enfin c'est un renversement de positions, ce n'est pas le chercheur qui a des idées, qui va développer ses idées et faire une expérience et éventuellement utiliser des citoyens pour faire son expérience, c'est... un citoyen peut générer un programme de recherche, orienter un nouveau programme et du coup l'intérêt de mettre les données les plus ouvertes possibles, il est là aussi. » ;

« Moi je pense qu'il faut quand même ouvrir les données. Tant pis, c'est un risque effectivement. Mais si on ouvre les données, on les ouvre aussi pour des... ils ne vont pas toujours gagner [les GAFAM]. Voilà le principe du web, c'est quand même, depuis son invention, c'était justement d'être ouvert et ça permet à plein de communautés de travailler. » ;

« C'en est vraiment très contraignant quand on a connu le passé avec des solutions, des systèmes où on n'était pas obligé de se loguer avec 36 000 mots de passe pour accéder à sa donnée d'intérêt ou bien par des conflits de, comment dire, des incompatibilités de système. » ;

« Tout ça, ça va continuer à évoluer, mais il ne faudrait pas que la sécurisation à outrance complexifie et entrave l'efficacité de la recherche ».

2.6 Attentes et besoins concernant la gestion des données de la recherche

On questionne ici les attentes des personnes interviewées par rapport à la problématique des données de la recherche, leur perception des difficultés individuelles et collectives sur cette question et les solutions pour y remédier. Il faut également modérer les résultats de cette catégorie : le placement en fin d'entretien ne facilite pas nécessairement les choses. On relève néanmoins des attentes globales sur la formation, la mise à disposition de supports, la mise en place d'une plateforme mutualisée. Il semble y avoir un besoin d'avoir un interlocuteur, au plus près de la discipline, de l'unité de recherche, pour prendre en charge la totalité des dossiers sur cette question. Les personnes interviewées identifient également un besoin de formation, pour elles, les doctorants, les étudiants de Master, même s'il n'y a pas vraiment de consensus sur la forme que devrait prendre un éventuel dispositif de formation. Le constat est le même pour la question d'une plateforme : on ressent un besoin sur le stockage, l'archivage, les métadonnées, et toutes les problématiques que l'on pourrait

y agréger, mais aucune forme réelle ne se dessine. Dans cette catégorie, là encore, les problématiques du libre et de l'assistance aux outils sont à noter.

« Il y a tellement d'autres choses importantes à améliorer au sein des universités [...] d'abord, il faudrait que les gens comprennent ce qu'on peut faire pour eux. Quel est leur intérêt ? Et après comprendre ces gens. » ;

« Évidemment, avoir un interlocuteur au moins. Si c'est un interlocuteur à la fac, moi, je sais que je n'irais pas. Si c'est un interlocuteur SHS, j'irais peut-être. » ;

« Je pense qu'on rate des choses quand même, au sens où je pense que les chercheurs, soit n'ont pas la formation, soit n'ont pas l'envie, soit n'ont pas le temps de le faire et je pense qu'il faudrait des référents en fait. Par laboratoire, qui puissent prendre en charge ces questions-là, qui sont en lien ensuite soit avec la BU, soit avec un référent HAL. » ;

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

« Mais voilà, il n'y a pas de formation à la gestion de données dans la formation que j'ai suivie donc rien que, peut-être, un cours, ça pourrait ne serait-ce que donner les grands noms qui permettent d'aller voir par soi-même. » ;

« Peut être que sur l'*open science*, il faut faire quelque chose comme ça, en format de formation ou un mini mooc, je ne sais pas. » ;

« Je pense que les enseignants-chercheurs ont besoin d'être formés. Et sur plein de choses par rapport au fait de diffuser à certains moments ou ne pas diffuser à d'autres pour éviter de se faire piquer leurs idées. » ;

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

« Alors, pour bien répondre, il faudrait déjà que je sois formé quoi, tout simplement. Je pense que c'est déjà ce que vous faites, de former le personnel. Voilà déjà expliquer, expliquer les enjeux, expliquer ce que c'est. Puis surtout, il faudrait proposer quelque chose de simple, qui ne demande pas trop de temps aux gens et qui soit facilement transférable quoi. Si c'est trop fastidieux, les gens ne le feront pas. » ;

« Les meilleures dispositions, ce serait sans doute justement d'avoir une base de données qui soit largement accessible, en tout cas qui parle à tout le monde. Le problème, c'est qu'aujourd'hui c'est difficile d'avoir par réflexe le nom d'une base de données scientifique qui soit accessible, donc il y a un problème déjà au départ de communication là-dessus. » ;

« Je trouverais ça bien qu'il y ait soit au niveau des laboratoires, mais ça peut être au niveau de l'université, des outils pour gérer le stockage des données de recherche en cours. Effectivement, après de faciliter la transition vers un archivage plus pérenne, à la fois pour

la mise à disposition et puis pour la pérennité des données. Et c'est une question que je me pose vraiment en ce moment, est ce que ça vaudrait le coup par exemple de développer un *cloud* ou une solution de stockage spéciale là pour toutes les données qui serait poussé par l'Université Paris Saclay ? Et en fait, j'ai quand même l'impression que la bonne réponse, ça serait plutôt de contribuer à des outils comme OSF. Ou comme des fondations ou des dispositifs qui existent déjà. Et du coup si on peut plutôt s'associer et mettre en commun des moyens pour avoir des plateformes qui soient interopérées et interopérables, ça serait vraiment bien. Plutôt que de redévelopper des serveurs internes, des trucs, et encore un nouvel entrepôt de données à indexer. Je me dis que c'est peut-être plus intelligent de mettre en commun les données. » ;

