La Source.
Institut et Haute
Ecole de la Santé

Partage des données de recherche.

Un enjeu capital pour une poursuite durable du développement des connaissances.

Rapport destiné au dicastère Recherche et innovation de la HES-SO – appel à projets *Open Data* HES-SO 2020

Myriam Guzman Villegas-Frei, MSc ScI

Jonathan Jubin, PhD

Annie Oulevey Bachmann, PhD

Lausanne, le 29.10.2021.

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

Table des matières

1. Résumé

L'*Open Data* vise à universaliser le partage libre de données de recherche. Cette proposition permet à la science de gagner en fiabilité et en réplicabilité tout en offrant aux chercheur·e·s l'opportunité de valoriser leurs travaux. Néanmoins, pour ce qui concerne le domaine de la santé, elle présente également certains risques : notamment le non-respect de la protection des données et leur utilisation détournée.

Dans le présent rapport est exposé le processus que nous avons suivi à l'occasion du partage de nos propres données. L'étude dont elles sont issues était longitudinale et portait sur le bien-être psychique des étudiant·e·s de Bachelor HES. Vu la nature des données recueillies (données de santé), elle avait nécessité l'aval d'une commission d'éthique pour être conduite. Cette commission nous avait autorisé·e·s à partager les données à la condition que par la suite, tout·e chercheur·e souhaitant les obtenir présente un document attestant de l'examen du projet par une commission d'éthique reconnue et de son autorisation.

Nous avons donc recherché une plateforme d'archivage en ligne (*repository*) acceptant les données de sciences humaines et sciences de la santé. Elle devait nous permettre, entre autres, de contrôler les accès à nos données. Notre choix s'est porté sur FORSbase (qui évolue actuellement vers SWISSUbase). Cette plateforme satisfaisait à tous nos critères en plus d'être gratuite.

Le processus de préparation des données a nécessité un travail conséquent afin de les rendre compatibles avec les normes FAIR. En nous y conformant, nous nous sommes assurés·e·s que nos données puissent être trouvées (*Findable*), soient accessibles (*Accessible*), interopérables (*Interoperable*) et réutilisables (*Reusable*). Cette étape a donc impliqué de décrire méticuleusement l'intégralité du projet de recherche ainsi que le contenu des jeux de données afin que quiconque y accédant puisse les comprendre et les utiliser aisément.

En conclusion, le partage de données, ici des données de santé, est un processus complexe qui implique des questionnements bien plus complexes que ce à quoi nous nous attendions. Les risques qu'il présente peuvent être contrôlés par l'adhérence stricte aux normes éthiques et à celles proposées dans la littérature scientifique. Comme l'*Open Data* représente un enjeu capital pour une poursuite durable et transparente du développement des connaissances, il est crucial que les équipes de recherche se préparent à ce qui est considéré comme un nouveau paradigme. Cette préparation passe par la formation des chercheur·e·s et l'intégration du concept de partage des données dès la conception des protocoles de recherche.

2. Introduction

Alors que la digitalisation de la recherche se développe massivement en sciences humaines et en sciences de la santé (Duca & Metzler, 2019; Nagy Hesse-Biber, 2011), les feuilles-réponses stockées sous clé dans les armoires des bureaux des chercheur·e·s apparaissent de plus en plus comme des reliques d'un autre temps. Les nouvelles technologies rendent notamment collecte de données plus aisée que jamais (Pang et al., 2018) : passation de questionnaire ou de sondages en ligne au moyen de logiciels attractifs et simples d'utilisation ; possibilité de paramétrer les réponses à des fins de qualité des données collectées ; ou encore exportation directe des données dans des logiciels d'analyses quantitatives ou qualitatives. Ces technologies permettent aussi d'atteindre de grandes quantités de participant·e·s potentiel-le-s : réponses où qu'ils ou elles se trouvent et sur divers supports (ordinateurs, tablettes, smartphones) ; accès facilité aux questionnaires via des hyperliens ou des QR codes. Il en résulte un accroissement potentiellement colossal de la quantité de données récoltées pouvant être mises facilement et rapidement à disposition de la communauté scientifique mondiale (International Science Council, 2017). C'est dans ce contexte que se développe le courant de l'*Open Data*.

