Haute école de travail social
et de la santé Lausanne

Projet n°81829

Mise en open data des données de l'enquête auprès des enfants, des jeunes et des acteurs jeunesse

Rapport scientifique à la HES-SO

ÉQUIPE DE RECHERCHE

Dr. Pierre Benz (HETSL | Lausanne | HES-SO)

Dr. Romaric Thiévent (HETSL | Lausanne | HES-SO)

29 OCTOBRE 2021

Haute école de travail social et de la santé Lausanne
Ch. des Abeilles 14 • 1010 Lausanne • Suisse
+41 (0)21 651 62 00 • www.hetsl.ch

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

Haute école de travail social
et de la santé Lausanne

Table des matières

Table des illustrations

1 INTRODUCTION

La Loi sur l'encouragement de l'enfance et de la jeunesse (LEEJ) permet aux cantons de bénéficier de soutiens financiers de la Confédération pour mettre sur pied des programmes qui visent la constitution ou le développement de leur politique de l'enfance et de la jeunesse. Le canton du Jura a saisi cette opportunité et initié un projet de réflexion et d'expérimentation baptisé *Jura Jeunes 4.0*[1]. Construit autour des thématiques de la « protection », de « l'encouragement » et de la « participation », ce projet vise l'amélioration des mesures de soutien en faveur des enfants et des jeunes, ainsi qu'une augmentation de la réactivité du réseau de professionnel·le·s et d'acteurs jeunesse pour répondre aux besoins de cette population.

S'étendant sur la période 2019-2021, *Jura Jeunes 4.0* comprend deux étapes successives : 1) d'abord dresser un état des lieux des problèmes et des besoins des enfants et des jeunes en s'appuyant sur des études quantitatives et qualitatives, pour 2) ensuite associer les acteurs jeunesse, tels que les clubs de sports, les groupements de jeunesse ou les espaces de loisirs, et les organisations d'aide à la jeunesse à l'élaboration de projets concrets.

Dans le cadre du premier volet du projet visant à disposer d'éléments empiriquement fondés sur lesquels baser le développement de mesures concrètes, le Service de l'action sociale du Canton du Jura a mandaté la Haute école de travail social et de la santé Lausanne et la Haute école de gestion Arc (HES-SO) pour la réalisation d'une enquête à large échelle auprès des enfants, des jeunes et des acteurs jeunesse sur l'ensemble du territoire cantonal. Menée d'avril 2019 à mai 2020, cette recherche par questionnaires avait pour objectifs d'identifier d'une part les besoins des jeunes jurassien·ne·s de 12 à 25 ans en matière de protection, de participation et d'encouragement et de déterminer d'autre part la manière avec laquelle les acteurs jeunesse (organisations de jeunesse et associations socioculturelles et sportives) détectent les besoins des jeunes, répondent aux problématiques identifiées et conçoivent leur rôle face à ces besoins[2].

La fin du mandat (mai 2020) a coïncidé avec l'ouverture de l'appel à projets *Open Data HES-SO* le 1er juin 2020. L'équipe de recherche, qui n'avait pas prévu de partager les données de l'enquête, a saisi la possibilité de financement offerte par la HES-SO pour se lancer pour la première fois dans le processus de partage de données quantitatives. Le présent rapport retrace la démarche et les activités déployées dans ce cadre. Il est structuré en 5 chapitres. Le chapitre 2 présente les échantillons des trois enquêtes (enfants, jeunes et acteurs jeunesse) et détaille les données exploitées dans le cadre du mandat. Le chapitre 3 propose un bref état de l'art des *Data repositories* existants dans le domaine des sciences humaines et sociales et expose les raisons du choix de FORSbase pour le dépôt des données de l'enquête. Le quatrième chapitre présente la méthode d'implémentation des principes FAIR

[1] https://www.jura.ch/DIN/SAS/Jeunesse/Jura-Jeunes-40.html

[2] Les résultats détaillés sont consultables dans le rapport final (Tironi et al. 2020). Une sélection a été publiée dans une revue professionnelle (Heim et al. 2021).

lors du dépôt et propose une synthèse des tâches effectuées en y associant les compétences mobilisées et le temps approximatif. Le chapitre 5 décrit les problématiques éthiques auxquelles l'équipe de recherche a été confrontée au cours du processus de partage des données et les solutions mises en œuvre pour y répondre. D'une portée plus générale, la conclusion revient tout d'abord sur les quelques éléments saillants qui ressortent de notre première expérience de mise en open data de donnée et dresse ensuite un rapide état des lieux de quelques contributions scientifiques significatives portant sur l'étude des pratiques de l'open science.

2 DESCRIPTION DES ECHANTILLONS ET DES DONNEES

La recherche est basée sur trois enquêtes par questionnaire : l'enquête auprès des enfants, l'enquête auprès des jeunes et l'enquête auprès des acteurs jeunesse. Dans cette partie, nous exposons brièvement la construction des trois échantillons, puis détaillons les données exploitées dans le cadre du mandat.

Enquête auprès des enfants (12-18 ans)

Le questionnaire destiné aux enfants a été administré au début de l'année scolaire 2019-2020 au sein d'écoles du secondaire 1 et 2 (CEJEF). Un échantillon stratifié a été constitué sur la base des statistiques scolaires 2018-2019, permettant de choisir un certain nombre d'établissements et de classes. En accord avec le canton du Jura, mandant de l'étude, il a été décidé d'interroger environ 900 élèves parmi les 5'451 que comptent les niveaux secondaires 1 et 2 pour l'année scolaire 2018-2019, soit 75 classes. L'échantillon compte finalement 781 enfants ayant rempli le questionnaire entre le 23 septembre et le 4 octobre 2019, durant leurs leçons et sous la supervision de leur enseignant·e ou d'un chargé de projet du Service de l'action sociale du Canton du Jura.

Enquête auprès des jeunes (19-24 ans)

Une procédure d'échantillonnage différente a été mise en œuvre dans le cas des jeunes. Sur la base du registre de la population du Canton du Jura, un échantillon aléatoire de 1'300 jeunes a été prélevé parmi les personnes nées entre le 30 septembre 1994 et le 9 septembre 2000 (soit 23.8% des personnes concernées). Les 1'300 individus sélectionnés ont reçu une lettre d'invitation à participer à l'enquête leur indiquant un lien d'accès au questionnaire en ligne. La récolte de données s'est déroulée entre le 19 septembre et le 21 octobre 2019. Finalement, 361 jeunes ont participé à l'enquête et rempli des questionnaires complets et exploitables, ce qui représente un taux de réponse de 28%. La participation n'ayant pas été égale dans toutes les catégories de jeunes, les résultats ont été pondérés en fonction de l'âge et du sexe.

Enquête auprès des acteurs jeunesse

La population cible de l'enquête auprès des acteurs jeunesse a été définie en référence à la Loi sur la politique de la jeunesse (art. 4 let. c et 15) et comprend l'ensemble des organisations de jeunesse et des associations socioculturelles et sportives qui encadrent par leur action les jeunes de 12 à 24 ans dans le canton du Jura. Les acteurs inclus partagent de manière cumulative les caractéristiques suivantes :

1) L'encadrement des enfants et des jeunes est principalement assuré par des bénévoles sur un principe de type « milice ».
2) L'accueil des enfants et des jeunes se fait de manière universaliste, c'est-à-dire qu'il s'adresse aux enfants et aux jeunes sans distinction.
3) L'accueil des enfants et des jeunes se fait sur une base régulière, c'est-à-dire que les activités proposées ont lieu sur une base hebdomadaire ou au moins plurimensuelle.
4) L'accueil des enfants et des jeunes se fait de manière collective, c'est-à-dire que les activités proposées sont, en principe, effectuées en groupe.
5) L'accueil des enfants et des jeunes repose sur une base volontaire, c'est-à-dire que les enfants et les jeunes sont libres de participer aux activités proposées.

