Le partage des données quantitatives en SHS : enjeux scientifiques et éthiques, conditions matérielles

Sébastien Oliveau

Le partage des données quantitatives en SHS : enjeux scientifiques et éthiques, conditions matérielles

Sébastien Oliveau,
IR* Progedo, CNRS
UMR Mesopolhis, CNRS, Université d'Aix-Marseille

Résumé :

A travers l'expérience de l'infrastructure de recherche Progedo, ce papier vise à éclairer par la pratique la question du partage des données. Il revient d'abord sur les enjeux de ce partage, d'un point de vue aussi bien scientifique qu'éthique. Il explique ensuite la manière dont les données sont partagées, d'un point de vue théorique, institutionnel, technique, et juridique. Il interroge enfin la nécessité de ce partage pour mettre en perspective les injonctions actuellement faites aux chercheurs.

Mots-clefs :

Données, diffusion, partage, FAIR, Infrastructure de recherche

Introduction

Nous nous proposons dans ce texte de présenter la question du partage des données scientifiques en France dans le contexte particulier des sciences humaines et sociales, en nous concentrant plus spécifiquement sur le cas des données quantitatives. Notre objectif est de porter un regard distancié sur les enjeux réels ou supposés de ce partage, mais aussi d'ancrer cette analyse dans une expérience pratique, pour essayer de dépasser les approches théoriques, voire normatives, qui prévalent aujourd'hui, en les confrontant aux usages observés. « Il faut ouvrir les données », certes, mais pourquoi et comment ?

Si nous avons choisi le champ assez restreint des données quantitatives en sciences humaines et sociales, c'est pour plusieurs raisons. De toute évidence, c'est le champ que nous maîtrisons le mieux, et nous souhaitons partir de notre pratique pour avancer nos réflexions. D'autre part, c'est un champ particulier que celui des sciences humaines et sociales, et il est toujours bon de le rappeler (Bourdeloie). En effet, l'objet de nos études est bien l'être humain, et cela conditionne plusieurs dimensions des données produites. Ainsi, l'expérimentation au sens des sciences expérimentales est plus compliquée pour au moins deux raisons. D'un point de vue éthique d'abord, nous partageons avec le secteur de la santé des limites à nos observations -qui ne doivent pas altérer les conditions de vie de nos sujets d'étude- mais que l'on doit envisager plus strictement en considérant que l'on ne peut pas enquêter sur des personnes sans leur consentement. Cela peut paraître trivial, mais il est nécessaire de s'en souvenir. D'un point de vue pratique ensuite : nous ne pouvons pas réitérer la même expérience deux fois sur un sujet, car sa perception est altérée par la première expérience. En allant plus loin, on rappellera aussi que l'observation est rendue difficile par la tendance du sujet à modifier son comportement en se sachant observé (c'est assez flagrant pour les enquêtes électorales par exemple).

Dans le vaste champ des données en sciences humaines et sociales, nous avons restreint notre réflexion au cas des enquêtes quantitatives et des données administratives. Elles forment un sous-champ spécifique, déjà très structuré dans leur production, leur diffusion, et leur traitement. Ces données posent néanmoins des questions que l'on peut retrouver ailleurs.

Dans un premier temps, nous allons revenir sur les enjeux du partage des données. Au-delà de la position de principe, que la tendance à la science ouverte a beaucoup mis en avant, nous souhaitons nous interroger sur

sa pertinence. Nous verrons ensuite comment ce partage s'effectue, quels en sont les cadres légaux et techniques. Nous nous appuierons pour cela sur l'expérience de l'équipe d'ingénieur·e·s du Centre National de la Recherche Scientifique des « archives de données issues de la statistique publique » (CNRS, France). Enfin, nous reviendrons sur les limites de ces injonctions au partage. Elles se situent dans le rapport à la donnée que la science entretient d'une part, mais aussi dans les moyens nécessaires pour sa mise en œuvre d'autre part.