Les données d'une étude peuvent donc aujourd'hui être exploitées au meilleur de leur potentiel. Leur partage à distance donne l'occasion de collaborer aisément à l'échelle nationale ou internationale (Lowndes et al., 2017). Lors de la soumission d'articles scientifiques, il permet aux *reviewers* d'inspecter minutieusement les sections résultats. Une fois un article publié, il offre la possibilité aux lecteur·rice·s de vérifier ou approfondir les analyses effectuées à partir des données mises à disposition. Grâce au partage des données, la recherche gagne donc en réplicabilité et en fiabilité. Les données peuvent aussi être réutilisées dans le cadre d'études secondaires, telles des méta-analyses. De plus, cette utilisation secondaire par des chercheur·e·s tiers permet à celles et ceux qui en sont à l'origine de gagner en visibilité et de voir leur travail de collecte valorisé. Elle permet aussi de contenir les coûts de la recherche en évitant de multiplier des collectes de données similaires. Finalement, elle évite de sursolliciter les populations d'intérêt (Piwowar, 2011).

Pour toutes les raisons évoquées ci-dessus, et depuis quelques années, de nombreux·ses scientifiques, organismes finançant la recherche, ou milieux académiques prônent le partage aussi large que possible des données. Néanmoins, cette proposition présente des risques (Goben & Sandusky, 2020). Par exemple, dans une enquête en ligne, les données sont récoltées *via* un logiciel de sondage et stockées sur un serveur rattaché à cet outil. Un·e chercheur·e les télécharge ensuite sur son ordinateur afin de les analyser et les transmet potentiellement aux personnes avec qui il ou elle collabore. Pour ce faire, un courriel ou la copie des données sur un support physique comme une clé USB peuvent être utilisés. Chaque ordinateur professionnel sur lequel les données sont copiées est probablement lié à un *cloud* institutionnel, ou externe, où elles seront sauvegardées. Ce *cloud* peut reposer sur des serveurs situés à plusieurs endroits différents, dans ou hors du pays où les données ont été collectées, afin de garantir l'intégrité des informations qu'ils recèlent en cas de dommages matériels.

En très peu de temps, il peut donc exister des dizaines de copies informatiques de données empiriques éparpillées aux quatre coins du monde. Cet état de fait est particulièrement problématique dans le cas de données sensibles contenant des informations privées, telles que des données de santé (Templ et al., 2014). Le développement de l'accès à ces données ne va donc pas sans risques : leur authenticité et leur intégrité peuvent être compromises, elles peuvent être détournées de leur finalité initiale, exploitées à des fins commerciales ou être volées alors qu'elles n'ont pas encore été publiées. Leur emploi par des chercheur·e·s tiers peut mener à des interprétations erronées voire abusives, si elles ne sont pas accompagnées d'une documentation claire.

Le développement rapide et exponentiel des connaissances nécessaires pour contrer la pandémie COVID-19, et les dérives dénoncées, l'illustrent magistralement depuis presque 18

mois : l'accès aux données de santé est hautement souhaitable, mais sans pour autant que n'importe qui puisse en faire n'importe quoi.

3. Contexte de notre étude

Notre équipe est constituée de chercheur·e·s du domaine de la santé appartenant à l'Institut et Haute Ecole de la Santé la Source, un site de la Haute Ecole Spécialisée de Suisse Occidentale (HES-SO). Nous nous intéressons à la manière dont il est possible de développer la prévention et de promouvoir la santé dans les activités humaines telles les activités professionnelles, estudiantines ou d'aide informelle. Nous avons obtenu un financement interne du domaine santé de la HES-SO en 2019 pour examiner le niveau de stress perçu, le bien-être psychique et les facteurs susceptibles de protéger ce dernier chez des étudiant·e·s Bachelor de la HES-SO, et comment ces paramètres évoluaient à 12 mois d'intervalle (acronyme de l'étude : HEalStud[1]).

Pour les raisons décrites ci-dessus, et parce que les données collectées pouvaient particulièrement intéresser la communauté HES-SO dans son ensemble, nous avons d'emblée annoncé vouloir partager nos données de recherche une fois le travail de valorisation effectué (publication d'articles scientifiques, professionnels et présentations orales). Collectant des données de santé, cette étude tombait sous le coup de la Loi relative à la recherche sur l'être humain (LRH). Elle a donc dû recevoir l'autorisation de la Commission cantonale d'éthique de la recherche sur l'être humain CER-VD (numéro de projet 2019-01379). Nous avons dû nous conformer aux directives concernant la gestion et la protection des données. En particulier, elles devaient être enregistrées en tout temps sur un serveur situé physiquement en Suisse et elles devaient être anonymisées ou codées (selon les situations) pour toute utilisation.