Au total, 370 acteurs jeunesse ont été recensés comme répondant à ces critères. Ils se répartissent en quatre grands domaines d'activité : le sport (clubs et associations sportives, y compris les écoles de danse), la culture (écoles de musique, fanfares, théâtre), l'accueil et l'animation (groupes jeunesse, scouts, centres et espaces de jeunesse) et la politique (conseil des jeunes de la ville de Delémont et sections jeunesse des partis politiques). Au sein des organisations éligibles, l'enquête a ciblé les président·e·s ainsi que toute personne, bénévole ou salariée, occupant des fonctions d'animation et d'encadrement auprès des jeunes de 12 à 24 ans[3].

Les acteurs jeunesse ont reçu un e-mail les invitant à remplir un questionnaire en ligne. L'enquête a été ouverte entre le 17 septembre et le 6 novembre 2019. Au total, 186 acteurs jeunesse sur les 370 contactés y ont participé, soit un taux de retour de 50.3%. En tout, 297 questionnaires ont été complétés par des membres des 186 acteurs jeunesse répondants.

Description des données

Les réponses aux trois questionnaires ont été traduites en trois jeux de données au format .sav compatible avec le logiciel de traitement statistique SPSS. L'opération suit une procédure standardisée, la conversion et l'export des données étant directement intégrés au logiciel Limesurvey[4]. Les données textuelles brutes ont ensuite été recodées en modalités chiffrées suivant un usage très répandu dans l'utilisation du logiciel SPSS. Cette démarche requiert l'établissement d'un dictionnaire des variables ainsi que d'un *codebook* visant à préciser les modalités [0 = non, 1 = oui, p. ex.]. Ces documents étant précisément ceux qui sont exigés lors d'un dépôt en ligne des données notamment sur FORSbase, il apparaît clairement que la méthodologie utilisée pour les trois enquêtes ainsi que le type de données quantitatives produites répondent très bien aux standards de *l'Open data*. On peut donc relativement facilement les formater dans cette perspective.

De plus, à l'exception d'une minorité de questions portant sur les problèmes rencontrés par les enfants et les jeunes au cours de l'année écoulée (harcèlement physique, cyberharcèlement, problème d'argent, consommation excessive d'alcool ou de drogue, par exemple), les données des trois enquêtes ne sont pas particulièrement sensibles et l'on peut garantir une anonymisation optimale des trois jeux de données. Nous reviendrons plus en

[3] Le nombre de personnes, bénévoles ou salariées, actives au sein de ces 370 organisations est toutefois inconnu.

[4] https://manual.limesurvey.org/Exporting_results

détail sur ces enjeux dans la partie 6 du présent rapport. Le Tableau **1** ci-dessous présente le nombre de questions, de variables et de cas pour les trois enquêtes.

Tableau 1 : Récapitulatif des caractéristiques principales des données des trois enquêtes

	Nombre de questions	Nombre de variables	Nombre de cas
Enquête auprès des enfants	16	111	781
Enquête auprès des jeunes	19	118	361
Enquête auprès des acteurs jeunesse	18	56	297

Les tables de données comportent quatre types de variables. Des variables binaires (1) sont utilisées (a) lorsque les modalités de réponse sont oui ou non, ou (b) lorsque plusieurs réponses sont possibles. Dans ce cas, chaque modalité constitue une variable binaire. Des variables de type échelle (2) [tout à fait d'accord, plutôt d'accord, etc.] sont utilisées pour les questions renvoyant à des évaluations subjectives ou à des fréquences [tous les jours, plusieurs fois par semaine, plusieurs fois par mois, jamais]. Les âges et années sont des variables numériques (3) et les réponses aux questions ouvertes ont été conservées dans le format orignal [texte] (4).

Les enquêtes auprès des enfants et des jeunes sont relativement similaires. Les questionnaires présentent la même architecture autour de trois thèmes (temps libre et activités de loisir, participation et consultation, protection et prévention). La plupart des questions sont identiques, bien que certaines modalités aient dû être adaptées lorsqu'elles concernaient uniquement les enfants (12-18 ans) ou les jeunes (19-24 ans). Le questionnaire auprès des jeunes comporte quelques questions supplémentaires concernant le niveau de formation, la situation familiale et le revenu mensuel net.

L'enquête auprès des acteurs jeunesse est en revanche tout à fait différente des deux autres, bien que les questions posées entrent directement en résonnance avec les questionnaires des enfants et des jeunes. Les questions s'articulent autour de trois thèmes : l'expérience en matière de détection des besoins des jeunes, les compétences en matière de détection des besoins des jeunes, et la connaissance des institutions et des associations d'aide à la jeunesse.

3 CHOIX DU DATA REPOSITORY

Notre choix s'est porté sur FORSbase[5], une plateforme suisse et disciplinaire spécialisée dans les sciences sociales[6]. La plateforme offre une remarquable accessibilité, renforcée par le dispositif d'encadrement et d'accompagnement *ad personam* déployé par FORS qui facilite grandement la compréhension des démarches à effectuer en vue d'un dépôt. Dans le cas du présent mandat, nous avons directement contacté une personne chez FORS que nous connaissions par ailleurs, et avons été dirigés vers une archiviste spécialiste avec qui nous

[5] https://forsbase.unil.ch/

[6] Les données et les métadonnées de l'enquête auprès des enfants, des jeunes et des acteurs jeunesse sont disponibles sur FORSbase à l'adresse suivante : https://forsbase.unil.ch/project/study-public-overview/17579/0/

avons pu discuter de nombreux enjeux liés au dépôt des données en ligne sur le plan non seulement administratif (procédures, guides, jeux de données), mais également technique (traitement des données manquantes, techniques d'anonymisation, traitement des recodages, notamment).

Un dépôt sur FORSbase présente de nombreux avantages : le centre est spécialisé dans l'accompagnement du dépôt de données et connaît les enjeux liés au partage des données en sciences humaines et sociales. Il existe par ailleurs de nombreux guides consacrés à ces enjeux établis par les spécialistes de FORS[7]. De plus, la plateforme offre un excellent degré de protection (enregistrement préalable pour avoir accès aux données, signatures de contrats d'utilisation, information transmise aux propriétaires des données lorsque l'accès à celles-ci est demandé), et un éventail de choix quant à la limitation de l'accès à un public spécifique. Sur la question des choix et stratégies institutionnelles en matière de dépôt de données, Guirlet (2020) évoque de nombreux critères de qualité des dépôts et identifie les informations et les paramètres importants pour le choix d'un dépôt. Y figure FORSbase, de même que les dépôts OLOS, DaSCH, GitHub et Zenodo évoqués ci-après.