1. Les enjeux scientifiques et éthiques du partage des données

a. Partager les données

Aborder le partage des données et ses enjeux nécessite que l'on s'arrête d'abord sur les termes utilisés et le contexte dans lequel ils le sont. La notion de données, comme celui de « données de la recherche » reste encore floue : c'est un terme « commode mais peu explicite » (Stérin et Noûs). C'est pourquoi nous proposons de distinguer « matériau » et « donnée ». Le matériau est ce qui est prélevé sur le terrain ou l'objet d'étude : ce peut-être un objet (matériel) ou une information (idéel). Certains qualifient parfois le matériau de « donnée brute », c'est-à-dire finalement non encore traitée (Borgman). Le matériau n'a pas vocation à être partagé, même si rien ne l'interdit, puisqu'il nécessite une médiation voire une transformation pour être compris. La donnée, quant à elle, résulte de la transformation ou de la médiation d'un matériau. Elle n'est plus simplement prélevée, elle est construite par son producteur (Gitelman). C'est d'ailleurs bien ce que nous dit le mot : elle est « donnée », et l'on retrouve en anglais et en allemand la même racine latine de *datum*. La donnée est une information codée, au sens où un chercheur lui a donné une signification spécifique définie par des normes éventuellement implicites, mais préférentiellement explicites.

La question du partage peut sembler plus simple. Même si l'étymologie de partage est bien celle de la découpe en part (autrement dit de la division en sous-élément), il est admis que l'information a paradoxalement comme caractéristique de ne pas se diviser lorsqu'on la partage, mais de se multiplier[1]. Le partage de l'information ne constitue donc pas un émiettement mais une diffusion de celle-ci, sa reproduction, à condition qu'elle soit bien conservée. À l'inverse, son absence de partage pose des problèmes vis-à-vis de l'éthique et questionne la première finalité de la recherche, à savoir accroître les connaissances.

La démarche de partage des données de la recherche n'est pas récente (Dasgupta et David) mais l'approche a été renouvelée, et surtout mise en lumière, par le développement depuis 20 ans des discours sur l'ouverture de la science, et particulièrement par son institutionnalisation depuis la fin de la décennie 2010. La science ouverte envisage que la science doit être le plus largement partagée et prône pour cela sa plus grande diffusion possible. Cela concerne d'abord la production de connaissance (les écrits scientifiques), mais induit aussi les conditions de cette production (données et codes informatiques).

b. Ouvrir les données

Dans le cas qui nous intéresse, à savoir les données quantitatives en sciences humaines et sociales, la tendance au partage des données est bien plus ancienne que les discours qui le prônent, et date d'avant la mise en place des politiques sur l'ouverture de la science et le partage des données. On peut distinguer différents cas de figure : les données produites par la recherche et les données utilisées par la recherche.

[1] La multiplication peut aussi entraîner une forme d'altération, c'est pourquoi la documentation de la donnée est primordiale.

Les données produites par la recherche sont le résultat de programmes de recherche dont les plus importants induisent dès leur création un partage parce qu'aucun individu seul ne peut les élaborer. Le partage se fait alors dès la création de l'enquête et dépasse le cadre de la donnée produite. Frédéric Gonthier[2], responsable pour la France de l'enquête ISSP parle ainsi d'« intellectuel collectif » pour souligner l'effort de coordination qui implique dès le début une logique d'harmonisation des dispositifs (modes d'administration) et de réflexion sur la comparabilité des indicateurs (questions).

Il s'agit typiquement de grandes enquêtes internationales qui ont vu le jour, pour les plus anciennes, au milieu du 20e siècle. On retient souvent l'enquête « *How Nations See Each Other* » (Buchanan et al.) comme la première du genre. Leur organisation et leur financement nécessite, dès leur conception des regroupements de chercheurs et d'institutions, et la question de leur partage est donc une des premières à laquelle il faut répondre. Si les producteurs ont une priorité sur les résultats, les programmes se sont tournés dès le départ vers une large diffusion qui renforce leur légitimité. Il existe aussi de nombreuses autres données produites par la recherche à des échelles plus fines, reposant sur des consortiums plus restreints, voire sur des petites équipes ou sur des individus. La diffusion et même la conservation de ces enquêtes pose question : la quasi-totalité des enquêtes ainsi produites ont disparu.

Les données utilisées par la recherche constituent un vaste champ de données possibles, depuis les productions locales (associations, collectivités territoriales, etc.) jusqu'à la statistique publique, c'est-à-dire les « registres administratifs » et les « enquêtes directes » produites par les gouvernements et leurs administrations (Desrosières, « Décrire l'État ou explorer la société : les deux sources de la statistique publique »). Si la qualité des premières doit être questionnée, on connaît l'utilité des secondes et leurs qualités ont été largement discutées (Desrosières, « Pour une sociologie historique de la quantification »). Néanmoins, l'accès à ces données n'a pas toujours été évident, et reste encore aujourd'hui sujet à difficulté (Bothorel). Longtemps accessibles seulement par un contact direct avec leur producteur (INSEE et services statistiques ministériels), leur numérisation a permis leur meilleure diffusion dès les années 1980 et le développement de l'internet a levé les dernières barrières techniques. Cette plus grande diffusion s'est accompagnée d'un questionnement sur les modalités de cette diffusion et les inégalités d'accès entre chercheurs. S'est aussi posée la question de la sauvegarde des données produites à moyen et long terme.