En mars 2020 (T0), les étudiant·e·s de tous, sauf deux, établissements d'enseignement et de recherche membres de la HES-SO ont été invité·e·s à remplir un questionnaire en ligne portant sur leur qualité de vie, leur exposition perçue à des stresseurs, leur sentiment d'auto-efficacité, leur capacité de résilience, leur disposition à la pleine conscience et le soutien social qu'ils et elles recevaient. Celles et ceux qui le souhaitaient ont pu donner leur accord d'être recontacté·e·s une année plus tard pour une seconde mesure visant à analyser l'évolution de ces éléments.

Or, la pandémie de COVID-19 a frappé la Suisse juste après T0 et entraîné l'instauration de mesures sanitaires drastiques : distanciation sociale, enseignement à distance, déploiement de certain·e·s étudiant·e·s dans les milieux de soins, limitation des ressources de santé offertes par les loisirs, etc. En conséquence, le deuxième temps de mesure de notre étude, qui s'est déroulé en mars 2021 (T1), permet de brosser un portrait de la condition estudiantine dans la HES-SO après le passage de deux vagues de pandémie, l'enseignement à distance et l'isolement vécus par les étudiant·e·s.

Face à cette situation exceptionnelle, et afin de comprendre au mieux le vécu de certain·e·s étudiant·e·s, nous avons effectué une mesure supplémentaire en septembre 2020 (T0.5). L'acronyme de ce bras de l'étude HEalStud est HEalS-Nu. Un questionnaire identique à celui de T0 (données quantitatives), auquel nous avons ajouté une échelle de croissance posttraumatique et quelques questions en lien avec le COVID-19, a été soumis : (i) aux étudiant·e·s en soins infirmiers qui avaient accepté d'être recontacté·e·s pour T1 ; (ii) à l'ensemble du corps estudiantin de l'Institut et Haute Ecole de la Santé La Source. Nous savions

[1] Degré d'exposition des étudiants Bachelor HES à des stresseurs, relations avec leur santé mentale perçue et exploration de facteurs susceptibles de la protéger : une étude longitudinale. Myriam Guzman Villegas-Frei, Claudia Ortoleva Bucher, Jérôme Pasquier, Meichun Mohler-Kuo, Annie Oulevey Bachmann. N° SAGE-X 95592.

qu'une partie de ces étudiant·e·s avait été déployée afin de soutenir les institutions de soins au plus fort de la première vague COVID-19. L'idée était d'explorer les liens entre les stresseurs auxquels ces personnes avaient été exposées durant cette période, et leur bien-être psychique. Certaines d'entre elles ont également participé à des *focus-groups* lors desquels elles ont pu nous faire part de la façon dont elles avaient vécu cette période (données qualitatives).

Concrètement, les données de ces deux études étaient donc de nature quantitative (HEalStud et HEalS-Nu) et qualitative (HEalS-Nu). Elles étaient constituées des réponses de 2534 participants à T0, 1223 participants à T1 et 418 à T0.5 ainsi que les retranscriptions des *focus-groups* s'étant déroulés à T0.5.

Ces données uniques pourraient profiter directement à toutes les recherches s'intéressant, par exemple, à l'impact de la pandémie ou à la santé des étudiant·e·s, en particulier des étudiant·e·s en soins infirmiers. Il était donc encore plus important que nous les partagions. Toutefois, comme requis par la CER-VD, toute utilisation secondaire des données devrait être soumise à autorisation d'une commission d'éthique compétente concernant l'utilisation qui pourrait être faite des données.

À ces contraintes juridiques et éthiques, s'ajoutaient des impératifs pratiques. Afin de pouvoir référencer les données dans nos publications, il était nécessaire qu'un *Digital Object Identifier* (DOI) leur soit attribué. Il s'agit d'une adresse internet durable renvoyant à la page permettant l'accès aux données. De plus, nous avions besoin de placer les données sous embargo, c'est-à-dire, tout en signalant leur existence au monde scientifique, bloquer temporairement leur accès jusqu'à ce que nous ayons terminé de valoriser nos résultats de recherche.