Il existe en effet de nombreux autres *repositories*. Le Fonds national suisse de la recherche scientifique (FNS) en dénombre 146, dont 24 répondent à la double exigence de remplir les critères *Open Research Data* (ORD) du FNS et d'avoir été mentionnés au moins 10 fois dans un rapport de monitoring du FNS sur *l'Open Research Data*, publié en février 2020[8]. Parmi elles figurent notamment GitHub et Zenodo[9]. L'utilisation de Github est très répandue parmi les utilisateurs des logiciels R et Python, qui entretiennent une forme de culture du partage de scripts de codage via la plateforme. Il s'agit d'un dépôt généraliste, orienté vers le partage de données et le développement de code. Tout comme GitHub, Zenodo est un dépôt généraliste, c'est-à-dire non orienté disciplinairement. En revanche, il diffère dans son organisation. Alors que GitHub fonctionne comme une communauté de développeurs de logiciels open source, Zenodo est un projet institutionnel développé et hébergé par le CERN dans le cadre du programme Horizon 2020. Parmi les dépôts généralistes, il faut aussi mentionner Olos, un nouvel arrivant lancé cette année par l'Université de Genève et la HES-SO dans le cadre du projet DLCM avec le soutien de Swissuniversities, en phase d'accréditation par le FNS[10]. À côté des data repositories généralistes, le FNS recense 13 dépôts disciplinaires dont 10 en sciences de la vie, un pour les humanités (DaSCH), et deux pour les sciences sociales (FORSbase et Harvard Dataverse)[11]. Des développements sont à suivre avec la création de SWISSUbase qui se veut un *« outil national central pour la conservation, la préservation et la diffusion des données de la recherche scientifique, ainsi qu'un outil d'information sur les projets de recherche en cours et terminés dans le pays »*[12].

[7] https://forscenter.ch/publications/fors-guides/?lang=fr

[8] https://www.snf.ch/fr/WtezJ6qxuTRnSYgF/dossier/points-de-vue-politique-de-recherche

[9] https://github.com/ ; https://zenodo.org/

[10] https://olos.swiss/ ; https://www.dlcm.ch/

[11] https://dasch.swiss/ ; https://dataverse.harvard.edu/

[12] https://info.swissubase.ch/fr/about-the-project/

4 METHODE D'IMPLEMENTATION DES PRINCIPES FAIR

Le temps et les compétences nécessaires à la préparation des données en vue d'un dépôt en libre accès sont extrêmement variables en fonction notamment d'aspects techniques (type de données, complexité de l'enquête, qualité du matériau, volume de données), mais aussi d'aspects liés à la quantité de métadonnées à produire. La présente enquête est un exemple de cas plutôt simple. Il s'agit de données quantitatives issues de questionnaires sans notable complexité, d'une méthode d'implémentation de transcription des modalités de réponses dans un format reconnu et largement utilisé par la communauté des chercheur·e·s (SPSS), en suivant une procédure connue (Limesurvey). Malgré ce contexte idéal – c'est à ce type d'enquête et de données que font en premier lieu référence les standards techniques de la mise en libre accès – ce processus a demandé des compétences spécifiques pour un temps relativement important. Il faut au moins deux semaines pour effectuer l'entier du travail et garantir l'intégrité de l'ensemble des (méta)données. La fin de cette partie présente un tableau récapitulatif des tâches et du temps approximatif dédié à chacune d'elles, en intégrant non seulement le travail sur les (méta)données, mais aussi l'ensemble du processus qui aboutit à la mise en ligne des données de l'enquête.

Dans ce qui suit, nous reprenons point par point les principes FAIR (trouvables, accessibles, interopérable, réutilisable)[13], décrivons les tâches effectuées correspondantes et évaluons le temps de travail et les compétences utilisées. Nous utilisons les sous-principes [F1, F2, A1, etc.] tels qu'ils apparaissent dans les « Principes directeurs pour la publication de données trouvables, accessibles, interopérables et réutilisables version b1[14] », voir également Wilkinson et al. (2016).

Trouvables (findable)

La première étape de la (ré)utilisation des données consiste à les trouver. Les métadonnées et les données doivent être faciles à trouver, tant par les humains que les ordinateurs. Les métadonnées lisibles par machine sont essentielles pour la découverte automatique des ensembles de données et des services, c'est donc un élément essentiel du processus de FAIRification[15].

Les données des trois enquêtes possèdent chacune un identifiant de citation et un DOI créés automatiquement par FORSbase (F1). Les données ont été décrites dans deux documents : un dictionnaire des variables et un codebook. Les métadonnées (F2) comprennent aussi les informations destinées aux répondant·e·s, renseignant ainsi sur les conditions de la passation des questionnaires, ainsi que les questionnaires eux-mêmes. Ces métadonnées sont enregistrées et indexées via leur publication dans FORSbase (F3), et elles spécifient les enquêtes auxquelles elles sont liées (F4).

La tâche ayant pris le plus de temps et nécessité des compétences particulières est l'établissement des métadonnées, spécifiquement les dictionnaires des variables et les codebook. Parce qu'elles sont directement dépendantes de la mise en forme des tables de

[13] https://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

[14] https://www.force11.org/fairprinciples

[15] https://www.go-fair.org/fair-principles/

données, elles ont été produites en parallèle et il est donc très difficile d'évaluer avec précision la répartition des deux semaines entre le traitement des données et le travail sur les métadonnées. Pour donner un ordre de grandeur, nous l'évaluons à deux tiers du temps pour les données et un tiers pour les métadonnées. Des mouvements d'aller-retour ont dû être effectués, d'abord lors de la préparation des tables, et surtout lors des contrôles finaux. Pour garantir l'intégrité des données, nous avons créé un script qui sélectionne, ordonne et recode les variables à partir du fichier de données source. L'écriture d'une telle syntaxe permet (a) de reproduire à l'infini l'ensemble des opérations, (b) contrôler chaque étape du processus, (c) générer les tables finales destinées à l'archivage en ligne sans intervention « humaine » dans celles-ci. Pour des raisons de compétences, l'entier des commandes a été codé en R, et non en SPSS.

On peut relever ici deux remarques importantes. Premièrement, les modalités des variables de type *factor* en R ne sont pas reconnues en SPSS par leurs labels, mais par leur position. Par exemple, les modalités [0 = non, 1= oui, 99 = manquant] en R seront interprétées de la manière suivante en SPSS : [0 = non, 1 = oui, 2 = manquant], soit [facteur 1, facteur 2, facteur 3]. Un traitement des données en R doit donc exclure l'utilisation de variables *factor* et leur préférer de simples caractères. Une seconde remarque technique est que l'import des labels (« valeurs » en SPSS) ne peut être fait si l'un d'eux dépasse la limite de caractères autorisée par le programme.

Une fois que l'entier des variables et des modalités ont été préparées, et les trois tables de données formées, un dictionnaire des variables a été établi. Celui-ci relie les questions des questionnaires aux variables, et indique le nom des variables, le type et les libellés. Ensuite, les codebooks mentionnent le nom des variables et leurs modalités (chiffre et label). Pour les deux types de documents, les informations nécessaires ont été exportées de R vers Excel, puis intégrées dans des documents Word.

Accessibles (accessible)

Une fois que l'utilisateur a trouvé les données requises, il doit savoir comment y accéder, ce qui peut inclure l'authentification et l'autorisation.

L'ensemble des exigences relatives à l'accessibilité des données engage la compétence de FORSbase. La traçabilité et l'accès des (méta)données via leur identifiant (A1), le protocole de communication concernant l'accès (A2) et/ou les autorisations d'accès (A3) sont très clairement indiqués au long du processus de mise en dépôt des données et des métadonnées.