Nous verrons plus loin comment les dispositifs issus des initiatives de chercheurs se sont institutionnalisés pour atteindre aujourd'hui la forme d'une infrastructure de recherche (Progedo), parce que la réutilisation des données a constitué un enjeu majeur pour les SHS (Silberman).

c. Réutiliser les données

La réutilisation des données, c'est-à-dire l'usage par une personne n'étant pas son producteur, est bien ancrée dans les pratiques en sciences humaines et sociales. L'utilisation abondante des statistiques publiques en témoigne, que ce soit pour cadrer son sujet ou développer des analyses personnelles. Elle reste néanmoins encore difficile à mesurer de manière fiable. En effet, la citation de la source n'a pas encore connu la standardisation nécessaire à la mise en place d'une mesure d'impact bibliométrique. On peut espérer que le rapide développement des identifiants uniques et pérennes sur les jeux de données permette d'améliorer la situation.

[2] Communication personnelle.

Dans le cadre de notre infrastructure, nous avons plusieurs indicateurs qui nous permettent d'apprécier l'utilité des dispositifs de partage mis en place. Le premier est le nombre de demandes de données suivies d'un téléchargement. L'usage de nos données étant restreint par la loi (voir plus bas), notre public est circonscrit : seuls les chercheurs et les étudiants menant des recherches peuvent accéder à nos données, après inscription et autorisation.

Nous connaissons ainsi le nombre et le profil des utilisateurs. En augmentation régulière, les demandes de téléchargement ont dépassé le millier l'an dernier, pour plus de 900 utilisateurs. Chaque utilisateur peut demander plusieurs jeux de données, et un utilisateur peut s'inscrire pour avoir le droit d'utiliser les données dans un projet sans pour autant les demander (partage local avec les membres de son projet).

Nos utilisateurs[3] sont pour un peu plus de la moitié des étudiants, un quart de chercheurs et enseignants chercheurs, 15% de doctorant et 5 à 10% d'ingénieurs et chargés d'études. Les disciplines les plus représentées sont l'économie (un peu moins de la moitié des utilisateurs, et particulièrement chez les étudiants) et la sociologie (un tiers des utilisateurs). Les autres disciplines de sciences sociales sont moins bien représentées, pour différentes raisons. Cela peut être lié à la moindre utilisation du type de données que nous proposons (histoire, anthropologie) ou à l'utilisation majoritaire de données libres absentes de notre catalogue (géographie par exemple). On notera d'ailleurs que les utilisateurs sont présents au-delà des SHS, puisque l'on trouve aussi des épidémiologues ou des mathématiciens.

2. Les conditions pratiques du partage

Si le dispositif institutionnel s'est lentement renforcé pour atteindre aujourd'hui sa maturité et dispose désormais d'une reconnaissance indéniable sous la forme d'une infrastructure de recherche, il convient de rappeler que cela s'est opéré dans un contexte technique, juridique et scientifique en évolution constante. Aujourd'hui, un certain nombre de questions restent posées, notamment au regard de l'accès aux données confidentielles dans un contexte législatif qui renforce la protection de citoyens tout en rappelant la nécessité pour la recherche de pouvoir disposer de l'information produite par les administrations.

a. La question de la FAIRisation des données

Le partage des données nécessite à la fois des conditions institutionnelles, techniques, juridiques sur lesquelles nous allons revenir après avoir décrit un élément devenu central : le FAIR dans les données. On utilise depuis 2016 l'acronyme FAIR[4] (Wilkinson et al.) pour décrire les conditions pratiques optimales de la diffusion des données. Les données doivent être Faciles à trouver, Accessibles, Interopérables et Réutilisables. En termes plus concrets, cela se traduit par leur bon catalogage qui permettra de les trouver, et la disponibilité de ce catalogue. L'usage d'un identifiant unique et pérenne (de type Handle ou DOI par exemple) augmente la qualité de ce catalogage les rendant plus faciles à trouver grâce au web. L'accessibilité décrit les conditions juridiques mais aussi matérielles pour obtenir ces données. C'est une spécificité de notre service que d'offrir aux chercheurs des données qui ne sont pas ouvertes (données protégées par la loi) mais qui sont accessibles via un processus transparent (https://data.Progedo.fr/). Des standards ouverts sont promus (fichiers de type ascii par exemple plutôt que des formats liés à des logiciels propriétaires) pour assurer l'interopérabilité entre les données et les logiciels. Enfin, la réutilisation renvoie à la capacité de prise

[3] Ces informations figurent dans le rapport d'activité annuel de l'infrastructure.