Aussi, nous avons examiné quelle pouvait être la meilleure manière de procéder, tout en préservant l'intégrité des données des deux recherches et leur sécurité. Nous décrivons par la suite plus précisément nos besoins, et les recherches effectuées, afin de trouver la solution de d'archivage à même de leur répondre au mieux.

4. Choix de la plateforme d'archivage

De très nombreuses plateformes d'archivage de données scientifiques existent dans le monde. Financé par la Fondation allemande pour la recherche (*German Research Foundation*) et recommandé par le programme Horizon 2020 de la Commission Européenne (European Commission, 2017), le site *www.re3data.org* recense plus de 2'600 répertoires de données (*Registry of Research Data Repositories*, 2021). Il permet d'effectuer des recherches selon divers critères, notamment le pays hôte et les domaines de recherche qui peuvent y soumettre leurs travaux. Nos données concernaient les sciences de la santé et les sciences humaines. De plus, comme mentionné précédemment, pour des raisons légales, elles devaient obligatoirement être enregistrées sur un serveur situé physiquement en Suisse. Les options se sont immédiatement réduites à onze possibilités (Tableau 1).

Tableau 1. Plateformes et adéquation par rapport à nos besoins.

Plateforme	Retenue	Arguments
FORS Data and Research Information Services (FORS DARIS)	Non	Equivalent à FORSbase
Openresearchdata, Platform Switzerland (ORD@CH)	Non	Plus en service
Data and Service Center for the Humanities (DaSCH)	Non	Adapté principalement aux données qualitatives et difficile d'accès
FORSbase (prochainement SWISSUbase)	**Oui**	
ETH Data Archive	Non	Pas d'accès public aux données et réservé à l'Ecole Polytechnique Fédérale de Zürich
Africa Centre for Population Health	Non	International et orienté vers la biologie
Communication Portal for Accessing Social Statistics (COMPASS)	Non	Plus en service
Health on the Net Foundation (HONmedia)	Non	Spécialisé dans les données audiovisuelles
Comparative Political Data Set (CPDS)	Non	Données politiques et institutionnelles
ETH Zürich Research Collection	Non	Réservé à l'Ecole Polytechnique Fédérale de Zürich
Sammlung Schweizerischer Rechtsquellen online (SSRQ-online)	Non	Adapté à l'Histoire et au Droit

Seule FORSbase répondait à nos besoins. A cette option, s'est ajoutée OLOS, une plateforme en phase finale de développement soutenue notamment par différentes universités suisses et la HES-SO. Elle n'était pas encore répertoriée dans le registre ci-dessus au moment de nos recherches. OLOS présentait, comme FORSbase, l'avantage d'être gérée et basée en Suisse, et a pour ambition de répondre au mieux aux besoins actuels des chercheurs.

Afin de les départager, nous avons rencontré des responsables de chacune de ces plateformes. Parallèlement, nous avons également suivi de nombreux webinaires portant sur la question du stockage et du partage des données.

À la suite de ces investigations et du développement de nos connaissances en la matière, nous avons défini un certain nombre de critères pour effectuer notre choix. Ils figurent, par ordre d'importance pour notre décision, dans le Tableau 2.

Tableau 2. Critères retenus et caractéristiques des plateformes de dépôt FORSbase et OLOS.

Critères	FORSbase	OLOS
Structure du répertoire	Adaptable	Unités organisationnelles (UO)
Prix	Gratuit	Au moins 67.50 CHF/année avec support technique
Gestion des données à long terme	Administrateur du projet de recherche	Administrateur de l'UO
Contrôle de l'accès aux données	Possible	En développement
Consultation des archives	Requiert une inscription	Libre
***Digital Objet Identifier* (DOI) attribué aux données**	OK	OK
Embargo	Possible	Possible
Recherche dans la base de données	Interface parfois peu claire	Aisée
Disponibilité	Immédiate bien qu'en transition vers SWISSUbase	Immédiate bien que pas lancée officiellement
Normes FAIR	OK	OK
Support technique	Oui	Oui

Note. Les informations listées dans ce tableau reflètent les caractéristiques des plateformes lorsque nous avons effectué nos investigations (de mai à août 2021) et sont sujettes à modification dans le futur.