Les conditions d'accès aux données déposées sur FORSbase sont définies dans le cadre d'un contrat d'utilisateur[16]. Celui-ci stipule que toute personne ou organisation peut y accéder pour autant qu'elle respecte certaines conditions, notamment l'utilisation des données uniquement à des fins de recherche scientifique ou d'enseignement académique sans possibilité d'usage à des fins commerciales, le respect des lois et normes en matière de protection des données ou encore l'obligation d'informer FORS en cas de publication utilisant les données. L'équipe de recherche a été particulièrement sensible à cette politique d'accessibilité potentielle des données à toutes et tous complétée par des règles strictes quant à leur utilisation.

[16] https://forsbase.unil.ch/media/general_documentation/fr/User_contract_F.pdf

Sur FORSbase, l'accès aux données peut être restreint à certaines utilisations (et utilisateurs·trices) selon trois niveaux de restriction d'accès, au choix : (1) *Aucune restriction*, (2) *Recherche et enseignement académiques*, (3) *Recherche académique uniquement*. Les modalités sont de plus en plus restrictives. *Aucune restriction* signifie que les données sont disponibles pour tous les types d'usages non exclus explicitement par le contrat utilisateur. Elles peuvent ainsi être utilisées dans le cadre de recherches menées par des groupes d'intérêt (partis politiques, associations, syndicats), d'articles journalistiques ou par des instituts privés effectuant des mandats d'intérêt public. La modalité *Recherche et enseignement académique* exclut les usages précités pour les réserver à la recherche et l'enseignement académique. Dans ce cas, seules les personnes rattachées à une haute école peuvent avoir accès aux données. *Recherche académique uniquement* est la modalité la plus restrictive et n'autorise l'usage des données que dans le cadre de la recherche académique. Il convient de rappeler que dans tous les cas, seul·e·s les utilisateurs et utilisatrices du système ayant préalablement créé un compte sur FORSbase et signé le contrat d'utilisation ont accès aux données.

Afin de garantir l'accès le plus large possible aux données, nous avons choisi de n'appliquer aucune restriction. La recherche ayant été financée par le Canton du Jura, il aurait été à notre sens paradoxal que les administrations publiques d'autres cantons ou des communes jurassiennes ne soient pas autorisées à accéder aux données.

L'institution qui dépose les données peut exercer en outre exercer un contrôle sur l'accès aux données. Lors du dépôt des données sur FORSbase, les auteurs doivent remplir un champ intitulé *« Permission spéciale »* en choisissant entre deux modalités : (1) *Aucune permission spéciale* et (2) *Accord préalable de l'auteur*. Avec le choix de la seconde option, toutes les demandes de téléchargement des données – incluant une brève description des intentions d'analyse – sont soumises aux auteurs pour accord préalable.

Nous avons décidé de ne pas demander de permission spéciale afin de garantir un accès immédiat et indifférencié aux données. Étant donné que les modalités d'accès sont déjà encadrées de manière rigoureuse et professionnelle par FORS, il nous a semblé inutile et contre-productif d'ajouter des barrières supplémentaires aux personnes intéressées. Premièrement, l'expérience a montré que tout ajout d'un obstacle dans l'accès aux données augmente la probabilité de non-usage de celles-ci. Deuxièmement, notre décision ne pourrait qu'être subjective – et contribuer à créer des inégalités – en l'absence de critères prédéfinis sur lesquels baser notre acceptation ou notre refus d'octroi d'accès aux données.

Cette étape n'a pas pris beaucoup de temps, car les consignes lors du dépôt sur FORSbase étaient clairement identifiables et faciles à suivre.

Interopérable (interoperable)

Les données doivent généralement être intégrées à d'autres données et donc être « interopérables » pour l'analyse, le stockage et le traitement. En d'autres termes, elles doivent être prêtes à être échangées, interprétées et combinées de manière (semi-)automatisée avec d'autres ensembles de données par des êtres humains ou des systèmes informatiques.

Les (méta)données des enquêtes possèdent un fort degré de standardisation, et peuvent très facilement être comprises, partagées et appliquées dans des logiciels de traitement statistique standard (SPSS, R). Le format .sav utilisé est standard, et toutes les informations nécessaires à la bonne compréhension des données ont été répétées directement dans les tables (noms

de variables et libellés clairs), en reprenant la formulation exacte des dictionnaires des variables et des codebook (I1). Le vocabulaire est également standardisé (I2). Les métadonnées sont en français, c'est-à-dire dans la langue des données et du rapport d'enquête. Le lien entre les données et les différentes métadonnées est assuré par l'architecture de la plateforme FORSbase, qui indique de manière très explicite le référencement des différents documents (I3).

Réutilisables (Reusable)

L'objectif ultime de FAIR est d'optimiser la réutilisation des données. Pour y parvenir, les métadonnées et les données doivent être bien décrites afin qu'elles puissent être reproduites et/ou combinées dans différents contextes.

Avec le travail mené, la réutilisation des données est assurée moyennant un effort minimal de compréhension du contexte des trois enquêtes. Les (méta)données sont garanties conformes aux trois points précédents, et sont suffisamment bien décrites et riches pour être automatiquement liées ou intégrées à d'autres sources de données (R1). Elles sont publiées avec une licence d'utilisation des données qui précise les conditions dans lesquelles les données peuvent être utilisées et qui est claire et accessible (R2). Les métadonnées attestent de la provenance des données et de leur construction (R3), et sont directement intelligibles, car elles correspondent aux standards de la communauté scientifique des sciences humaines et sociales (R4).

De manière générale, le temps nécessaire à la préparation des fichiers finaux à partir des données sources a été d'environ deux semaines : une semaine pour l'établissement du script et des documents, et une deuxième semaine de rodage, de contrôle et de corrections. Il est impossible de décrire exactement le temps passé pour chacun des points précités ; encore moins pour les sous-points. En effet, ceux-ci sont fortement liés et la préparation des données englobe les différents aspects des principes FAIR de façon non linéaire[17]. La question sous-jacente est de savoir jusqu'où il est nécessaire de connaître la recherche pour effectuer le travail de dépôt des données et de préparation de la documentation. Dans le cas de la présente recherche, bien que la personne ayant effectué l'essentiel du travail sur les (méta)données ne faisait pas partie de l'équipe de recherche, le coût d'entrée était relativement bas, conséquence encore une fois de la standardisation de la méthodologie. Deux jours ont néanmoins été nécessaires pour se familiariser avec l'enquête et les données. Mais cela pourrait fortement varier dans le cas d'un autre type d'enquête.

La création du script a demandé de bonnes compétences du programme R et la connaissance de fonctions spécifiques au traitement des données d'enquête par questionnaire. Une expertise avérée dans ce genre de procédés a également été profitable. De plus, une bonne connaissance de la suite Office a été un atout pour gagner du temps dans la production de la documentation. La conjonction de ces compétences a permis de travailler directement sur une version finale, ce qui a permis de gagner du temps.

[17] https://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

Synthèse

Dans cette dernière partie, nous proposons une synthèse des tâches effectuées en y associant les compétences mobilisées et le temps approximatif (en jours). Le Tableau **2** ci-dessous présente un récapitulatif de ces éléments répartis en quatre axes : familiarisation avec l'enquête et les données, mise en forme et nettoyage des données, création des métadonnées et dépôt des données sur FORSbase. Pour l'ensemble du processus, le temps nécessaire est estimé à un mois (ci-après 20.5 jours).