[4] En anglais le jeu de mot est évident mais difficile à bien traduire. La notion de « *fair* » renvoie à des valeurs très positives regroupant à la fois ce qui est juste, équitable, loyal mais aussi raisonnable.

en main des données par des utilisateurs qui ne les ont pas produites, voire par les producteurs des données après plusieurs années. Pour cela, la documentation des données est essentielle. Il s'agit de créer de la donnée sur la donnée, la « métadonnée ». On peut d'ailleurs aller jusqu'à considérer qu'une donnée sans métadonnée n'est pas vraiment une donnée (mais plutôt une forme de matériau), puisqu'elle ne sera pas compréhensible. Pour ce faire, les sciences humaines et sociales ont mis en place au milieu des années 1990 un standard de documentation qui permet de décrire les données statistiques : la Documentation Data Initiative, communément appelée DDI (Vardigan, Heus et Thomas 2008).

On notera enfin que cette question de la « fairisation » des données est encore largement à promouvoir. L'*European Open Science Cloud* (EOSC), qui porte au niveau européen le partage des données de toute la recherche, a mis au cœur de stratégie le développement de données et de services FAIR (https://eosc-portal.eu). Des projets comme FAIR-IMPACT[5], auquel Progedo participe, œuvre à développer ces pratiques auprès de l'ensemble des communautés scientifiques.

b. Le contexte institutionnel

C'est au sein du Laboratoire d'Analyse Secondaire et de Méthodes Appliquées à la Sociologie (LASMAS), créé en 1986, que sont constituées les premières archives de données issues de la statistique publique (ADISP). 15 ans plus tard, et dans la continuité des préconisations du rapport sur « les sciences sociales et leurs données » (Silberman), est créé le Comité de Concertation pour les Données en Sciences Humaines et Sociales (Chenu) qui aura la responsabilité du « centre Quetelet » créé la même année[6].

En 2010 l'accord-cadre de coopération « réseau Quetelet » prend la forme d'un Groupement d'Intérêt Scientifique qui regroupe le CNRS, l'INED (son service d'enquête), l'EHESS (qui accueille alors l'ADISP au sein du centre Maurice Halbwachs), l'Université de Caen-Base Normandie, l'université des sciences et technologies de Lille 1, la fondation nationale des sciences politiques, l'école d'économie de Paris et l'IRDES.

Une simplification du dispositif s'opère graduellement au cours de la décennie 2010. La feuille de route des infrastructures de recherche française (2008) acte l'existence d'une Très Grande Infrastructure de Recherche nommée Progedo (PROduction et GEstion des DOnnées). Elle se matérialisera en 2012 par la création de l'Unité Mixte de Service « Quetelet Progedo » qui reprend l'activité du réseau Quetelet. En 2017, le CNRS décide d'affecter l'équipe de l'ADISP au sein de Quetelet-Progedo.

c. Quelques éléments de techniques

Avant même que la notion de FAIR ait vu le jour, les pratiques des archives de données en sciences sociales travaillaient déjà dans cette perspective. Le FAIR a néanmoins permis d'expliciter les attendus et de renforcer les bonnes pratiques.

Concrètement, cela se traduit par l'usage de la spécification DDI. DDI est un ensemble de spécifications écrites en langage XML (Extensible Markup Language) qui propose un ensemble de métadonnées visant à décrire l'enquête et ses variables : documentation de l'enquête, dictionnaire des variables, documentation des variables, voire questionnaires de l'enquête. Dans notre cas, la grande majorité des 1.500 enquêtes

[5] Le projet FAIR-IMPACT est financé par le programme Horizon Europe de la commission européenne (GA 101057344).

