Comment j'ai déposé les données de ma recherche (sans savoir ce qui m'attendait)

Célia Bouchet

ARTICLE - VERSION AUTEUR

Comment j'ai déposé les données de ma recherche (sans savoir ce qui m'attendait)

Célia Bouchet

CRIS, Centre de recherche sur les inégalités sociales (Sciences Po, CNRS)
LIEPP, Laboratoire interdisciplinaire d'évaluation des politiques publiques (Sciences Po)
CEET, Centre d'études de l'emploi et du travail

In: *Genèses*, 2023/4, n°133, p. 113-128.
DOI : 10.3917/gen.133.0113

Loi du 7 octobre 2016 pour une République numérique, Plan national pour la science ouverte de 2018, Directive européenne du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public… Au fil de la dernière décennie, une forte production juridique a encouragé voire, dans certains cas[1], rendu obligatoire la mise à disposition d'une partie des données de la recherche (Stérin et Noûs 2019). L'ouverture des données de la recherche n'est pas une perspective nouvelle dans les sciences sociales : au XX^e^ siècle, le mouvement de la science ouverte l'a érigée comme un objectif central, et au tournant du XXI^e^ siècle, plusieurs initiatives d'archivage ont contribué à diffuser une « culture de la revisite » (Scot 2006). Mais la bureaucratisation de ces démarches soulève plusieurs problèmes de fond, que la recherche récente a visibilisés. Alors qu'aux États-Unis, les politistes ont notamment exprimé leur exaspération à se voir opposer un paradigme de réplicabilité pensé dans les termes des sciences dures (Duchesne et Noûs 2019), les ethnographes en France ont insisté sur les défis déontologiques posés par l'ouverture des données (Mouton 2018 ; Revelin *et al.* 2021). Comme un numéro récent de *Genèses* l'a souligné, au sein d'un monde académique en prise avec une judiciarisation croissante, la conservation de données qualitatives pose des risques de rupture de l'anonymat – susceptibles d'affecter les personnes enquêtées mais aussi les chercheur·es (Laurens 2022 ; Siméant-Germanos 2022). Le droit porte la trace de cette tension, entre injonction à l'ouverture des données de la recherche et volonté de protection des données personnelles (Stérin et Noûs 2019).

Au-delà de ces légitimes inquiétudes de principe, les défis pratiques de l'ouverture des données restent peu renseignés. De fait, si la mise à disposition de données est une pratique émergente parmi les chercheur·es (Gay 2021), l'essentiel de cet effort est assuré – et documenté – par des équipes de soutien à la recherche (Duchesne et Garcia 2014 ; Caporali, Morisset et Legleye 2015 ; Cornilleau et Duwez 2022). Leur travail offre des perspectives précieuses sur les rouages des dépôts de données, que nous mobiliserons tout au long de cet article. Mais, par leur distance aux enquêtes effectuées et/ou leur registre technique, ces retours d'expérience peuvent banaliser et parfois invisibiliser des implications méthodologiques, juridiques et scientifiques du dispositif.

Jeune docteure en sociologie, j'ai travaillé pendant six mois (de janvier à juin 2022) à déposer des données quantitatives et qualitatives issues de ma recherche de thèse sur un entrepôt institutionnel (Encadré). Cet article propose un retour réflexif sur mon expérience, rendant compte de la confusion que j'ai éprouvée à cette occasion, et des ressorts ambivalents de mon engagement dans ce travail. Que révèle le trouble d'une jeune chercheuse qui ouvre certaines de ses données quant aux normes pratiques de cette activité ? Je distingue trois dimensions qui ont contribué à rendre mon expérience déroutante : les temporalités et exigences contrastées du versement des différents types de données issues de ma thèse (variabilité des normes) ; les zones d'ombre dans les règles juridiques et les protocoles scientifiques de l'entreposage (défaut de normes) ; mon appropriation ambivalente de la démarche d'auto-dépôt, entre perméabilité à des incitations contextuelles, détermination à mener à bien ce travail au nom de diverses valeurs et doutes sur le ratio coûts/bénéfices de mon investissement (marges de rationalisation des normes).

Méthodologie de l'enquête

L'enquête sur laquelle porte cet article a été menée de septembre 2018 à novembre 2021, dans le cadre d'une thèse de doctorat en sociologie. Son but était de documenter l'empreinte du handicap sur les parcours de vie dans le contexte français, depuis l'enfance jusqu'aux positions sociales à l'âge adulte. Pour ce faire, l'enquête s'est intéressée à différentes franges de personnes vivant depuis la naissance, l'enfance ou l'adolescence avec un « handicap » au sens de la loi du 11 février 2005[2] (limitations durables dans les activités quotidiennes liées à un problème de santé, un trouble ou une déficience) ; et elle les a comparées à des personnes sans limitations, dites « valides ».

Il s'agit d'une enquête par méthodes mixtes. Son volet quantitatif se fonde sur une exploitation de la vague 2011 de l'Enquête emploi en continu (EEC) et de son module *ad hoc* « Insertion professionnelle des personnes handicapées » (Insee et Dares). Entre 2018 et 2021, j'ai réalisé des statistiques bivariées, des régressions et des analyses factorielles pour caractériser les situations scolaires, professionnelles et familiales des 3 185 personnes sondées ayant grandi avec un « handicap au sens large », c'est-à-dire soit une reconnaissance administrative de handicap, soit un problème de santé durable et une limitation durable dans les activités quotidiennes et/ou vis-à-vis de l'emploi ; et pour comparer ces situations à celles de 16 168 personnes sans handicap au sens large, dites « valides ». Toutes les analyses statistiques ont été réalisées avec le langage de programmation R.

Le volet qualitatif repose sur une campagne d'entretiens semi-directifs menés avec 37 personnes de 30 à 55 ans ayant grandi « avec un handicap ou des difficultés handicapantes » (termes de l'annonce), dont 20 ayant grandi une déficience visuelle (partielle ou totale) et 17 ayant grandi avec des troubles dys (dyslexie, dyscalculie, dyspraxie, dysphasie, etc.). Pour les atteindre, des annonces sous plusieurs formats (braille, gros caractères, numérique) ont été diffusées par divers vecteurs : associations, forums et groupes sociaux, institutions spécialisées, boutiques de matériel adapté, etc. Les entretiens se sont tenus entre décembre 2019 et juin 2020, d'abord dans 6 villes de France métropolitaine et leur banlieue puis (à partir du confinement de mars 2020) par téléphone ou visioconférence. La grille d'entretien commençait par une question large sur la « situation en ce moment », puis suivait un format biographique. J'ai enregistré les entretiens et les ai intégralement retranscrits puis codés à l'aide du package RQDA de R. À l'issue de chaque entretien, j'ai également consigné 2 à 5 pages d'informations non enregistrées dans des fichiers de notes méthodologiques.