La mise en open data d'une enquête est un processus relativement long et qui demande certaines compétences. Malgré une prise en main relativement aisée de l'outil FORSbase rendue possible par la qualité des guides et la disponibilité du personnel de FORS, le travail de mise à disposition des données nécessite de bonnes compétences en traitement des données et la maîtrise d'un logiciel statistique (R, SPSS). Afin de garantir l'intégrité des (méta)données, il est absolument nécessaire de travailler par scripts pour assurer la systématicité des opérations ainsi que leur traçabilité. Cette compétence technique doit se doubler d'une bonne connaissance des données de l'enquête. En effet, il est toujours impératif d'effectuer tous les contrôles nécessaires « à la main », la prise en main technique ne pouvant se soustraire au regard des personnes ayant mené la recherche. Ces allers-retours prennent du temps, mais sont incontournables. Il est aussi primordial que l'équipe de recherche possède une bonne connaissance générale de la méthodologie d'enquête et des enjeux éthiques pour répondre à un bon nombre de décisions relatives à l'octroi des licences et des droits d'accès, ainsi qu'à l'anonymisation des données.

En résumé, le travail de mise en ligne des données d'enquête nécessite un temps d'environ un mois, et des compétences techniques (maîtrise des techniques d'enquête et du traitement des données par logiciel) et théoriques (pour procéder aux choix en matière d'éthique, gérer les droits d'accès, décider des métadonnées à produire). Une bonne coordination de ces compétences est nécessaire, et l'intervention des personnes ayant effectué la recherche est indispensable, même dans le cas où le travail de mise en open data est confié à une tierce personne, comme dans le cas présent.

Tableau 2 : Récapitulatif des tâches, des compétences et du temps nécessaires à la mise en open data des données

Tâches	Compétences	Temps en jours
Familiarisation avec l'enquête et les données		**2**
Lecture du rapport final	Connaissance des techniques de recherche et d'échantillonnage, connaissances en méthodologie d'enquête par questionnaire	0,5
Prise de connaissance des données		0,5
Séances de discussion		1
Mise en forme et nettoyage des données		**7**
*Recodage uniforme des données**	Compétences en traitement des données à l'aide d'un logiciel statistique (R, SPSS), gestion des scripts, maîtrise des enjeux théoriques et éthiques des méthodes quantitatives en sciences sociales	3
Anonymisation		1
Nettoyage		1
Export des tables SPSS		-
Contrôle		2
Création des métadonnées		**6**
Création et mise en page des dictionnaires des variables	Compétences en traitement des données à l'aide d'un logiciel statistique (R, SPSS), gestion des scripts, maîtrise des enjeux théoriques et éthiques des méthodes quantitatives en sciences sociales	1
Création et mise en page des codebook		1
Contrôle		2
Contrôle de l'intégrité de l'ensemble des (méta)données		2
Dépôt des données FORSbase		**5.5**
Prise de connaissance de l'interface et lecture des guides	Pas de compétence particulière, même si une bonne connaissance en méthodologie est un atout.	1
Discussions avec une archiviste de FORS		0,5
Création du « projet »		0,5
Saisie des champs dans FORSbase et dépôt des (méta)données		1
Prise en compte des commentaires de l'archiviste de FORS		2.5
Temps total		**20.5 jours**

** Cette tâche inclut l'écriture, les tests et l'exécution du script en R qui permet d'assurer la traçabilité de l'ensemble des opérations menées.*

5 DIFFICULTES JURIDIQUES ET ETHIQUES

Nous l'avons mentionné, la nature des données de l'enquête auprès des enfants, des jeunes et des acteurs jeunesse et les compétences techniques mobilisables ont rendu le processus de préparation techniquement fluide. En d'autres termes, peu de questions se sont posées quant à la forme que devaient prendre les données et les documents les accompagnant, tant elles sont issues d'un processus d'enquête très standardisé et largement connu de la communauté. Le processus de partage des données en revanche a soulevé des problématiques éthiques relatives (1) à l'anonymisation des données, (2) au consentement éclairé au partage des données et (3) l'usage ultérieur des données par d'autres acteurs·trices. Nous les présentons succinctement ci-après en précisant les solutions mises en œuvre.

L'anonymisation des données

Le premier enjeu éthique auquel l'équipe de recherche a été confrontée est celui de l'anonymisation des données. On trouve chez Stam et Kleiner (2020, p. 3) une analyse approfondie des enjeux et des stratégies d'anonymisation définie comme *« un processus par lequel les éléments permettant l'identification d'une personne sont définitivement supprimés des données et de la documentation connexe, de sorte qu'un individu ne puisse être identifié sans effort important. »* Elle est à ne pas confondre avec la *pseudonymisation*, qui consiste en *« la suppression ou le remplacement des identifiants par des pseudonymes ou des codes, lorsque les identifiants sont conservés séparément et sécurisés par des mesures techniques et organisationnelles »* (Stam et Kleiner, 2020, p. 4).

Dans notre cas, en l'absence d'identifiants directs (nom, prénom, adresse, numéro de téléphone, p. ex.), l'anonymisation des données des enquêtes auprès des enfants et des jeunes aurait pu être compromise par le recoupement des variables de lieu de résidence, de nationalité et d'âge. Nous avons opté pour une stratégie d'anonymisation par *généralisation* en modifiant *« l'échelle des attributs des jeux de données, ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes. Cette technique permet d'éviter l'individualisation d'un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d'autres.* »[18]. La stratégie retenue a donc consisté à diminuer la précision du lieu de résidence et de la nationalité par un recodage du lieu de résidence en districts de résidence (Porrentruy, Delémont, Franches-Montagnes) et de la nationalité en trois zones géographiques (Suisse, pays de l'Union européenne, pays hors Union européenne).

Les contrôles effectués par l'archiviste de FORS lors du premier dépôt des données[19] ont révélé une anonymisation insuffisante des variables string (contenant du texte) dans les trois enquêtes. Dans certains cas, il était en effet théoriquement possible d'identifier les individus à partir d'informations laissées dans les questions ouvertes. Pour résoudre ce problème, trois stratégies différentes ont été mise en œuvre en fonction des variables concernées : (1) la suppression de *toutes* les valeurs d'une variable string, (2) le recodage des variables string de

[18] https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

[19] Avant leur mise en ligne publique, les données et la documentation associée sont soumises à des contrôles stricts par les archivistes spécialisé·e·s de FORS. Ces contrôles ont notamment pour objectifs d'assurer une saisie uniforme des champs de FORSbase, la correspondance entre les données et les métadonnées, ainsi que l'anonymisation effective des données.

type « autre » en variables binaires et (3) la suppression sélective de certaines valeurs des variables string permettant un long développement. Ces trois stratégies sont autant de manières différentes de gérer la tension existant entre la perte d'information et la nécessité d'anonymiser les données.

La première stratégie consistant à *supprimer toutes les valeurs d'une variable string* a été appliquée à la variable « Nom de l'association » de l'enquête auprès des acteurs jeunesse. Croisées avec l'âge et à la fonction occupée au sein de l'association, cette variable aurait permis d'identifier certain·e·s répondant·e·s. Pour l'éviter, nous avons supprimé toutes les valeurs de cette variable. La solution choisie n'a pas de conséquences particulières sur les potentialités d'analyse étant donné que le jeu de donnée contient une variable distinguant les organisations en fonction de leur domaine d'activité (sport, culture, politique, accueil et animation).