Les critères les plus importants étaient la structure du répertoire de données et le coût. OLOS proposait un modèle payant dans lequel était créée ce qui est appelé « unité organisationnelle (UO) ». Une UO peut contenir des données déposées par une institution, une équipe de recherche ou même un·e chercheur·e seul·e. Elle est caractérisée par un espace et une durée de stockage définis à l'avance. Ces deux paramètres déterminent le tarif appliqué. L'espace minimal à louer était de 50 Giga-octets (Go), ce qui dépassait largement les quelques Go dont nous avions besoin pour notre projet. Ainsi, cette offre nous plaçait face au dilemme de devoir soit utiliser une portion congrue de l'espace pour lequel nous aurions payé ou de le partager avec d'autres équipes de recherche. La seconde option aurait nécessité une coordination au niveau institutionnel aussi bien concernant le financement du stockage que l'organisation des permissions d'accès à la plateforme pour déposer ou télécharger des données. Il en aurait donc résulté un investissement en temps très élevé pour une solution non optimale puisque notre équipe de recherche n'aurait finalement pas contrôlé complètement tout l'espace de stockage.

Au contraire, FORSbase nous permettait d'ouvrir un espace de stockage dédié spécifiquement à notre projet de recherche sans contrainte organisationnelle. La gratuité de cette plateforme, financée notamment par le Fonds national suisse pour la recherche scientifique et l'Université de Lausanne, représentait également un gain de temps substantiel. Nous n'avions pas besoin de lancer des démarches administratives pour trouver un financement supplémentaire. Enfin, même si la consultation et la recherche dans les archives nous ont semblées plus intuitives sur la plateforme OLOS, cela nous a paru un inconvénient mineur. En effet, FORSbase est actuellement en transition vers SWISSUbase et ce problème gagnerait à être corrigé.

C'est donc à partir de ces réflexions que nous avons choisi FORSbase plutôt que OLOS. Signalons tout de même qu'aucune des deux plateformes ne se démarquait clairement sur les autres points évalués. Elles exigeaient toutes deux un ou une administrateur·trice des données à llong terme. Elles permettaient toutes deux de gérer les accès aux données afin de contrôler que les demandeurs possèdent l'autorisation d'une commission d'éthique. Elles permettaient également les deux d'obtenir un DOI, de placer un embargo sur les données, de recevoir de l'assistance pour le formatage et le téléchargement des données et elles se conformaient aux normes FAIR que nous aborderons dans la section suivante.

Une fois notre choix effectué, nous avons contacté FORS et nous sommes entretenus avec leur personnel afin de planifier le dépôt des données.

5. Préparation et dépôt des données

Afin de rendre les données partagées aussi claires et accessibles que possible, Wilkinson et al. (2016) ont proposé l'utilisation des normes FAIR (*FAIR Principles*, 2021)[2]. Selon ces dernières, les données de recherches partagées doivent être trouvables (*Findable*), accessibles (*Accessible*), interopérables (*Interoperable*) et réutilisables (*Reusable*). Chacun de ces critères englobe plusieurs caractéristiques explicitées ci-après. Nous avons, pour chacune, décrit comment nous nous y sommes conformés.

5.1 Principes FAIR

5.1.1 Findable

Caractéristiques attendues	Réponses
Les données doivent se voir attribuer un identifiant[3] unique et persistant.	FORSbase remplit cet objectif en attribuant un DOI aux données.
Les données doivent être accompagnées de métadonnées complètes. Ces dernières ont pour but de décrire les données de recherche de manière aussi détaillée que possible.	Un fichier contenant les métadonnées requises a été créé. Il comporte la description détaillée de la section méthode des études HEalStud et HEalS-Nu, des fichiers de données, de leur organisation et de leur traitement.
Les métadonnées doivent inclure l'identifiant des données (DOI) clairement et explicitement.	Effectué
Les données sont indexées sur une plateforme qu'il est possible de trouver en utilisant les moteurs de recherche courants.	FORSbase apparaît dans les résultats des principaux moteurs de recherche sur internet (tests effectués avec Google, Bing, etc…)

[2] Site internet de l'initiative GO FAIR, composée d'individus issus des mondes politique et scientifique, qui promeut l'application des principes FAIR.

[3] Nous avons privilégié l'utilisation du terme identifiant pour le terme anglais *identifier* dans ce rapport. A noter que FORSbase lui préfère le terme identificateur.