Pour l'analyse du travail d'auto-dépôt, je m'appuie sur quatre types de supports : une prise de notes réflexives débutée en mars 2022 ; des archives de correspondance électronique ; ma mémoire immédiate, ayant commencé cet article fin juin 2022 soit peu après la mise en ligne du second jeu de données ; et, à l'issue de la rédaction d'une première version de cet article, la consultation des protagonistes que je cite le plus souvent pour trianguler leurs souvenirs de nos interactions avec les miens. Les personnes enquêtées citées dans cet article sont désignées par des pseudonymes.

Entre formalisme et improvisation

Reproduisant une opposition quali/quanti typique (Moore 2007 ; Ruggiano et Perry 2019), le volet qualitatif de la recherche mobilise des données primaires, récoltées lors d'une campagne d'entretiens, tandis que le volet quantitatif exploite les données d'une enquête produite par la statistique publique. Ce paramètre a façonné la temporalité et le degré de formalisation de chaque processus d'auto-dépôt avec, pour le volet qualitatif, de

premières démarches avant même le début de la campagne d'entretiens et, pour le volet quantitatif, une décision de versement à la toute fin de la recherche.

Volet qualitatif : un partage prévu en amont et des étapes balisées

L'idée d'une mise à disposition de certaines données de ma recherche doctorale a émergé en novembre 2019. En thèse depuis un peu plus d'un an, j'ai déjà avancé sur le volet quantitatif de l'enquête et je m'apprête à débuter son volet qualitatif : une campagne d'entretiens ciblant des trentenaires, quadragénaires ou quinquagénaires ayant grandi avec une déficience visuelle ou avec des troubles dys.

Avant le lancement de l'enquête, sur la suggestion de ma directrice de thèse et d'une doctorante plus âgée, je prépare une fiche d'information à remettre aux personnes rencontrées. Ce document sert à la fois de récapitulatif des principaux renseignements sur la recherche (objectifs, conditions de participation, adresse de contact) et de support pour le recueil des préférences et autorisations (souhait de retours sur les résultats, accord pour être recontacté·es pour une recherche ultérieure). Je compte communiquer cette fiche à la déléguée à la protection des données (DPD) de mon université, identifiant bien son rôle après qu'elle est intervenue dans mon centre de recherche. Pour préparer cette prise de contact, je soumets le document à ma directrice de thèse. Elle en valide l'essentiel mais y ajoute une troisième question : « Autorisez-vous la chercheuse à transmettre, une fois la recherche terminée, les données pseudonymisées à un fonds institutionnel sécurisé d'archive des données de la recherche, en vue d'une mise à disposition sous conditions à d'autres chercheurs ? » Elle m'explique que le recueil de ce consentement m'ouvrira l'option d'un dépôt des entretiens sur une plateforme dédiée – sans caractère contraignant, si la démarche s'avère finalement trop complexe. Son argument me convainc. C'est cette fiche enrichie que je communique à la DPD quelques semaines plus tard, et que je remets aux personnes rencontrées au fil du travail de terrain (généralement lors des rencontres, en amont si elles posent beaucoup de questions ou en cas d'entretien vidéo).

Lors de la campagne d'entretiens, la perspective de réutilisations scientifiques des entretiens pseudonymisés recueille davantage d'adhésion que je ne m'y attendais. Parfois dès la remise des fiches, le plus souvent lorsque j'explique le contenu de la rubrique « autorisations » à l'issue des entretiens, 33 des 37 personnes rencontrées me donnent leur accord[3]. Le plus contraignant est de m'assurer de conserver une trace de cet accord sans sembler excessivement formelle, typiquement en profitant de l'enregistrement (donc en faisant attention à ne pas l'arrêter trop tôt) ou à l'occasion d'échanges par emails. Je commets quelques erreurs. Une fois, réalisant avoir coupé l'enregistrement de façon prématurée, je le relance. À deux autres reprises, je n'ose pas rallumer par peur de sacraliser la demande ; je me contente de consigner la réponse sur mon carnet de terrain (en signalant cette prise de note verbalement : « je le note ») et d'en créer une copie informatique dès mon retour.

Après la fin de la campagne d'entretiens (juin 2020), la mise à disposition des données sort de mes préoccupations pendant près de 18 mois. Même lorsque je transcris les entretiens, j'ai en tête mes propres perspectives d'analyse, et non d'éventuelles réutilisations. C'est seulement une fois ma thèse soutenue que, en janvier 2022, je repense au devenir des transcriptions. Sur cette période, une collègue doctorante qui travaille sur un sujet proche du mien commence à fatiguer d'enchaîner les entretiens et j'imagine les économies d'échelle que permettrait la réutilisation d'une partie de mon corpus – pour elle ou pour d'autres. Je contacte les personnes ressources de mon université en matière de gestion des données (listées sur le site de la bibliothèque) pour m'enquérir des options d'entreposage envisageables. Je leur demande notamment si l'entrepôt de données de l'université permet

de stocker des fichiers protégés qui seraient transmis seulement sous conditions – conformément à l'engagement pris auprès des personnes enquêtées – et en obtiens confirmation.

Je réfléchis alors aux transformations à apporter aux transcriptions avant de les verser. Outre la rectification de quelques coquilles introduites à la transcription (comme des mots manquants ou des fautes d'orthographe), se pose la question des modalités d'anonymisation. La fiche remise aux personnes enquêtées n'évoque qu'un changement de leur nom, une pseudonymisation. Mes recours passés à des enquêtes du catalogue de la banque d'enquêtes qualitatives beQuali, dans le cadre d'enseignements, m'ont appris que cette option était classique. Mais j'ai également garanti aux personnes rencontrées que, dans mes publications scientifiques, je changerais aussi les autres éléments susceptibles de les rendre identifiables par des inconnu·es – comme les noms de villes, d'organisations professionnelles, etc. Je n'ai pas ré-explicité que le degré d'anonymisation serait moindre pour l'entreposage, et présume que les personnes ont pu assimiler les deux. Dans cette perspective, il me semble plus prudent de changer tous les noms propres dans les transcriptions, ou de donner la consigne expresse aux personnes réutilisant les données de procéder à ces ajustements avant publication. Je consulte ma directrice de thèse, qui m'encourage à « supprimer tous les noms propres par sécurité ». Elle estime que les réutilisations représentent des « *process* encore très mal définis », entendant par là qu'une consigne d'anonymisation serait compliquée à faire respecter. Je me résous donc à effectuer le travail moi-même, ce qui implique de modifier plusieurs centaines de mots-clés et des milliers d'occurrences sur un ensemble de presque 1 400 pages[4].

Le temps de réaliser cette pseudonymisation renforcée, c'est finalement début mars que je sollicite le *data librarian* (bibliothécaire responsable des données de la recherche) pour créer un compte sur l'entrepôt de données de la recherche de mon université. Au fil du mois, en mode hors-ligne, je charge les transcriptions en protégeant les fichiers (accès conditionné à une demande justifiée) et je les complète par des documents méthodologiques : annonce diffusée pour le recrutement, grille d'entretien, fiche d'information, etc. Je publie finalement le jeu de données début avril 2022, presque deux ans et demi après les premières formalités.