[6] L'Unité Mixte de Service « Centre d'archivage et de diffusion des données en sciences humaines et sociales – Quetelet » (UMR T2419) est créée en partenariat par le CNRS, l'EHESS, l'INED et l'université de Caen (Anonyme) et associe comme unité partenaire le LASMAS et le CIDSP (Centre d'Informatisation et de Données Socio-Politiques).

disponibles comprennent toutes ces informations, mais certaines ne sont pas décrites au niveau des variables. Ce sont néanmoins plus de 380.000 variables qui sont documentées dans le catalogue de Quetelet-Progedo-Diffusion. Le langage XML a été développé dans une perspective d'interopérabilité entre système d'informations hétérogènes.

À cet usage du DDI s'ajoute un catalogue en ligne (https://data.Progedo.fr) qui intègre pour chaque jeu de données un identifiant numérique unique (DOI) qui permet de rendre ces données faciles à trouver. L'usage du protocole OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) permet en outre de rendre ce catalogue moissonable par des métacatalogues. L'OAI-PMH est au format XML et repose sur l'utilisation de la norme ISO 15836 dite « Dublin Core ». Le site de DDI rappelle que « la DDI est destinée à se conformer à la norme Dublin Core » (https://ddialliance.org/resources/ddi-profiles/dc consulté le 07/05/2022).

On notera qu'un logiciel comme Zotero permet de récupérer les métadonnées pour les organiser comme on le fait avec des publications. On peut donc utiliser Zotero pour gérer son corpus de données (Fabre) puisqu'il est capable de convertir ceux fournis par Datacite, comme c'est le cas pour le catalogue Quetelet-Progedo-Diffusion.

Le catalogue Quetelet-Progedo-Diffusion est un logiciel libre construit sur mesure pour nos besoins par l'équipe de DBnomics (https://db.nomics.world). Le choix d'un développement *ad hoc* a reposé sur deux contraintes qui n'étaient pas prises en charge par les entrepôts standards existants de type « Dataverse ». La première était la volonté de pouvoir accéder aux métadonnées au niveau des variables (c'est-à-dire d'exploiter pleinement les possibilités de DDI), la seconde était la nécessité de pouvoir protéger les données non ouvertes par un système d'enregistrement et de contrôle de l'identité des requérants (obligation qui nous est imposée par la loi).

Il permet dans une interface unique d'explorer l'ensemble des métadonnées, aussi bien au niveau de l'enquête que des variables. Rapidement, il permettra aussi de commander les données à la volée, par un système de panier.

Nos équipes ont en parallèle commencé à implémenter des vocabulaires contrôlés pour la documentation. Les vocabulaires contrôlés sont des lexiques qui s'appuient sur une terminologie partagée et qui imposent l'usage de termes prédéfinis. Encore peu mis en avant, les vocabulaires contrôlés devraient être plus développés, car ils permettent d'optimiser la recherche d'information. La mise en place de vocabulaires contrôlés autorise en outre de lier des concepts (fonction essentielle pour le déploiement du web sémantique) et de proposer des traductions multilingues de manière automatisée. Nous nous appuyons pour cela sur le travail fourni au niveau du CESSDA : *Cessda Vocabulary Service* (https://vocabularies.cessda.eu) et le thesaurus ELSST (https://elsst.cessda.eu).

d. Le contexte juridique français

Pour comprendre les contraintes qui s'imposent à nous en termes de diffusion de données, et qui constituent la particularité de notre service, il nous faut revenir sur la législation qui encadre leur accès et l'originalité des dispositions françaises[7].

[7] Pour une analyse juridique détaillée concernant spécifiquement les données de la recherche, voir l'ouvrage d'Agnès Robin (Robin).

Pour aller vite, on peut dire que deux grands textes encadrent la diffusion des données en France. Le premier est le règlement général sur la protection des données (RGPD - règlement UE 2016/679 du Parlement européen et du Conseil du 27 avril 2016), qui encadre le traitement des données pour protéger les citoyens sur l'ensemble du territoire de l'Union européenne. Le second est la loi française « pour une république numérique » (loi n° 2016-1321 du 7 octobre 2016) qui travaille à ouvrir les données tout en renforçant la protection des personnes.

La première vise à protéger les citoyens et restreint donc l'accès aux données les concernant, la seconde vise à ouvrir les données pour qu'elles puissent servir au plus grand nombre. La loi vise donc à ce que les données soient, selon la formule maintenant classique « aussi ouvertes que possible, aussi fermées que nécessaire ». Cela produit deux grands types de données : les données anonymes, qui doivent être ouvertes et accessibles au plus grand nombre, et les données confidentielles, dont l'accès doit être soumis à des autorisations. Dans le contexte français, il y a d'une part le site officiel https://data.gouv.fr qui donne accès aux données ouvertes, et d'autre part un comité du secret statistique (https://www.comite-du-secret.fr/) qui autorise l'accès aux données confidentielles sous la forme de fichiers anonymes ou pseudonymes.