La seconde stratégie consistant à *recoder des variables string en variable binaires* a été appliquée à l'ensemble des variables « Autre » des enquêtes auprès des jeunes et des acteurs jeunesse permettant aux répondant·e·s d'indiquer par du texte une modalité de réponse non prévue. Dans certains cas, les réponses fournissaient des indications permettant d'identifier les répondant·e·s, car elles contenaient des informations, par exemple sur le lieu de résidence, une configuration familiale particulière ou des loisirs particuliers. Pour pallier ce problème, les 12 variables concernées (3 dans l'enquête enfants et 9 dans l'enquête jeunes) ont été recodées en variables binaires, 1 indiquant que le champ textuel était rempli, 0 qu'il était vide (Tableau **3**). Au total 252 valeurs ont été recodées. La faible fréquence du choix des modalités de réponse [Autre] et l'apport limité d'information qu'elles contiennent font que la perte d'information engendrée par le recodage pour anonymisation est faible.

Tableau 3 : Récapitulatif du traitement pour anonymisation des questions ouvertes de type « autre » des enquêtes auprès des enfants et des jeunes

Enquête	Variables	Nombre de valeurs supprimées
Enfants	L'activité (ou les activités) que vous pratiquez est (sont) de type : [Autre]	8
	Pourquoi ne pratiquez-vous pas d'activités proposées par un club de sport, une société ou un Espace-Jeunes ? [Autre]	39
	Qu'est-ce qui vous empêche de vous rendre plus souvent dans l'espace public ? [Autre]	42
Jeunes	L'activité (ou les activités) que vous pratiquez est (sont) de type : [Autre] [Recodée pour anonymisation]	5
	Pourquoi ne pratiquez-vous pas d'activités proposées par un club de sport, une association ou une société ? [Autre]	44
	À quelle activité ou quelles activités participez-vous ? [Autre]	15
	Pourquoi ne participez-vous pas à ces activités ? [Autre]	12
	Qu'est-ce qui vous empêche de vous rendre plus souvent dans les espaces publics jurassiens ? [Autre]	23
	Quelle est votre situation familiale actuelle ? [Autre]	32
	Quel type de diplôme envisagez-vous obtenir (dans le cadre de votre formation actuelle) ? [Autre]	3
	Quelle est la plus haute formation que vous avez achevée ? [Autre]	5
	Quelle est votre situation familiale ? [Autre]	24

La troisième stratégie d'anonymisation consistant à *supprimer uniquement certaines valeurs des variables string* a été appliquée aux quatre questions ouvertes permettant un long développement (1 dans le questionnaire enfants[20], 1 dans le questionnaire jeunes[21], 2 dans le questionnaire acteurs jeunesse[22]). Étant donné que ces variables contiennent des informations détaillées dont la suppression aurait considérablement réduit la richesse et les potentialités d'analyse des trois jeux de données, la suppression de toutes les valeurs ou le recodage des variables string en variables binaires n'étaient pas des stratégies pertinentes. Bien qu'étant une entreprise chronophage, nous avons décidé de contrôler *une à une* chacune des 1'344 valeurs des 4 questions ouvertes afin d'identifier les réponses contenant des

[20] Finalement, si vous aviez une baguette magique, qu'est-ce que vous souhaiteriez changer dans votre vie ?

[21] Finalement, si vous aviez une baguette magique, qu'est-ce que vous souhaiteriez changer dans votre vie ?

[22] D'après vous quels sont les problèmes, difficultés ou dangers principaux auxquels les jeunes jurassien·ne·s de 12 à 24 ans sont confronté·e·s aujourd'hui ?; D'après vous, quelle sont les envies, désirs ou attentes principales des jeunes jurassien·ne·s de 12 à 24 ans ?

indications permettant potentiellement d'identifier les répondant·e·s. L'analyse a montré que 31 réponses contenaient des indications pouvant mettre à mal la garantie de l'anonymat. Sept types de cas ont été identifiés (Tableau 4). Les 31 valeurs problématiques ont été supprimées.

Tableau 4 : Typologie des cas rencontrés dans le cadre de l'anonymisation des questions ouvertes des enquêtes auprès des enfants, des jeunes et des acteurs jeunesse

Types de cas	Nombre de cas		
	Enfants	Jeunes	Acteurs jeunesse
Indications sur le lieu de domicile des répondant·e·s ou de leur famille proche	6	5	0
Indications sur la formation planifiée ou la profession exercée par les répondant·e·s	1	2	0
Indications sur une maladie rare des répondant·e·s ou d'un parent proche	6	3	0
Informations sur une pratique sportive ou de loisir particulière des répondant·e·s	2	1	0
Informations concernant un projet personnel particulier des répondant·e·s	1	2	0
Informations concernant un·e ami·e proche des répondant·e·s	1	0	0
Informations sur le club ou l'association des répondant·e·s	0	0	2
Total	**16**	**13**	**2**

Le consentement éclairé à la réutilisation des données

Une autre difficulté relative aux questions juridiques et éthiques est celle de l'information aux répondant·e·s. Si le consentement éclairé des participant·e·s a pu être validé dans la procédure d'enquête, celles et ceux-ci n'ont pas été informé·e·s de la publication des données en libre en accès. En effet, à aucun moment du processus de conceptualisation et de réalisation du mandat il n'a été prévu de mettre en open data les données de l'enquête. L'équipe de recherche s'est lancée dans ce processus de manière opportuniste, car un financement était proposé par la HES-SO. Cette situation, qui est l'expression d'une culture de partage de données encore peu répandue dans les institutions des membres de l'équipe de recherche, n'est pas sans poser quelques questions éthiques en lien avec le consentement éclairé des participant·e·s.

En effet, si l'invitation à participer au questionnaire transmise aux enfants, aux jeunes et aux acteurs jeunesse précisait le cadre dans lequel était administré le questionnaire, le caractère anonyme des réponses, les institutions responsables de la collecte et de l'analyse des données, celle-ci ne mentionnait pas la possibilité de partage de donnée ni l'éventuel usage hors du cadre spécifique du mandat.

La spécification de l'utilisation future des données est pourtant importante, car susceptible d'influencer la décision des participant·e·s (Krügel, 2019). Il serait donc toujours préférable d'obtenir le consentement des participant·e·s lors de la récolte des données, même si une obtention rétrospective du consentement est également admissible. Dans le cas qui nous concerne, l'obtention rétrospective du consentement à la mise en open data des données était impossible à obtenir, car nous ne disposons d'aucun moyen de recontacter personnellement les répondant·e·s. Comme le relève cependant Krüger, *« if the opportunity to gain retrospective consent is not feasible, sharing and re-use of the data collected is legally permissible under certain circumstances, specifically if the original consent or the information conveyed at the time of the data collection does not explicitly preclude sharing, if no harm to participants is to be expected, and if the data are sufficiently anonymized* (Krügel, 2019, p. 8). N'ayant jamais explicitement exclu un partage des données, celles-ci étant totalement anonymisées et aucun préjudice pour les participant·e·s ne pouvant être légitimement attendu, nous avons considéré que la mise en open data était pas idéale, mais admissible.