Volet quantitatif : une ouverture envisagée sur le tard

Je n'envisage l'éventualité d'un partage des statistiques issues de ma recherche doctorale qu'en juillet 2021, soit presque trois ans après le début du volet quantitatif et 18 mois après les premières démarches visant à l'ouverture de mes données qualitatives (voir la chronologie en annexe). Je suis alors en pleine rédaction de thèse, et je dois procéder à des coupes substantielles dans mon manuscrit pour en alléger la lecture. Pour guider cette démarche, ma directrice de thèse me conseille d'enlever certains tableaux de statistiques du corps du manuscrit et de les diffuser sur d'autres supports – par exemple sous la forme d'annexes, d'annexes électroniques, ou en les versant eux aussi à l'entrepôt de données de la recherche de mon université. Cette dernière option a l'avantage de fournir un identifiant pérenne (*digital object identifier* ou DOI). Pour autant, le flou juridique qui l'entoure m'invite à l'écarter sur le moment. Au moment du dépôt de ma thèse à l'automne, je mobilise des annexes électroniques stockées sur un site personnel, reléguant l'auto-dépôt de statistiques additionnelles au rang de vague projet à plus long terme.

En janvier 2022, suite à ma soutenance de thèse, je m'attelle d'abord au dépôt du volet qualitatif, pour lequel j'identifie bien la marche à suivre. Pour les exploitations statistiques, tout reste à définir. Y a-t-il du sens à créer un jeu de données, alors que les résultats essentiels de ma thèse figurent déjà dans le manuscrit ou les annexes électroniques ? Le cas échéant, que verser ? Reproduire les thématiques et indicateurs dont je traite dans ma thèse, ou mettre en forme les statistiques que je n'ai finalement pas exportées ? Mobiliser des statistiques

bivariées seulement, ou aussi les techniques d'analyse plus poussées que je mets en œuvre dans ma thèse – régressions, analyses factorielles et analyses lexicométriques ? Réaliser de simples copies des sorties de commandes, ou effectuer un travail de mise en forme (tableaux, graphiques) ? Au cours de l'hiver et du printemps 2022, des échanges avec ma collègue doctorante évoquée précédemment m'aident à clore certaines options et viennent en ouvrir d'autres. Par ses questions et ses remarques, je comprends qu'elle aurait l'utilité de données de cadrage variées, davantage que de statistiques techniques au coût d'entrée plus important. Par ailleurs, en partageant mes lignes de code R avec cette doctorante, je me rends compte que la mise à disposition de mes scripts pourrait aussi avoir un intérêt, en facilitant la prise en main de l'EEC et de ses variables liées au handicap. Je décide donc de mettre en ligne deux types de contenus : des compilations de statistiques bivariées, tableaux et graphiques[5], qui abordent le handicap à travers des thématiques diverses ; et les scripts R ayant permis de les produire. Ce programme – on le verra plus loin – m'occupe plusieurs heures par jour durant trois mois.

Courant juin 2022, moins d'un an après mes premières réflexions sur un éventuel auto-dépôt de statistiques et à peine quelques mois après le début du travail concret dans ce sens, j'aboutis à un ensemble de 153 documents : trois compilations thématiques, les trois scripts R ayant permis d'en produire les 81 tableaux et 64 graphiques, les sorties directes de ces supports depuis R, et deux documents méthodologiques[6]. Le corpus ainsi constitué couvre les modes de délimitation de populations et sous-populations ayant grandi avec un handicap ; leurs caractéristiques socio-démographiques en comparaison avec la population valide (sexe, âge, profession des parents dans l'enfance, statut d'occupation du logement, etc.) ; et les situations scolaires, professionnelles et familiales des différents groupes.

Incertitudes juridiques et scientifiques

L'auto-dépôt des contenus d'entretiens comme celui des exploitations statistiques a soulevé de nombreux questionnements pour moi, comme je découvrais une activité nouvelle sans en maîtriser les normes. L'équipe de soutien à la recherche m'a été d'une grande aide pour m'initier aux dimensions matérielles de l'entreposage : création d'un jeu de données sur le répertoire institutionnel, chargement de fichiers, complétion des métadonnées, etc. Deux types de questions se sont avérées plus épineuses, dénotant les défauts de régulation sur ces aspects (absence de normes ou publicisation insuffisante de ces normes) : le périmètre précis des contraintes juridiques, et la gestion scientifique des enjeux de catégorisation induits par la transformation des données.

Contraintes et zones troubles dans les prescriptions juridiques

Dans la loi pour une République numérique, l'injonction d'ouverture des données de la recherche s'arrête là où commencent les droits de tierces personnes (Stérin et Noûs 2019). Parmi ces droits, figurent notamment le droit des personnes enquêtées à la protection de leurs données personnelles (assurée par le règlement général sur la protection des données, ou RGPD), et le droit de la statistique publique à la propriété de ses enquêtes (encadré par des contrats ou des conventions lors de mises à disposition des données).

Les entretiens réalisés pour ma recherche doctorale relèvent du premier cadre, et mes usages de l'EEC et de son module, du second. Cet environnement juridique pose certaines contraintes claires, par exemple la nécessité d'obtenir le consentement des personnes interrogées à l'entreposage des contenus d'entretien, ou encore l'obligation de citer l'Insee à chaque mobilisation de l'EEC[7]. Mais des zones troubles demeurent, comme l'illustrent deux épisodes survenus lors des auto-dépôts.

Une première incertitude a porté sur le droit d'entreposer des données secondaires. Lors de mes réflexions initiales, au moment de la rédaction de ma thèse, je doute que cette démarche soit légale. Selon les termes de la

convention que j'ai signée avec Progedo-Adisp pour l'accès aux données de l'EEC et de son module *ad hoc* 2011, je suis autorisée à « utiliser les données [exclusivement] dans une finalité de recherche », mais non à « céder ces données, sous quelque forme que ce soit, à une tierce personne ». Autrement dit, la *production* scientifique est autorisée, alors que la *reproduction* est interdite. Mais où s'arrête la première, et où débute la suivante ? Je contacte les ingénieur·es et bibliothécaires de mon université en charge de la gestion des données pour clarifier ce point ; il apparaît que l'équipe n'a encore jamais été confrontée à ce cas de figure. Une ingénieure me transmet un contact à l'Adisp.