Les données ouvertes sont facilement téléchargeables en ligne mais posent pour la recherche deux problèmes. Le premier concerne la documentation qui les accompagne, qui peut être de qualité très variable et demeure trop succincte. Le second concerne leur préservation, qui n'est pas garantie. Dans ce contexte, il est nécessaire d'avoir un dispositif qui documente les données de manière plus détaillée et qui propose une conservation de plus long terme. On comprend déjà l'utilité d'un service comme celui offert par les archives de données issues de la statistique publique (Adisp-Progedo). Les données confidentielles sont quant à elle accessibles selon différents canaux. Le plus fréquent et le plus sécurisé est aujourd'hui le dispositif d'accès sécurisé aux données (CASD) qui permet de travailler sur des données confidentielles via une bulle informatique sécurisée. Le dispositif a néanmoins un coût non négligeable, avoisinant les 4.000€ annuel, à charge des chercheurs. Il existe aussi des fichiers dits « Fichiers Production et Recherche » (FPR) qui sont des versions comportant un niveau de détail moindre pour quelques variables potentiellement ré-identifiantes. Cela correspond à des données pseudonymisées[8]. L'accès à ces données nécessite aussi une autorisation du comité du secret statistique (via une procédure allégée). L'Adisp-Progedo gère l'accès et la diffusion de ces données à travers le dispositif « Quetelet-Progedo-Diffusion » (https://data.Progedo.fr), auquel participe aussi le service d'enquête de l'INED pour ses propres données pseudonymes.

3. Et si tout n'était pas si simple ?

La science ouverte repose sur des principes généreux et qu'il est a priori difficile de rejeter. Il s'agit de l'accès à la connaissance pour toutes et tous d'une part et du partage des données et des codes informatiques qui ont permis de les élaborer d'autre part (Rentier). Cela amène à une autre vertu, celle de la transparence dans l'élaboration des savoirs, qui passe par la reproductibilité des résultats obtenus, et constitue un levier important de la confiance faite à la science.

Notre rôle de scientifique est néanmoins d'interroger les fondements et les conséquences de cette plus grande ouverture de la science à l'instar du travail de Denis et Goëta (2017) ou du hors-série n°19 de la revue *tracés* (Galonnier et al.). On pourrait ainsi rappeler que la science repose depuis son origine sur la circulation des idées, et qu'elle n'a, en ce sens, jamais été complètement fermée. Il s'agit aujourd'hui de l'ouvrir totalement.

[8] « La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires » https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data, consulté le 09 mai 2022.

Dans le cadre qui nous intéresse (l'ouverture des données quantitatives en SHS), cela soulève au moins trois questions.

La première question concerne la pertinence d'étendre ce modèle d'ouverture des données. Nous avons indiqué en première partie que les matériaux n'ont pas forcément vocation à être ouverts, parce qu'y accéder n'a pas forcément de sens. Un certain nombre de données qualitatives peuvent aussi être ouvertes (les entretiens par exemple), mais sans que leur réutilisation ne soit possible. Quelle plus-value existerait alors à leur diffusion, alors que les coûts de mise en œuvre de la diffusion sont importants ?

Cette première question entraîne la seconde : quel est l'intérêt de la donnée dans la recherche en SHS ? On peut considérer en effet que ce qui compte le plus est avant tout l'analyse de la donnée. De ce point de vue, il est légitime de s'interroger sur la nécessité de conserver et de rendre accessible toutes les données.

On arrive enfin à la question de la conservation des données, qui, comme tout archivage, nécessite de faire des choix : faut-il conserver l'intégralité des données, au détriment de la qualité de leur description (faire passer la quantité avant la qualité) ou au contraire conserver moins de données, mais mieux documentées (faire passer la qualité avant la quantité) ? Une troisième voie serait aussi de réduire la quantité de données produites pour permettre de tout conserver de manière bien documentée.

Nous n'avons pas vocation à trancher ce débat ici. Tout au plus pouvons-nous rappeler la politique qui anime notre infrastructure de recherche. Il s'agit d'une part de promouvoir les entreprises collectives et la réutilisation des données, c'est-à-dire finalement de réduire en partie la production générale pour l'orienter vers une production de plus grande qualité. C'est le rôle du soutien aux grandes enquêtes et aux dispositifs collectifs nationaux. Il s'agit dans le même temps de conserver dans les meilleures conditions les données pour qu'elles soient réutilisables. Cela passe par l'usage et la mise en avant de pratiques « FAIR ».