5.1.2 Accessible

Caractéristiques attendues	Réponses
Les données doivent pouvoir être retrouvées grâce à leur identifiant en utilisant un protocole de communication standardisé, tels par exemple, http ou ftp. Autrement dit, l'accès aux données ne doit pas reposer sur des protocoles propriétaires qui pourraient ne pas être accessibles à tout le monde.	FORSbase utilise des protocoles https compatibles avec ce critère.
Le protocole de communication standardisé doit être gratuit et *open-source* afin que n'importe qui disposant d'un ordinateur et d'une connexion internet puisse accéder au moins aux métadonnées.	C'est le cas.
Le protocole de communication standardisé permet d'authentifier et autoriser l'accès aux données si nécessaire.	FORSbase offre cette possibilité en exigeant une inscription pour accéder aux données et en donnant la possibilité aux chercheur·e·s qui ont déposé les données d'en contrôler les accès.
Les métadonnées doivent être accessibles même si les données ne le sont plus.	FORSbase ne fixe pas de durée de stockage déterminée. Les métadonnées restent disponibles en tout temps.

5.1.3 Interoperable

Caractéristiques attendues	Réponses
Les données doivent être transmises dans un langage formel, accessible, partagé et généralisé. Elles doivent pouvoir être lues par un ordinateur sans recourir à un algorithme ou un traducteur.	Les données ont été enregistrées au format .csv lisible pour n'importe quel ordinateur.
Les données utilisent un vocabulaire respectant les principes FAIR.	Les données et métadonnées ont fait l'objet d'un travail très conséquent de mise en forme et de description.
Les données doivent inclure des références qualifiées à d'autres données, par exemple en expliquant comme différents jeux de données sont liés entre eux.	Effectué.

5.1.4 Reusable

Caractéristiques attendues	Réponses
Les données doivent être décrites par de nombreux attributs pertinents sans essayer d'anticiper l'identité ou les besoins des utilisateurs secondaires.	L'intégralité des données à notre disposition a été partagée à l'exception de celles permettant d'identifier les participant·e·s.
Les données doivent être diffusées avec une licence d'utilisation claire et accessible.	La diffusion des données se fera sous une licence Creative Commons qui autorise leur réutilisation (sous condition que la source originale soit citée), mais pas leur redistribution à des tiers.
La provenance exacte des données doit être spécifiée. Leur origine et leur histoire doivent être spécifiée.	Le projet de recherche et son bras secondaire ont été relatés précisément, de leur conception aux publications auxquelles ils donneront lieu.
Les données doivent correspondre aux standards de la communauté du domaine sur lequel elles portent.	Les données ont été traitées conformément aux usages requis dans les sciences de la santé et sciences humaines et en respectent les standards de présentation habituels.

5.2 Données quantitatives

Concrètement, l'adaptation de nos données aux principes FAIR s'est déroulée en plusieurs étapes.

5.2.1 Nettoyage des données

En premier lieu, le gestionnaire des données a procédé à une nouvelle extraction des données brutes et les a anonymisées. Elles ont ensuite été nettoyées afin d'en faire disparaître tout élément permettent d'identifier les participant·e·s. Bien que le gestionnaire de données ait supprimé les identifiants directs, tels que le nom et l'adresse de certain·e·s participant·e·s, il restait possible d'en identifier certain·e·s en croisant certaines informations, en particulier lorsque des valeurs extrêmes étaient impliquées. Par exemple, si une participante avait répondu avoir 65 ans et étudier le Design, il était probable qu'elle soit la seule étudiante de cet âge dans ce domaine : ses caractéristiques auraient permis de l'identifier. Par conséquent, les variables continues comme l'âge ont été recodées en catégories, afin d'éviter que certaines valeurs ne soient uniques. Par exemple, si trois participant·e·s ont respectivement 42, 56 et 65 ans[4], ils sont maintenant tou·te·s classé·e·s dans la catégorie d'âge « 35 ans et plus », rendant impossible d'isoler l'un·e ou l'autre sur la base de son âge.

Un certain nombre de variables comportant la même valeur pour tou·te·s les participant·e·s (par ex : indication de l'enregistrement à la fin du questionnaire) ont été supprimées. Les scores et moyennes de chacune de nos variables numériques ont été calculées selon les instructions fournies par les personnes ayant conçus les questionnaires utilisés.

[4] Ces valeurs sont fictives.

Enfin, nous avons nettoyé et rendu le plus accessible possible le script du programme d'analyse statistique R utilisé afin d'effectuer toutes les opérations listées ci-dessus. Partager ce script permettra à quiconque utilisera ces données, de contrôler l'exactitude de nos démarches.