La correspondance électronique et téléphonique qui s'ensuit avec l'équipe de l'Adisp joue un rôle essentiel de production normative. Lors de nos échanges, dans un premier temps, mon interlocutrice peine à comprendre ce que je veux faire, tandis que je peine à comprendre où elle place le problème. Elle insiste sur le fait que la ligne rouge est l'interdiction de créer une « sous-base » de l'EEC et de son module. Or, ce risque me semble complètement écarté, puisque je n'entends copier aucune des données primaires. C'est seulement suite à des questions plus ciblées que je réalise que, dans sa définition, la notion de « sous-base » n'englobe pas seulement la copie des données primaires mais aussi tout recours à des tableurs (Excel, LibreOffice Calc ou assimilé). Interloquée, j'indique alors à mon interlocutrice que le partage de tableaux et graphiques sous un format ou un autre – tableur ou document Word – ne fait aucune différence à mes yeux. Elle rebondit en soulignant la différence essentielle que cela constitue pour l'Adisp : les tableurs constituent des « sous-bases » là où les documents textuels valorisent des « matériaux scientifiques » originaux. Je m'engage volontiers à favoriser le second type de support. C'est donc au final la proximité entre le format de stockage des données primaires et celui sous lequel je restitue mes données secondaires qui a servi de critère pour qualifier un usage juridiquement mal défini.

Le deuxième flou juridique auquel je me suis heurtée concerne la possibilité de verser mes notes méthodologiques sur les entretiens : interactions avant ou après l'enregistrement, langage non verbal, description du logement de la personne en cas d'entretien à son domicile, etc. Les manuels de méthodologie qualitative s'accordent sur l'importance de ces notes de terrain détaillées, même en cas d'enregistrement (Bertaux 1997 ; Beaud et Weber 2003 ; Merrill et West 2009). La plupart relèvent l'enjeu de consigner les renseignements que les personnes transmettent après la fin de l'enregistrement. Quelques travaux confèrent à ces notes un rôle méthodologique plus important : retracer l'historique de la relation d'enquête, de sorte à replacer la conversation audible à l'enregistrement dans l'ensemble des interactions depuis la première prise de contact (Beaud et Weber 2003 : 229) ; ou encore caractériser la situation d'entretien « comme cadre d'observation » (*ibid.* : 183), c'est-à-dire décrire le lieu où se déroule la rencontre, les personnes présentes et leurs caractéristiques observables, les modes de présentation de soi (notamment non verbale) adoptés par les unes et les autres, etc. Pour autant, les usages possibles de ce matériau sont mal définis. Il n'existe pas de jurisprudence sur le sujet, et aucun débat scientifique majeur ne les a problématisés – alors que la déontologie vient couramment pallier les failles du droit (Stérin 2018).

En l'absence de ligne de conduite, les dépôts d'enquêtes qualitatives ont à arbitrer entre les apports scientifiques des notes méthodologiques et la non-connaissance de leur existence par les personnes enquêtées. Mes recours à des enquêtes beQuali dans le cadre d'enseignements m'ont appris que le premier enjeu y primait : le versement de notes de terrain pseudonymisées au titre de fichiers « descriptifs », « fiches » ou « synthèses » y est la norme. Après quelques hésitations, je m'apprête à suivre ce modèle pour mon propre dépôt quand, à quelques jours de la mise en ligne du jeu, un message de ma directrice de thèse réveille mes doutes : « Après coup je m'interrogeais juste un peu sur le fait de verser les métadonnées [notes méthodologiques], qui sont un peu en dehors

du contrat avec la personne, mais je ne sais pas ce qui se fait d'habitude. » Je me sens prise entre deux injonctions contradictoires. D'un côté, le « contrat » implicite de la relation d'enquête (Weber 2008) ne couvre que les échanges verbaux du temps de questions/réponses : les personnes ne se doutent pas que d'autres informations sont collectées et elles seraient susceptibles de s'y opposer. D'un autre côté, les pratiques professionnelles courantes (« ce qui se fait d'habitude ») banalisent cette collecte. Je retiens finalement du message de ma directrice un argument en faveur du non-versement : le rapport d'ignorance – et donc, par défaut, de non-consentement – des personnes enquêtées vis-à-vis de ces données. En amont de la mise en ligne, je retire donc du jeu les fichiers de notes méthodologiques sur chaque entretien, laissant seulement quelques lignes introductives dans les transcriptions (lieu, durée de l'entretien). Ce revirement me laisse une sensation mitigée, entre soulagement et frustration : il me libère de scrupules tenaces, mais repose sur une logique de responsabilisation individuelle qui, à mes yeux, révèle certaines limites des réflexions déontologiques au sein des sciences humaines et sociales (SHS) (Bendjaballah *et al.* 2018).

Dilemmes scientifiques dans les transformations de données

Le partage des transcriptions comme celui des exploitations statistiques a nécessité un travail de transformation des données de la recherche. L'anonymisation implique le retrait de certains éléments directement ou indirectement identifiants – l'option ici retenue étant le changement de tous les noms propres – tandis que la réalisation de compilations statistiques appelle à sélectionner certaines variables plutôt que d'autres et certains modes de codage plutôt que d'autres. Dans un cas, l'objectif est de brouiller l'identité des personnes interrogées ; dans l'autre, de constituer des catégories statistiques intelligibles et fiables (respect des contraintes d'effectifs limitant les subdivisions possibles). Si les motifs des transformations divergent, les implications sont convergentes : les données sont altérées, dans le sens d'une réduction des informations.

Là où les administrations valorisent cette standardisation des contenus (transformation de termes jargonants, uniformisation des formats…) comme une forme de « brutification » (Denis et Goëta 2017), la sociologie des sciences critique le concept de « données brutes » et s'attache plutôt à documenter les opérations de transformation réalisées (Latour et Woolgar 1979 ; Bowker 2008). Envisager les données de sciences sociales comme réutilisables au-delà d'une recherche singulière n'implique pas de nier leur construction dans le cadre d'un questionnement spécifique. Il s'agit au contraire de partager, outre les produits de la recherche, les processus de production, avec le double enjeu de dés-essentialiser les contenus mis à disposition et de permettre aux sciences sociales une réflexion incrémentale sur leurs pratiques (Laferté 2006). Selon ce même paradigme, je cherche à visibiliser les modifications que j'ai réalisées pour en rendre compte : dans les transcriptions anonymisées, je place les noms changés entre crochets, et dans les fichiers de statistiques, je détaille les modes de construction des variables.

Si, au regard des normes des sciences sociales, la nécessité de documenter mon traitement des données est évidente, le choix du type de traitement à adopter ne l'est pas. L'anonymisation comme le recodage de variables réduisent les informations disponibles : la dés-identification nécessaire à l'anonymisation entre en tension avec une caractérisation scientifique fine des personnes enquêtées (Monge 2016 ; Coulmont 2017 ; Bendjaballah *et al.* 2018), tandis que la construction d'indicateurs statistiques oblige à renoncer aux détails de variables à 10, 100, voire 1 000 modalités. Dans un cas comme dans l'autre, des arbitrages doivent être faits pour choisir les caractéristiques à restituer et celles à sacrifier. Dans les transcriptions anonymisées, que conserver des informations véhiculées par les noms de villes, d'organisations professionnelles, d'associations – entre autres ? Mettons,

exemple fictif, que je veuille anonymiser Perpignan. J'ai conscience que le département constitue un marqueur essentiel pour les analyses des politiques du handicap, celles-ci étant déployées en France à l'échelle départementale par les Maisons départementales des personnes handicapées (MDPH). Je pourrais donc indiquer [Ville des Pyrénées-Orientales]. En revanche, pour appréhender les inégalités urbaines, une indication de la taille de la ville, [Grande ville], ou de sa situation économique d'ensemble, [Ville pauvre], serait plus pertinente. Or, conserver les trois informations n'est pas envisageable : les Pyrénées-Orientales étant peu densément peuplées, cela réidentifierait immédiatement la ville. Pour ces substitutions comme pour le codage des variables de l'EEC et de son module, se pose également la question des frontières et des seuils. À partir de combien d'habitants parler de « grande ville », de quel niveau de vie médian ou taux de pauvreté parler de « ville pauvre » ? Pour le volet quantitatif, dans une optique de cumulativité des savoirs, il me semble souhaitable de favoriser des typologies statistiques courantes, comme les tranches de taille d'unité urbaine fournies par l'Insee. Mais ces classifications sont-elles pertinentes et suffisantes pour mon matériau qualitatif ?