« Alors, il y a deux questions différentes. Est-ce qu'on pourrait avoir besoin d'espace de stockage ? Et la question des projets qui est différente, mais nous on a aussi cette question-là. Sur la question de l'espace de stockage, je pense qu'il y a un gros besoin. [...] On aurait besoin d'avoir des lieux de stockage effectivement pour les thèses, pour les postdocs, etc. Mais là, il y a une vraie réflexion à avoir sur "comment est-ce qu'on fait quand les jeunes s'en vont ?" » ;

« On est toujours en tant que chercheur, producteur de données et utilisateur de données, à la recherche d'espace de stockage, que ce soit bien conçu, qu'on puisse trouver les données et aussi je dirais un aspect assez crucial de mon point de vue une forge intelligente performante de type GitLab pour gérer les logiciels. Et moi je suis un friand utilisateur de Renater et en particulier de SourceSup. » ;

« Ce qui serait intéressant, c'est de dire, voilà, je dépose des données, ces données sont des données de microscopie, je connais tel ou tel logiciel qui pourrait les traiter. Je dépose ces données, je suis massiste, je dépose des données de masse, voici les logiciels qui seraient susceptibles de les traiter, ou en tout cas, dont je connais la compatibilité. Vous vous rendez compte du temps qu'on gagnerait ? Vous allez chercher des données et on vous donne le mode d'emploi pour vous en servir. » ;

« Et il nous faut des solutions simples de stockage de données, d'étiquetage de données. Au fond, des choses que moi je fais parce que je sais explorer un système. Et donc je me retrouve en face de GitLab, bon bah voilà, je m'ouvre un compte et puis je fais, j'explore le truc et je découvre les fonctionnalités comme un grand parce que tout ça est très logique pour moi. Mais les gens avec lesquels je collabore, bah ils n'y sont pas du tout quoi. » ;

« Je pense que c'est très important que les doctorants soient formés dès les premiers mois de leur thèse sur tous ces sujets-là. Parce que même pour eux, pour mener son projet de recherche, bah voilà, il y a un certain nombre de choses auxquelles au début on ne pense pas et puis en fait à un moment on se rend compte que c'est très important, qu'on a besoin, et qu'il aurait fallu être plus rigoureux sur ces choses. » ;

« Former au début de la thèse ou avant. Effectivement, je pense que les étudiants en Master devraient être sensibilisés à ces questions aussi. Qu'ils continuent en thèse ou qu'ils fassent d'autres choses. Juste à partir du moment où ils ont un travail en lien avec la science, la recherche d'une manière ou d'une autre. Bah effectivement, c'est mieux s'ils sont sensibilisés, même s'ils travaillent en R et D en entreprise ou qu'ils deviennent cadres dans

une entreprise où ils utilisent en fait des données de la recherche. Une formation sur ces sujets-là, je pense que oui c'est utile. » ;

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

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

« Je ne sais pas comment dire, j'ai l'impression que ça reste un peu des préoccupations individuelles et qu'il n'y a rien qui se formalise au niveau de l'équipe, au niveau du laboratoire pour impliquer les autres, en particulier les doctorants, sur ces thématiques-là. » ;

« En gros, on voit bien que l'Université Paris Saclay ne fait pas un choix tout à fait libre sur ses systèmes informatiques. Et donc là il y aurait peut-être quelque chose qui serait que, par exemple... Donc moi j'utilise pas mal Cirrus. Cirrus, c'est développé par un collègue informatique. Enfin voilà ça a été mis en place parce qu'il y a des chercheurs bénévoles qui ont mis ça en place. La plateforme de programmation Jupyter, c'est pareil, c'est des collègues du département de Math. Et les services informatiques de l'université, bon ils ont développé e-campus. Ils ont mis *collaborate* dedans qui est un logiciel propriétaire. Enfin voilà, c'est peut-être avoir un appui aussi sur tous les outils. » ;

« Enfin je ne sais pas, vous pouvez probablement ne rien y faire, mais souvent, quand même, on a l'impression qu'on se bat... Enfin voilà, on parle beaucoup de données ouvertes, etc. Les logiciels ouverts et les systèmes d'information ouverts, c'est important aussi et là on se bat quand même contre les institutions. » ;

« Un serveur, ça coûte 15 000 €, ça s'achète dans un projet ANR. Il y a des *data center* un petit peu partout en France. Et ce qu'il faut, c'est l'administration du *data center*. Et le problème quand l'administration du *data center* elle est centralisée, c'est compliqué pour les gens d'y accéder. Et bon, ça c'est encore un problème politique hein. Ce n'est pas la philosophie des tutelles actuellement d'augmenter le nombre d'emplois dans la fonction publique. Ils aiment bien acheter des serveurs, ça c'est sûr. Trouver de l'argent, il y en a. On n'a aucun problème pour trouver de l'argent pour acheter des grosses machines et des trucs comme ça. Mais notre besoin, il n'est pas là, il est dans les gens. » ;