5.2.2 Préparation des métadonnées

L'étape suivante était la création des fichiers de métadonnées. Il s'agit de fichiers permettant de comprendre le contenu des jeux de données décrits au point précédent. Les métadonnées contiennent une description détaillée des deux projets de recherche. Le but du projet HEalStud et du bras HEalS-Nu QUANTI, sont décrits, de même que leur contexte, les méthodes et procédures suivies ainsi que les résultats et les interprétations que nous en avons tiré.

Les métadonnées incluent également la liste détaillée des manipulations effectuées au cours du nettoyage des données, la liste de toutes les variables présentes dans les jeux de données, ainsi que les questions auxquelles elles correspondent dans le questionnaire rempli par les participant·e·s. Relevons ici un point digne d'attention : certains questionnaires sont soumis à un droit d'auteur. C'était le cas de l'une des échelles du questionnaire, la *Connor-Davidson Resilience Scale* (CD-RISC ; Connor & Davidson, 2003). Les items la constituant n'ont ainsi pas pu être partagés dans le fichier de métadonnées. Nous avons simplement indiqué l'URL du site permettant de contacter les auteurs de cette échelle en lieu et place.

Afin de rendre les données accessibles pour le plus grand nombre, ces métadonnées ont été rédigées en anglais.

5.3 Données qualitatives

Les futurs utilisateurs des données qualitatives (HEalS-Nu quali) doivent également disposer du maximum d'informations pour utiliser ces données de manière pertinente lors d'analyses secondaires. La préparation du matériel est donc également essentielle. Nous avons suivi les recommandations de FORSbase guidant la préparation des données qualitatives pour leur dépôt.

Cette préparation débute déjà lors de la soumission à une commission d'éthique dans le cas où, comme ici, des données de santé sont collectées. Cette instance doit en effet pouvoir se prononcer sur la manière dont la confidentialité et la sécurité des données seront assurées. De plus, elle exige que les participant·e·s potentiel·le·s soient informé·e·s pour qu'ils ou elles puissent donner leur accord formel préalable à la réutilisation de leurs données par d'autres équipes de recherche, et ce moyennant le respect de conditions qui doivent leur être explicitées.

5.3.1 Nettoyage des données

Comme les entretiens en groupe (*focus-groups*) avaient eu lieu en français, tout le matériel partagé est mis à disposition dans cette langue.

Les données qualitatives ont été anonymisées et nettoyées en ôtant tout identifiant direct et indirect (noms des participant·e·s, de personnes, d'hôpitaux, d'institutions ou de lieux spécifiques) permettant de les relier avec un parcours de vie personnel ou professionnel. Nous les avons remplacés par « XXX ».

La transcription des entretiens a été effectuée en format Word. Chacune de ces transcriptions s'est vu attribuer un identifiant unique, une en-tête comprenant la date, le lieu, le nom de la personne qui a réalisé l'interview et des informations socio-démographiques de base sur les participant·e·s. La mise en page est uniforme. Une distinction est faite entre les différent·e·s locuteurs·trices ainsi qu'entre les séquences de questions-réponses. Et finalement, les pages sont numérotées.

5.3.2 Préparation des métadonnées

Les notes, rapports et tout matériel comprenant des informations en lien avec la méthodologie de recherche et l'utilisation des données sont partagées dans le fichier de métadonnées. En particulier, des informations sont fournies concernant le projet (nature, temporalité, situation géographique et objectifs) et la méthode suivie (devis, population, échantillonnage, cadre). Des renseignements précis sont aussi proposés à propos de la manière dont les données ont été collectées (courriels d'information et d'invitation à participer aux *focus-groups* ; formulaire de consentement, consignes aux intervieweuses ; guide d'entretien ; *mindmaps* construites avec les participant·e·s). Enfin, nous avons mis à disposition des indications sur la structure des fichiers de données, les relations entre les fichiers (soit ici entre les transcriptions des *focus-groups* en format Word et les *mindmaps* en format JPEG) ainsi que la description des mesures prises pour anonymiser ces transcriptions et assurer la confidentialité des données.