Aboutissant à la conclusion que mes décisions seraient inévitablement marquées par mes perspectives de recherche, je décide de les assumer comme telles. Ainsi, je choisis les catégories d'analyse en fonction des grilles de lecture que je mobilise moi-même. Par exemple, pour l'anonymisation des noms d'organisations professionnelles dans les transcriptions, je retiens des critères de statut d'emploi, de domaine d'activité et de nombre de salarié·es[8] – plutôt que le chiffre d'affaires ou encore la date de création. Ce cadrage me sert à restituer certaines catégories structurantes dans mes analyses de thèse, comme l'emploi associatif et médico-social, ou dans la littérature que je discute, comme le seuil de 20 salarié·es dans les travaux sur les quotas de handicap. De la même manière, ma sélection et mon codage des variables de l'EEC s'inspirent des angles de recherche qui me sont familiers. Ce protocole admet dans une large mesure des découpages standardisés, comme ceux des statuts d'emploi[9] ou la nomenclature d'activités française[10]. Mais il peut conduire à subdiviser certaines modalités, par exemple en isolant l'emploi associatif des autres emplois privés.

En mobilisant ces modes de catégorisation, je privilégie donc implicitement certaines orientations scientifiques. Mais il m'importe que mes données puissent également servir à d'autres champs de recherche, ou d'autres disciplines. Dans cette optique, je cherche à ouvrir des possibilités de catégorisations alternatives. Pour mon jeu de données qualitatives, je réalise une table des correspondances des noms anonymisés (à l'exception des noms de personnes), non versée au jeu de données mais présentée dans la documentation avec l'indication que je pourrai la transmettre sur demande aux projets de recherche qui auraient besoin d'extraire d'autres informations. En parallèle, dans la documentation de mon jeu de données quantitatif, j'invite le public de l'entrepôt à faire remonter ses souhaits pour de futurs versements complémentaires.

Une appropriation ambivalente

Dans mon expérience de l'ouverture des données, la confusion induite par les difficultés rencontrées s'est doublée de perplexité face à mon propre comportement. Mais que suis-je allée faire dans cette galère ? Un retour réflexif sur mon travail d'auto-dépôt m'invite à distinguer les déterminants contextuels qui m'ont menée à entamer la démarche et les valeurs qui ont soutenu mon engagement au fil du temps – sans empêcher certains doutes, que je présenterai.

Les incitations contextuelles : prescriptions et conditions de possibilité

Mon initiation d'un auto-dépôt est à lire à la lumière d'une époque, d'un environnement professionnel et d'une relation de travail. J'entame ma thèse en 2018, alors que les textes en faveur de l'ouverture des données de

la recherche se succèdent. De plus, je suis affiliée à un laboratoire interdisciplinaire, soit un milieu plus accoutumé que la majorité des centres de SHS au langage des « données de la recherche » et à la perspective de leur ouverture (Revelin *et al.* 2021). Enfin, c'est ma directrice de thèse, membre actif de ce laboratoire, qui – pour les deux volets de la recherche – soulève l'idée d'une mise à disposition des données. La relation hiérarchique colore ma réception de cette proposition, venant à la fois stimuler la démarche et la permettre. D'un côté, je reconnais une forte légitimité à cette encadrante, ce qui me pousse à suivre ses conseils. D'un autre côté, l'avis de ma directrice de thèse signale qu'elle est convaincue de la qualité de mes données, ce qui contribue à désamorcer certaines appréhensions. Les défauts supposés d'anonymisation (Siméant-Germanos 2022) et la réfutation des résultats (Duchesne et Garcia 2014) exposent les chercheur·es sans poste davantage que leurs collègues titulaires. Plus indirectement, mes jeux de données sont susceptibles de servir d'outils d'évaluation sur un marché académique soumis à une très forte pression. Ma conduite d'entretien est-elle suffisamment bonne pour que je la donne à voir ? Et mes pratiques de codage ? Le jugement implicite de ma directrice de thèse rend cette perspective plus facile à envisager.

Le renouvellement de l'engagement à la croisée de plusieurs valeurs

Si ces déterminants sont importants, ils ne permettent pas à eux seuls de comprendre que j'aie mené le processus d'auto-dépôt à son terme. Quand, quelques semaines après le début de mon entreprise de pseudonymisation des entretiens, mon ancienne directrice de thèse mesure le coût en temps de cette démarche, elle me recommande d'arrêter. Je ne suis pas son conseil, alors même que je vis le processus comme long et fastidieux. Le renouvellement de mon engagement s'est adossé à certaines valeurs, qu'il s'agit maintenant de présenter.

Distinguons trois grands groupes de valeurs. En premier lieu, il m'importe de partager mes données avec mes collègues, à commencer par la jeune collègue précédemment évoquée. Deuxièmement, utilisant Ubuntu, LibreOffice et R au quotidien, je m'inscris également dans une « culture du libre » favorable au partage des données. Ces deux motivations sont bien documentées parmi les chercheur·es (Revelin *et al.* 2021). Elles portent la trace de socialisations familiales et professionnelles ; dans mon cas, l'installation de Linux par ma sœur aînée, ou l'usage répandu de R dans mes centres de recherche. Troisième type de valeur, moins exploré par la littérature : je vois l'entreposage des données de ma recherche comme une action en faveur de la mise à l'agenda académique d'un sujet qui me semble important et insuffisamment traité. Cette conception entremêle des logiques scientifiques, éthiques et politiques qui méritent d'être détaillées.