L'usage ultérieur des données

Une troisième difficulté importante se situe à un niveau plus global. Même en multipliant les métadonnées et les annexes, aucun contrôle de la prise en compte de ces éléments ne peut être garanti dans l'absolu. Dans le cas qui nous concerne, bien que le contrat d'utilisateur de FORSbase stipule au point 5 que par sa signature l'utilisateur·trice s'engage à *« utiliser ces données de manière consciencieuse et informée, notamment en consultant la documentation, et dans le respect des règles de l'éthique scientifique »*[23], il est tout à fait envisageable qu'un·e utilisateur·trice ne prenne en compte qu'une partie du contexte de production, voire s'en distancie totalement. Le lien entre les données et l'enquête s'appuie, même lorsque toute la documentation est fournie, sur des normes et bonnes pratiques partagées par la communauté scientifique (Merton, 1973). Si ces normes sont effectivement partagées, car elles constituent le fondement même de la crédibilité du travail scientifique et de la légitimité des chercheur·e·s, elles sont aussi l'objet de luttes à la fois à l'interne de la sphère académique, mais aussi envers les pouvoirs publics et les pressions extérieures (Bourdieu, 1984, 2001 ; Gingras, 2013 ; Gingras et Gemme, 2006).

Le corollaire de cet argument est qu'il est également impossible de contrôler les connaissances des futur·e·s utilisateur·trice·s. S'il est possible d'exclure les entreprises privées des ayants droit aux données, il reste d'autres types d'utilisateur·trice·s potentiel·le·s, dont les ONG et les médias. Les mécanismes de fonctionnement de ces « champs » étant différents de ceux du champ académique – ne serait-ce que pour des questions de temporalité – l'on ne peut garantir le bon usage des normes implicites partagées par les chercheur·e·s. Ce dernier constat qui présuppose qu'il existe effectivement une communauté qui partage ces normes ne prend pas en compte les différences disciplinaires qui sont à la base du fonctionnement de la profession académique (Abbott, 1988, 2001).

D'importantes différences existent à l'égard du partage des données au sein des sciences sociales, que ce soit d'ailleurs d'un point de vue technique ou éthique. La procédure de dépôt des données implique de nombreux choix dont certains peuvent être effectués une fois l'enquête terminée (recodage des données quantitatives, définition de l'accès aux données,

[23] https://forsbase.unil.ch/media/general_documentation/fr/User_contract_F.pdf

anonymisation des identifiants dans les tables de données, etc.), alors que d'autres découlent de choix effectués en amont, et ne peuvent pas être effectués librement. Cet aspect est fortement présent dans les recherches qualitatives, où le travail sur les données fait partie intégrante de la production des données (Diaz, 2021). Le terme « données » fait d'ailleurs débat auprès de ces chercheur·e·s qui lui préfèrent le terme de « matériau ». Pour pousser la réflexion plus loin, le critère de reproductibilité des résultats ne saurait être applicable que dans certaines circonstances. Les normes de l'open science ont en effet été développées sur le modèle des *Big sciences* telles que la physique et la biologie fonctionnelle, qui répondent à des ontologies, épistémologies et méthodes très différentes des enquêtes par observation, entretiens, etc. Outre les méthodes de recueil d'information, les enjeux éthiques sont liés aux publics des enquêtes, encore une fois tout à fait différents des objets d'étude traditionnels de la physique et de la biologie (Roca i Escoda et al., 2020).

Comment par exemple publier un journal de terrain contenant toute la réflexivité nécessaire à l'interprétation d'une enquête, c'est-à-dire toute une série d'éléments non publiables en libre accès. Dans une prise de position intitulée « *Open Science, Data Management and Ethics in Anthropological Research Position Paper of the Swiss Anthropological Association (SAA)* » (2020), la Société suisse d'ethnologie revient sur un certain nombre d'enjeux fondamentaux pour la discipline qui, en définitive, rappellent que les enjeux « éthiques » ne peuvent pas être transposés sans adaptation préalable.

6 CONCLUSION

Ce rapport avait comme objectif de rendre compte des principales étapes effectuées pour la mise en open access des données de l'enquête auprès des enfants, des jeunes et des acteurs jeunesse. À la suite de cette première expérience de mise en open data de données d'enquête, nous souhaitons mettre en évidence quatre points de vigilance. Le premier se rapporte au caractère extrêmement chronophage et donc coûteux de processus. Si l'expérience accumulée dans ce domaine permet sans aucun doute de rendre le processus encore plus fluide et rapide, le temps nécessaire à la mise en open data de données restera toujours conséquent. Il apparaît deuxièmement que la mise en open data de données quantitatives demande des compétences techniques pointues sans lesquelles l'intégrité des données et leur correspondance avec les métadonnées ne peuvent être assurées. Ces compétences ne sont pas (ou plus) toujours disponibles au sein des équipes ayant mené les recherches et leur nécessité implique parfois l'engagement d'un spécialiste comme dans le cas présent. Un troisième élément à relever est que le processus d'anonymisation des données textuelles est plus complexe et prend plus de temps, notamment car il ne peut pas être automatisé. La qualité du travail est également assurée par la qualité des contrôles effectués par le personnel de FORS, dont nous relevons encore ici la compétence en matière d'encadrement.

Ces observations soulèvent inévitablement la question de la source et du montant du financement. Pour des recherches mandatées par les collectivités publiques, il apparaît illusoire que le coût du partage des données soit supporté par l'entité qui finance la recherche. Tout d'abord parce que les budgets disponibles pour ce type de mandats sont généralement juste suffisants pour réaliser le travail demandé et ensuite parce que le coût de la mise en

open data pourrait paraître disproportionné en regard du budget des recherches. Dans le cas qui nous concerne, les coûts nécessaires au partage des données sur FORSbase (hors rédaction du rapport scientifique) représentent plus du tiers (35%) du montant accordé par le Canton du Jura pour la réalisation du mandat. Le partage de données de qualité issue de recherches mandatées dépend ainsi fortement de source de financement externe telle que proposée par la HES-SO. Comme déjà mentionné, nous nous sommes lancés de manière opportuniste dans la mise en open data en raison de la possibilité de financement offerte par la HES-SO et ne pourrons envisager de le faire à l'avenir que si nous pouvons obtenir un financement spécialement dédié à cette tâche.

Dans cette conclusion, nous voulons ouvrir encore un peu plus le spectre de la généralisation et dresser un très rapide état des lieux de quelques contributions à l'étude des pratiques de l'open science. Un certain nombre d'études empiriques ont en effet tenté de mesurer de manière « objective » les attitudes de la communauté académique face au partage des données (Jeng et al., 2016 ; Kim et Stanton, 2013 ; Tenopir et al., 2011 ; Van den Eynden et al., 2016). De manière générale, on retrouve une vision positive de l'open data. Il apparaît cependant que la proportion de celles et ceux qui pensent qu'une culture de partage des données est importante est bien plus élevée que celles et ceux qui ont effectivement partagé des données. Il existe donc un décalage important entre les intentions et la pratique. L'hypothèse selon laquelle la pression des éditeurs ou des institutions de financement de l'activité scientifique conduirait à une augmentation des pratiques de partage (Kim et Stanton, 2013) n'a jusqu'à maintenant pas pu être confirmée (Jeng et al., 2016). Les études semblent attester de cette faible proportion de la production scientifique effectivement mise en ligne. À titre d'exemple, une étude portant sur des articles scientifiques de 50 revues à impact factor élevé a révélé que seuls 9% de ceux-ci avaient été déposés avec leurs données, malgré les exigences des revues (Alsheikh-Ali et al., 2011).