À la suite de Christine Borgman (2020) nous pouvons reprendre la distinction proposée par le rapport de la National Science Foundation (National Science Board 2005, annexe D) qui propose de différencier trois catégories de données. La première est constituée des « research database », c'est-à-dire des données collectées en vue d'une utilisation immédiate par de petits groupes de chercheurs, souvent peu documentées et dont la sauvegarde au-delà du projet n'est pas spécifiquement envisagée. La seconde, nommée « ressource or community data », réunit les données utiles à une communauté spécifique et constituées par des groupes assez larges et coordonnés. Leur documentation répond aux standards établis dans la communauté en question ce qui facilite leur sauvegarde. Elles peuvent d'ailleurs finir par devenir des « reference collections » (la troisième catégorie de données) qui se caractérisent par la qualité et la robustesse des données produites. Ces « collections de référence » sont souvent soutenues par des financements récurrents, leur documentation répond aux standards et est harmonisée, leur sauvegarde est envisagée dès l'origine, et leur accès très ouvert.

Au sein de Progedo, les données de type *research database* sont peu nombreuses, mais nous les accueillons lorsque leurs qualités et la qualité de leurs métadonnées permettent une réutilisation. C'est par exemple le cas de la base « mariages parisiens » (Garden). Les données *ressource or community data* sont elles aussi présentes. On pourrait y ranger les « enquêtes ménages déplacements » (https://data.Progedo.fr/series/adisp/enquetes-menages-deplacements-emd) qui sont produites selon un même protocole pour différentes villes à différentes périodes. Elles ne constituent donc pas des *reference collections*, car leur programmation n'est pas régulière et qu'elles n'ont pas été envisagées initialement pour être sauvegardées. Les *reference collections* pour Progedo prennent d'abord la forme de grandes enquêtes internationales, comme l'enquête SHARE, l'enquête ESS ou encore ISSP, EVS ou GGP. Maintenues dans la

durée par de larges consortiums internationaux de recherche, leur qualité répond aux exigences des communautés scientifiques qui les soutiennent. Leur documentation, leur diffusion et leur sauvegarde ont été pensées en amont et respectent les normes en vigueur. Le catalogue étant très varié, il est néanmoins difficile de classer définitivement dans une catégorie telle ou telle jeu de données.

On conclura ce trop court développement en rappelant que si des économies sont possibles grâce à la mutualisation des productions et à la réutilisation des sources existantes, cela a un coût, difficile à supporter par des chercheurs isolés. Il y a donc nécessité à se regrouper et à se coordonner. C'est bien ce qui a conduit à l'émergence des dynamiques collectives décrites dans la seconde partie et qui a abouti aux différents dispositifs promus par l'infrastructure de recherche Progedo. Néanmoins tout cela doit passer par des investissements collectifs, car l'ouverture de la science a un coût. Dans un monde aux ressources finies, le temps passé à ouvrir la science est autant de temps pris sur le temps de sa production.

Conclusion : former et accompagner, la nécessaire étape vers un partage des données efficace

Si l'ouverture de la science et, dans ce cadre le partage des données, correspond aujourd'hui à une nécessité évidente, les coûts engendrés (surcharge de travail au détriment de la production de connaissance), doivent être interrogés au regard des gains obtenus. On peut penser néanmoins qu'il existe une possibilité de réduire cette charge, qui passe par l'intégration le plus en amont possible des contraintes liées à l'ouverture des données. Il est nécessaire de l'envisager le plus tôt possible. La mise en place des plans de gestion des données constitue une opportunité intéressante pour mener cette réflexion de manière non dogmatique. Les plans de gestion permettent en effet de réfléchir à l'ensemble de la chaîne de traitement des données depuis leur élaboration jusqu'à leur dissémination (Cartier et al.).