Depuis que j'ai commencé à travailler sur le handicap, en master, j'ai pu constater que cet angle de recherche n'était pas courant. Le constat que les études sur la stratification sociale couvraient peu le handicap a d'ailleurs été le point de départ de mon projet de thèse. Ce défaut de connaissances est scientifiquement dommageable. Plusieurs des personnes rencontrées en entretien lui prêtent aussi des conséquences sociales et politiques, ce qui a coloré leur réception de la proposition d'entreposage : « Je pense que c'est bien : parce que… enfin, en général, le sujet du handicap est très mal traité[11] » ; « Si ça peut être utile à d'autres, et que ça permet de faire avancer le sujet […] tous les gens qui pourront participer à faire avancer la *(dans un rire tendu)* la "cause", entre guillemets… je dirai toujours oui[12] ! ». La mise à disposition des entretiens pseudonymisés au sein de la communauté académique est ainsi investie comme un vecteur pour contribuer à un meilleur traitement scientifique du « sujet » du handicap ; et, par ricochet, pour le visibiliser comme « cause » dans l'espace politique. Dans cette perspective, non seulement une publicisation contrôlée de leurs données ne pose pas de problème déontologique mais elle répond le plus souvent à une attente. L'entreposage des entretiens participe donc à mon contre-don aux

personnes enquêtées pour leur temps et leur confiance. Enfin, le versement de mes données a une portée politique, dans la mesure où il contribue à intégrer le handicap aux réflexions collectives (*disability mainstreaming*). Cet aspect est particulièrement saillant pour le volet quantitatif de ma thèse : en proposant des relectures d'indicateurs multiples (diplôme, situation professionnelle, logement, vie familiale…) sous l'angle du handicap, je crée un réservoir de statistiques[13] qui pourra non seulement informer les travaux de sciences sociales, mais aussi nourrir le débat public – à l'image de la forte politisation des statistiques de l'égalité entre hommes et femmes (Blanchard et Pochic 2021). Le partage de données quantitatives marque aussi mon attachement à une approche statistique sur le handicap, en tant qu'outil pour objectiver et analyser certaines inégalités sociales. Ce faisant, je me positionne face à des points de vue plus hostiles à ces statistiques, qui craignent que la catégorisation ne réifie le handicap (Barnes et Oliver 2012) – selon une opposition qui traverse également les controverses sur les statistiques ethniques (Simon 2008).

Des finalités mal définies et en reconfiguration

L'auto-dépôt des deux jeux de données a nécessité un volume de travail que je n'imaginais pas en amont, de l'ordre de plusieurs heures par jour durant six mois. Réalisant un CDD de recherche en journée, je grappille ce temps sur mes pauses déjeuner, mes soirées, mes weekends. Si cela rend malaisé d'en chiffrer l'ampleur, je sais que l'ensemble se compte en centaines d'heures. Cette charge de travail me pèse lourdement. Le soir de juin où je mets en ligne mon deuxième jeu de données, j'écris à des amies doctorantes : « Fini le gouffre de perte de temps sans fin ».

Tout au long de l'entreposage, mon vif ressenti de son coût temporel soulève parfois des doutes sur ses débouchés concrets. Je me demande si cela en vaut vraiment la peine, si l'absence de rémunération (travail gratuit) s'accompagne en plus d'une absence de réutilisations effectives (travail inutile). Ces inquiétudes sont d'autant plus tenaces que je n'ai qu'une vague idée des publics qui pourraient avoir l'usage des jeux, au-delà de l'avatar de mon amie doctorante. En outre, même pour elle, je doute qu'une réutilisation des entretiens soit possible, même en complément d'un corpus déjà fourni. En France, la norme de récolte de ses propres données reste en effet prégnante parmi les sociologues qualitativistes, en tension avec l'injonction à réaliser un nombre d'entretiens toujours plus élevé, par l'application de normes quantitatives aux enquêtes qualitatives (Beaud 1996). Lorsque, dès janvier 2022, je partage ces doutes à demi-mot avec ma directrice de thèse (sans évoquer le cas spécifique de la doctorante), elle me répond en mettant en avant l'intérêt qu'aura de toute façon ma démarche « pour des historiens dans cent ans ». Valorisant ma contribution à un projet patrimonial défendu par les archivistes (Both et Garcia 2014), cette remarque se veut un encouragement. Elle me laisse néanmoins une impression mitigée, érodant mon espoir d'être utile à mes collègues proches.

Une autre finalité, inattendue, m'apparaît en revanche fin mars 2022. À l'occasion d'un échange concernant l'entreposage des notes méthodologiques issues des entretiens, ma directrice de thèse commente :

> « Pendant que c'est encore frais, ce serait super intéressant que tu écrives sur l'expérience du travail à fournir pour permettre cet archivage : demande d'autorisation en amont, travail d'anonymisation, documents de support… Je trouve que ça ferait un super article méthodologique, je ne sais pas exactement dans quel support de publication, mais à partir du moment où tu as déjà fait tout ce travail, je trouve que ça vaudrait vraiment le coup de faire ce petit effort supplémentaire […] en tout cas je suis sûre que cela intéressera la collectivité. » (Réponse de mon ancienne directrice de thèse à un email intitulé « Réutilisation d'entretiens », le 29 mars 2022)

Sur le moment, la piste me paraît irréaliste tant je croule sous le travail. Mais l'argument fait mouche, me donnant à voir un éventuel débouché au moment où je lutte contre un sentiment de vacuité. Je commence à ce

moment à prendre quelques notes réflexives. À l'issue de l'auto-dépôt, en juin 2022, celles-ci me fournissent le squelette de ce qui deviendra ce manuscrit.

Dans le même temps, je reçois aussi un soutien actif de l'équipe de soutien à la recherche. Après m'avoir orientée sur le versant technique de l'entreposage, plusieurs membres de l'équipe me proposent des pistes pour perfectionner les jeux et les visibiliser suite à leur mise en ligne : préparation de contenus supplémentaires, participation à des événements scientifiques, rédaction d'un *data paper*… Nous fixons plusieurs rendez-vous entre juin 2022 et septembre 2022 pour discuter de ces idées. J'abandonne rapidement l'idée du *data paper*, comprenant que mes jeux de données correspondent peu aux normes de l'exercice – conçues pour des données statistiques primaires sans restriction d'accès et généralement en langue anglaise. En revanche, m'obstinant dans l'idée de rendre mes jeux de données opérationnels pour des collègues, je formule l'idée de réaliser un court questionnaire à l'usage des personnes qui ont consulté ces jeux de données ou qui envisagent de mobiliser ce type de données (statistiques ou entretiens biographiques) dans le cadre de recherches sur les inégalités sociales ou sur le handicap. Les réponses recueillies serviraient à prioriser les mises à jour ou les prolongements à effectuer. Stimulée par l'approbation du personnel de soutien à la recherche, je prépare ce questionnaire en septembre 2022 et j'organise sa diffusion par plusieurs vecteurs : sur une boucle mail thématique sur le handicap (environ 250 abonné·es), sur les réseaux sociaux de ma bibliothèque universitaire (environ 8 000 abonné·es), par des groupes d'information et de documentation accompagnant l'ouverture des données de la recherche (trois listes de diffusion totalisant des centaines d'abonné·es invité·es à relayer auprès de leurs communautés de recherche respectives)… L'initiative aboutit en tout et pour tout à 6 réponses[14]. Ce très faible taux de participation me semble confirmer l'intérêt limité des chercheur·es en SHS pour la réutilisation de mes jeux. Ironiquement, la publicité induite par la diffusion du questionnaire a en revanche amplifié l'intérêt des personnels de soutien à la recherche pour mon expérience d'auto-dépôt : rencontre avec le pôle communication du Programme prioritaire de recherche sur l'autonomie, invitation à la semaine DataSHS… Ces opportunités réorientent radicalement les débouchés scientifiques de mon travail d'auto-dépôt : davantage que mes données elles-mêmes, c'est leur processus d'ouverture qui suscite l'intérêt.