Un bilan plus nuancé ressort d'une enquête mandatée par le FNS en 2018 (von der Heyde, 2019) pour évaluer les pratiques et les besoins en matière de partage des données en Suisse : 75% des chercheur·e·s partagent leurs données d'une manière ou d'une autre – toute discipline confondue. Cependant, il apparaît que seulement 44% le font par l'intermédiaire de data repositories publics. C'est en fait le partage de personne à personne qui demeure le moyen le plus favorisé, et ce malgré l'effort mené en termes de politique scientifique incitative. De plus, la même étude relève aussi que le partage dans les sciences humaines et sociales n'est pas encore aussi courant que dans les autres disciplines (von der Heyde, 2019, p. 18). Une enquête menée par FORS en 2017 (Heers et al., 2017) indique que seuls 19% des chercheur·e·s en sciences sociales avaient déjà partagé leurs données via un dépôt en ligne (libre accès ou institutionnel), alors qu'ils et elles étaient environ 80% à considérer le partage des données comme très important.

Pour conclure, nous voulons rappeler qu'il est absolument nécessaire d'adapter les pratiques du partage des données aux spécificités ontologiques, épistémologiques et méthodologiques des sciences sociales. Les études précitées ont tendance à traiter de nombreuses disciplines scientifiques et n'abordent le cas des sciences sociales que marginalement. Ces dernières regroupent un éventail très large de disciplines et de méthodes qui présentent chacune des défis particuliers qui demandent un traitement spécifique, en particulier les méthodes et les approches « qualitatives », notamment parce qu'elles montrent un taux faible de partage, mais aussi de réutilisation des données (Faniel et Jacobsen, 2010 ; Wallis et al., 2013).

7 BIBLIOGRAPHIE

Abbott, A. (1988). *The system of professions: An essay on the division of labor*. The University of Chicago Press.

Abbott, A. (2001). *Chaos of disciplines*. The University of Chicago Press.

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., & Ioannidis, J.P.A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE, 6*(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Both, A. & Garcia, G. (2014). Le chercheur, l'archiviste et le webmaster : la polyphonie patrimoniale ? Le cas de beQuali, banque d'enquêtes qualitatives en sciences sociales. In B. Saou-Dufrene (Ed.), *Heritage and digital humanities: how should training practices evolve ?* (pp. 353-363). Lit Verlag.

Bourdieu, P. (1984). *Homo Academicus*. Editions de Minuit.

Bourdieu, P. (2001). *Science de la science et réflexivité*. Raisons d'Agir.

Davis-Kahl, S. (2016). Faculty self-archiving. In B.B. Callicott, D. Scherer, & A. Wesolek (Eds), *Making institutional repositories work* (pp. 143-158). Purdue University Press. https://digitalcommons.iwu.edu/ames_scholarship/113/

Diaz, P. (2021). Introduction: Archiving Qualitative Data in Practice: Ethical Feedback. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 150*(1), 7–27. https://doi.org/10.1177/0759106321995678

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW), 19*(3-4), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Feldman, S., & Shaw, L. (2019). The Epistemological and Ethical Challenges of Archiving and Sharing Qualitative Data. *American Behavioral Scientist, 63*(6), 699–721. https://doi.org/10.1177/0002764218796084

Gingras, Y. (2013). *Sociologie des sciences*. PUF.

Gingras, Y., & Gemme, B. (2006). L'emprise du champ scientifique sur le champ universitaire et ses effets. *Actes de la recherche en sciences sociales*, *164*(4), 51-60. https://doi.org/10.3917/arss.164.0051

Guirlet, M. (2020). Ouverture des données de recherche dans le domaine académique suisse : outils pour le choix d'une stratégie institutionnelle en matière de dépôt de données. *Ressi*, *21*. http://www.ressi.ch/num21/article182

Heers, M., Ferrez, E., & DePaula, E. M. (2017). *Data sharing and re-use: Researcher practices, attitudes and needs. FORS survey of social science researchers in Switzerland*. FORS. https://www.semanticscholar.org/paper/Data-sharing-and-re-use%3A-Researcher-practices%2C-and-Heers-Ferrez/9ba2e552f32f36683c0921f367ee57ee84ea79c8

Henry, G. (2014). Data Curation for Humanities. Perspectives from Rice University. In J. M. Ray (Ed.), *Research data management: Practical strategies for information professionals*. Purdue University Press. https://doi.org/10.2307/j.ctt6wq34t

Heim, J., Ischer, P., Thiévent, R., Kühr, J. & Tironi, Y. (2021). Plus on grandit, moins on utilise l'espace public. *Reiso : revue d'information sociale,* (18 mars). https://www.reiso.org/document/7160

Jeng,W., He, D., & Oh, J.S. (2016). Toward a conceptual framework for data sharing practices in social sciences: A profile approach. *Proceedings of the Association for Information Science and Technology, 53*(1), 1-10. https://doi.org/10.1002/pra2.2016.14505301037

Kim, Y., & Stanton, J.M. (2013). Institutional and individual influences on scientists' data sharing behaviors: A multilevel analysis. *Proceedings of the Association for Information Science and Technology, 50*(1), 1-14. https://doi.org/10.1002/meet.14505001093

Krügel, S. (2019). *The informed consent as legal and ethical basis of research data production*. (FORS Guide 05, Version 1.0). Swiss Centre of Expertise in the Social Sciences FORS. https://doi.org/10.24449/FG-2019-00005

Larouche, J. M., Genard, J. L., Roca i Escoda, M., & Diaz Venegas, P. A. (2020). Le contexte, les partenaires et le processus : les contraintes éthiques dans les recherches collaboratives. *SociologieS, La recherche en actes,* 1-10. https://doi.org/10.4000/sociologies.15268

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. The University of Chicago press.

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

Roca i Escoda, M., Burton-Jeangros, C., Diaz, P. & Rossi, I. (éds). (2020). Enjeux éthiques dans l'enquête en sciences sociales. *Sociograph*, 45. https://www.unige.ch/sciences-societe/socio/fr/publications/dernierespublications/sociograph-45-sociological-research-studies/

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

Stam, A., & Kleiner, B. (2020). *Data anonymization: legal, ethical, and strategic considerations*. (FORS Guide 11, Version 1.0). Swiss Centre of Expertise in the Social Sciences FORS. https://doi.org/10.24449/FG-2020-00011

Société suisse d'ethnologie. (2020). Open Science, Data Management and Ethics in Anthropological Research Position Paper of the Swiss Anthropological Association (SAA)

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE, 6*(6). https://doi.org/10.1371/journal.pone.0021101

Tironi, Y., Thiévent, R., Kühr, J., Ischer, P. & Heim, J. (2020). *Enquête auprès des enfants, des jeunes et des acteurs jeunesse* (Rapport de recherche). HETSL. https://www.hetsl.ch/fileadmin/user_upload/rad/recherche/Rapports/81486_Rapport-Jura_ACTEURS-JEUNES.pdf

Van den Eynden, V., Knight, G., & Vlad, A. (2018). *Open Research: practices, experiences, barriers and opportunities*. UK Data Archive. https://doi.org/10.5255/UKDA-SN-852494

von der Heyde, M. (2019). *Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future* (Report to the SNSF). https://doi.org/10.5281/zenodo.2643460

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE, 8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., …, & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(160018). https://doi.org/10.1038/sdata.2016.18