6. Dépôt sur FORSbase

Dans un premier temps, nous avons procédé à l'inscription du projet sur la plateforme FORSbase et déposé le descriptif des deux études liées en français et en anglais. Au moment de la reddition du présent rapport (octobre 2021), nous sommes en voie de finaliser la préparation des fichiers de données et de métadonnées pour les parties quantitatives et qualitatives (délai à fin 2021). Cette étape est particulièrement astreignante et coûteuse en temps. Elle nécessite une très bonne systématique de travail pour éviter tout oubli. Lorsque les fichiers de données et métadonnées seront prêts, ils seront téléchargés sur FORSbase. Le personnel scientifique dédié les inspectera afin de s'assurer que les contenus correspondent aux normes requises. Une fois les articles scientifiques et professionnels acceptés pour publication, et les communications orales approuvées pour présentation, nous lèverons l'embargo. Cela permettra ainsi à toute équipe de recherche d'accéder à ces données sous réserve, bien-sûr, de l'approbation du protocole par une commission d'éthique.

Il restera enfin à nommer un ou une gestionnaire du projet, soit une personne issue de l'équipe de recherche.

7. Conclusion

Les évolutions technologiques et la digitalisation ont amené de nombreux bénéfices à la recherche, en particulier ce qui concerne la collecte, la gestion et les possibilités de partage simple et rapide des données. Mais comment maximiser leur portée et leur visibilité tout en protégeant ces données, en particulier celles de santé considérées comme sensibles ?

Les questions soulevées vont bien au-delà de ce que nous avions imaginé avant de débuter le processus de partage de nos données. Le choix de la plateforme où les déposer est stratégique, il peut être contingent de restrictions ou de particularismes liés à la nature des données. Les propriétés et les services proposés par ces plateformes doivent en outre correspondre à la nature spécifique du projet, au contexte institutionnel et légal.

Le travail généré par la mise à disposition des données est conséquent, exige une attention soutenue et la capacité de les présenter de la manière la plus évidente et détaillée possible. Une grande rigueur et une connaissance poussée des enjeux entourant ce partage sont donc requises. Aussi, puisque la diffusion de données est désormais incontournable, il importe de penser aux moyens nécessaires pour garantir des pratiques de qualité. A l'avenir, outre des moyens financiers et des formations en la matière, les équipes de recherche auraient tout à gagner en s'adjoignant des collaborateurs·trices spécialisé·e·s dans la gestion, la préservation,

le partage et la valorisation des données. Cela nous paraît d'une importance cruciale pour que le partage des données de recherche, l'*Open Data*, soit en mesure de contribuer à la poursuite durable et transparente du développement des connaissances.

8. Remerciements

Nous tenons à remercier vivement M. André Jelicic, directeur de la plateforme OLOS qui nous a aimablement renseigné sur les possibilités qu'elle pouvait nous offrir. Notre gratitude va également à Mme Eliane Ferrez et MM. François Loretan et Nicolas Fedrigo, expert·e·s scientifiques, qui nous ont renseigné sur la plateforme FORSbase et accompagnent désormais notre dépôt. Enfin, merci à Mme Rym Vivien, chargée de projet Open Data A.I. au dicastère Recherche et innovation de la HES-SO qui a répondu à nos questions concernant le choix de la plateforme.

9. Références

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*(2), 76–82. https://doi.org/10.1002/da.10113

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

European Commission. (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

FAIR Principles. (2021). GO FAIR. https://www.go-fair.org/fair-principles/

Goben, A., & Sandusky, R. J. (2020). Open data repositories: Current risks and opportunities. *College & Research Libraries News*, *81*(2), 62. https://doi.org/10.5860/crln.81.1.62

International Science Council. (2017). *"Open Data in a Big Data World" accord passes 120 endorsements*. Council Science. https://council.science/current/news/open-data-in-a-big-data-world-accord-passes-120-endorsements/

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(6), 0160. https://doi.org/10.1038/s41559-017-0160

Nagy Hesse-Biber, S. (2011). *The handbook of emergent technologies in social research* (Oxford University Press).

Pang, P. C.-I., Chang, S., Verspoor, K., & Clavisi, O. (2018). The Use of Web-Based Technologies in Health Research Participation: Qualitative Study of Consumer and Researcher Experiences. *Journal of Medical Internet Research*, *20*(10), e12094. https://doi.org/10.2196/12094

Piwowar, H. A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE*, *6*(7), e18657. https://doi.org/10.1371/journal.pone.0018657

Registry of Research Data Repositories. (2021). Re3data.Org. https://doi.org/10.17616/R3D

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014). Introduction to Statistical Disclosure Control (SDC). *IHSN Working Paper*, *7*, 25.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18