*****

Dans cet article, la confusion ressentie durant l'auto-dépôt d'une enquête sociologique par méthodes mixtes a servi d'entrée pour questionner le processus d'ouverture des données : les variations de temporalité et de degré de formalisation selon les caractéristiques des données, les impensés des référentiels juridiques et scientifiques, les tenants et aboutissants d'un engagement ambivalent.

L'usage méthodologique que je fais de mon sentiment de confusion, en le traitant comme une entrée pour l'analyse d'un champ en évolution rapide, ne doit pas occulter les enjeux qui entourent cette charge mentale elle-même. L'auto-dépôt est un travail non rémunéré, et pourtant éprouvant. Les exigences émotionnelles et les conflits de valeurs constituent des formes de pénibilité (Gollac 2012) qui s'ajoutent à l'investissement temporel à fournir. Si la communauté académique a su se positionner face aux risques de bureaucratisation de l'ouverture des données, une réflexion collective semble maintenant indispensable pour encadrer le déroulement de cette activité : lui allouer les moyens économiques et humains nécessaires, mutualiser des arbitrages scientifiques et déontologiques qui, pour le moment, reposent sur la responsabilisation individuelle[15].

Ouvrages cités

BARNES, Colin et Mike OLIVER. 2012. *The New Politics of Disablement*. New York, Palgrave Macmillan.

BEAUD, Stéphane. 1996. « L'usage de l'entretien en sciences sociales. Plaidoyer pour l'"entretien ethnographique" », *Politix. Revue des sciences sociales du politique*, nº 35 : 226-257.

BEAUD, Stéphane et Florence WEBER. 2003. *Guide de l'enquête de terrain. Produire et analyser des données ethnographiques*. Paris, La découverte.

BENDJABALLAH, Selma, Sarah CADOREL, Émilie FROMONT, Guillaume GARCIA, Émilie GROSHENS et Emeline JUILLARD. 2018. « Anonymat et confidentialité des données : l'expérience de beQuali », *in* Véronique Ginouvès et Isabelle Gras (dir.), *La diffusion numérique des données en SHS. Guide de bonnes pratiques éthiques et juridiques*. Aix-en-Provence, Presses universitaires de Provence : 207-222.

BERTAUX, Daniel. 1997. *Le récit de vie*. Paris, Nathan.

BLANCHARD, Soline et Sophie POCHIC. 2021. *Quantifier l'égalité au travail. Outils politiques et enjeux scientifiques*. Rennes, Presses universitaires de Rennes.

BOTH, Anne et Guillaume GARCIA. 2014. « Le chercheur, l'archiviste et le webmaster : la polyphonie patrimoniale ? Le cas de beQuali, banque d'enquêtes qualitatives en sciences sociales », *in* Bernadette Saou-Dufrene et Benjamin Barbier (dir.), *Heritage and Digital Humanities: How Should Training Practices Evolve?*. Berlin, LIT Verlag : 353-364.

BOWKER, Geoffrey C. 2008. *Memory Practices in the Sciences*. Cambridge, MIT Press.

CAPORALI, Arianna, Amandine MORISSET et Stéphane LEGLEYE. 2015. « La mise à disposition des enquêtes quantitatives en sciences sociales. L'exemple de l'Ined », *Population*, vol. 70, nº 3 : 567-597.

CORNILLEAU, Anne et Emmanuelle DUWEZ. 2022. « Elipss, un dispositif inédit d'enquêtes pour la recherche en sciences sociales », *in* Emmanuelle Duwez et Pierre Mercklé (dir.), *Un panel français. L'étude longitudinale par Internet pour les sciences sociales (Elipss)*. Paris, Ined éditions : 15-44.

COULMONT, Baptiste. 2017. « Le petit peuple des sociologues. Anonymes et pseudonymes dans la sociologie française », *Genèses. Sciences sociales et histoire*, nº 107 : 153-175.

DENIS, Jérôme et Samuel GOËTA. 2017. « Les facettes de l'Open Data : émergence, fondements et travail en coulisses », *in* Pierre-Michel Menger et Simon Paye (dir.), Big data *et traçabilité numérique. Les sciences sociales face à la quantification massive des individus.* Paris, Collège de France (Conférences) : 121-138.

DUCHESNE, Sophie et Guillaume GARCIA. 2014. « BeQuali : une archive qualitative au service des sciences sociales », *in* Marie Cornu, Jérôme Fromageau et Bertrand Müller (dir.), *Archives de la recherche. Problèmes et enjeux de la construction du savoir scientifique*. Paris, L'Harmattan : 35-56.

DUCHESNE, Sophie et Camille NOUS. 2019. « Apories de la mise en banque : retour d'expérience sur la réutilisation d'enquêtes qualitatives », *Tracés*, nº 19 : 89-100.

GAY, Victor. 2021. « Un *data paper* en SHS : pourquoi, pour qui, comment ? » (en ligne), communication à un congrès. URL : https://hal.science/hal-03434216.

GOLLAC, Michel. 2012. « Les risques psychosociaux au travail : d'une "question de société" à des questions scientifiques. Introduction », *Travail et emploi*, nº 129 : 5-10.

LAFERTE, Gilles. 2006. « Des archives d'enquêtes ethnographiques pour quoi faire ? Les conditions d'une revisite », *Genèses. Sciences sociales et histoire*, nº 63 : 25-45.

LATOUR, Bruno et Steve WOOLGAR. 1979. *Laboratory Life: The Construction of Scientific Facts*. Princeton, Princeton University Press.

LAURENS, Sylvain. 2022. « L'ethnographie en procès. Enjeux contemporains autour de l'éthique de l'enquête de terrain », *Genèses. Sciences sociales et histoire*, nº 12 : 7-13.

MERRILL, Barbara et Linden WEST. 2009. *Using Biographical Methods in Social Research*. Londres, Sage Publications.

MONGE, Julia. 2016. « Écrire sans trahir. Les impératifs scientifiques du doctorant face aux contraintes éthiques », *Le sociographe*, nº 54 : 73-86.

MOORE, Niamh. 2007. « (Re)Using Qualitative Data? », *Sociological Research Online*, vol. 12, nº 3 : 1-13.

MOUTON, Marie-Dominique. 2018. « Dématérialisation et valorisation des matériaux de terrain des ethnologues L'archiviste face aux questions éthiques », *in* Véronique Ginouvès et Isabelle Gras (dir.), *La diffusion numérique des données en SHS. Guide de bonnes pratiques éthiques et juridiques*. Aix-en-Provence, Presses universitaires de Provence : 73-87.

REVELIN, Florence, Alix LEVAIN, Morgane MIGNON, Marianne Noël, Betty Queffelec, Pascal Raux et Hervé SQUIVIDANT. 2021. « L'ouverture des matériaux de recherche ethnographiques en question », rapport de recherche, Centre national de la recherche scientifique.

RUGGIANO, Nicole et Tam E. PERRY. 2019. « Conducting Secondary Analysis of Qualitative Data: Should We, Can We, and How? », *Qualitative Social Work*, vol. 18, nº 1 : 81-97.

SCOT, Marie. 2006. « Les archives britanniques des sciences sociales. Deux études de cas : UK Data Archive (UKDA) et Qualidata », *Genèses. Sciences sociales et histoire*, nº 63 : 46-65.

SIMEANT-GERMANOS, Johanna. 2022. « Qui protéger, consentir à quoi, enquêter comment ? Les sciences sociales face à la bureaucratisation de la vertu scientifique », *Genèses. Sciences sociales et histoire*, nº 129 : 66-87.

SIMON, Patrick. 2008. « Les statistiques, les sciences sociales françaises et les rapports sociaux ethniques et de "race" », *Revue française de sociologie*, vol. 49, nº 1 : 153-162.

STERIN, Anne-Laure. 2018. « Diffuser des données de la recherche dans le respect du droit et de l'éthique », *in* Véronique Ginouvès et Isabelle Gras (dir.), *La diffusion numérique des données en SHS. Guide de bonnes pratiques éthiques et juridiques*. Aix-en-Provence, Presses universitaires de Provence : 19-29.

STERIN, Anne-Laure et Camille NOUS. 2019. « Ouverture des données de la recherche : les mutations juridiques récentes », *Tracés*, nº 19 : 37-50.

WEBER, Florence. 2008. « Publier des cas ethnographiques : analyse sociologique, réputation et image de soi des enquêtés », *Genèses. Sciences sociales et histoire*, nº 70 : 140-150.

Annexe

Chronologie de la production et de la mise à disposition des données

Septembre 2018 : Demande d'accès aux données de l'Enquête emploi en continu (EEC) et de son module *ad hoc* auprès de l'infrastructure de production et gestion des données, via le service Archives de données issues de la statistique publique (Progedo-Adisp).

Octobre 2018 à juin 2022 : Réalisation de statistiques issues de ces deux sources.

Novembre 2019 : En préparation du lancement de la campagne d'entretiens, rédaction d'une fiche d'information à remettre aux personnes rencontrées, incluant une demande d'autorisation au « versement à un fond institutionnel sécurisé ».

Janvier 2020 : Rendez-vous avec la DPD, validation du protocole.

Décembre 2019 à juin 2020 : Entretiens, collecte progressive des autorisations au versement (sous forme écrite ou orale).

Décembre 2019 à février 2021 : Transcriptions des entretiens (seuls les noms des personnes sont immédiatement changés).

Juillet 2021 : Prise de contact avec les ingénieur·es et bibliothécaires de mon université en charge de la gestion des données, puis avec l'Adisp, concernant les possibilités de versement de données secondaires à l'entrepôt de données de la recherche de l'université.

Novembre 2021 : Dépôt du manuscrit de thèse, marquant la fin de l'enquête.

Janvier 2022 : Après la soutenance, reprise de contact avec l'équipe en charge de la gestion des données et avec la DPD, pour préparer l'entreposage des deux volets de la thèse.

Janvier 2022 à mars 2022 : Reprise des transcriptions pour en corriger les coquilles et supprimer tous les noms propres.

Mars 2022 : Création des jeux de données en version hors-ligne, à l'aide du *data librarian* de mon université.

Avril 2022 : Mise en ligne du jeu de données associé au volet qualitatif.

Avril 2022 à juin 2022 : Production de statistiques complémentaires et rédaction détaillée des scripts associés. Création de documents de synthèse compilant les tableaux et graphiques par thématique.

Juin 2022 : Mise en ligne du jeu de données associé au volet quantitatif.

---

[1] Sont concernés notamment les programmes financés par appels à projets sur fonds publics français et certaines recherches financées sur fonds publics européens.

[2] Loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées. URL : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000809647/.

[3] Une personne s'est déclarée indécise et deux autres ont demandé à ce que les demandes leur soient directement soumises ; je les ai finalement comptées comme des refus.

[4] Nous devons y voir un enjeu déontologique plutôt que juridique ; comme la DPD me le confirme quelques jours plus tard, le changement de tous les noms propres ne correspond toujours qu'à une « pseudonymisation » simple au sens du RGPD.

[5] Le couplage de formats graphiques et de tableaux s'inscrit dans une série de mesures destinées à faciliter la consultation des statistiques par des publics divers, y compris des personnes concernées par différentes formes de troubles ou déficiences.

[6] J'incorpore également des précisions méthodologiques nombreuses au fil des autres documents : informations sur les variables et leurs modalités dans les compilations, balises de commentaires au fil des scripts R, source et champ en bas de chaque tableau.

[7] Le service chargé de diffuser les enquêtes et bases de données produites par la statistique publique auprès de la communauté scientifique, Progedo-Adisp (Archives de données issues de la statistique publique, de l'infrastructure de recherche Production et gestion de données), rappelle sur son site Internet les obligations des chercheur·es auxquel·les il transmet des données. URL : http://www.progedo-adisp.fr/acces_engagements.php.

[8] Je détaille les trois informations à la première occurrence du nom puis les abrège lors des suivantes, pour alléger la lecture.

[9] Nomenclature disponible sur le site de l'Insee, URL : https://www.insee.fr/fr/metadonnees/definition/c1792 (consulté le 11/12/2022).

[10] Nomenclature disponible sur le site de l'Insee, URL : https://www.insee.fr/fr/metadonnees/nafr2/ (consulté le 11/12/2022).

[11] Entretien avec Jackie Raynal, trentenaire dys, avril 2020.
[12] Entretien avec Gregory Prigent, trentenaire dys, juin 2020.
[13] Un projet international porté par Fordham University, la Disability Data Initiative, œuvre également en ce sens. URL : https://disabilitydata.ace.fordham.edu/.
[14] À sa clôture le 23 décembre 2022, plus de quatre mois après sa mise en ligne.
[15] Cet article a bénéficié du soutien apporté par l'Agence nationale de la recherche (ANR) et l'État au titre du programme d'Investissements d'avenir dans le cadre du Labex LIEPP (Laboratoire interdisciplinaire d'évaluation des politiques publiques, ANR-11-LABX-0091, ANR-11-IDEX-0005-02) et de l'Idex Université Paris Cité (ANR-18-IDEX-0001). Je remercie Anne Revillard pour l'idée de cet article méthodologique, et Cyril Heude, Guillaume Garcia, Paul Colin et le comité de rédaction de *Genèses* pour leurs remarques sur différentes versions du manuscrit.