De manière plus importante encore, nous pensons qu'il est nécessaire d'accompagner le changement pour qu'il se réalise le plus rapidement possible. Cela passe par l'appui aux chercheurs et la formation (Oliveau et al.). C'est dans cet esprit qu'a démarré la première Plateforme Universitaire de Donnée (PUD) en 2002-2003 à Lille. Elle sera suivie en 2009 par le dispositif PANELS à Lyon et par une nouvelle PUD à Caen en 2011. Il en existe aujourd'hui 16 à travers toute la France, dont le développement a été très fortement poussé par Progedo depuis 2015, et accéléré par un fort soutien ministériel entre 2018 et 2021. Les PUD sont des dispositifs d'accompagnement des chercheurs sur les sites académiques, qui visent à faire découvrir les données existantes, former les étudiants et les chercheurs et accompagner les projets pour les mener vers les meilleures pratiques permettant la réalisation et la diffusion de leurs enquêtes quantitatives en sciences sociales. L'impact des PUD est assez remarquable si on le juge à l'aune des commandes de données auprès de l'ADISP. L'augmentation des demandes de fichier de données a progressé de 20% en 2020 et 2021 pour passer la barre des 1.000 demandes annuelles.

Il nous reste aujourd'hui la question de l'impact de cette distribution sur la production de connaissances. Nous ne possédons pas encore les outils pour mesurer le nombre de publications issues de la réutilisation de nos données. La mise en place d'outils spécifiques pour cela aura un coût aussi, mais il semble important de pouvoir mieux évaluer le bénéfice de l'ouverture des données.

Bibliographie

Anonyme. « Le centre Quetelet- Un nouveau centre pour les données en sciences sociales ». *BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique*, n° 75, 2002, p. 70‑72.

Borgman, Christine L. « Qu'est-ce que le travail scientifique des données ? : Big data, little data, no data ». *Qu'est-ce que le travail scientifique des données ? : Big data, little data, no data*, traduit par Charlotte Matoussowsky, OpenEdition Press, 2020. *OpenEdition Books*, http://books.openedition.org/oep/14692.

Bothorel, Éric. *Pour une politique publique de la donnée*. Mission à la demande du Premier ministre du 22 juin au 22 décembre 2020, Cabinet du premier ministre, 2020, p. 214.

Bourdeloie, Hélène. « Ce que le numérique fait aux sciences humaines et sociales. Epistémologie, méthodes et outils en questions ». *tic&société*, vol. 7, n° 2, 2013. *journals.openedition.org*, https://doi.org/10.4000/ticetsociete.1500.

Buchanan, William, et al. *How Nations See Each Other: A Study in Public Opinion*. Greenwood Press, 1953.

Cartier, Aurore, et al. « Chapitre 11. Gestion des données, partage et conservation pérenne avec le Data Management Plan ». *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*, édité par Étienne Cavalié et al., Presses de l'Université de Montréal, 2018, p. 159. *OpenEdition Books*, http://books.openedition.org/pum/11132.

Chenu, Alain. « Une infrastructure pour les données en sciences humaines et sociales ». *Courrier des statistiques*, n° 107, 2003, p. 29‑31.

Dasgupta, Partha, et Paul A. David. « Toward a New Economics of Science ». *Research Policy*, vol. 23, n° 5, septembre 1994, p. 487‑521. *ScienceDirect*, https://doi.org/10.1016/0048-7333(94)01002-1.

Denis, Jérôme, et Samuel Goëta. « La fabrique des données brutes : Le travail en coulisses de l'open data ». *Ouvrir, partager, réutiliser : Regards critiques sur les données numériques*, édité par Clément Mabi et al., Éditions de la Maison des sciences de l'homme, 2017. *OpenEdition Books*, http://books.openedition.org/editionsmsh/9050.

Desrosières, Alain. « Décrire l'État ou explorer la société : les deux sources de la statistique publique ». *Genèses*, vol. 58, n° 1, 2005, p. 4‑27, https://doi.org/10.3917/gen.058.0004. Cairn.info.

---. « Pour une sociologie historique de la quantification : L'Argument statistique I ». *Pour une sociologie historique de la quantification : L'Argument statistique I*, Presses des Mines, 2013. *OpenEdition Books*, http://books.openedition.org/pressesmines/901.

Fabre, Chloée. « Chapitre 3. Zotero : la gestion de références bibliographiques et de corpus documentaires ». *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*, édité par Étienne Cavalié et al., Presses de l'Université de Montréal, 2018, p. 55. *OpenEdition Books*, http://books.openedition.org/pum/11105.

Galonnier, Juliette, et al. « Ouvrir les données de la recherche ? » *Tracés. Revue de Sciences humaines*, n° #19, #19, décembre 2019, p. 17‑33.

Garden, Maurice. *Étude « Mariages parisiens - 1885 » | Catalogue*. 1997, https://data.progedo.fr/studies/doi/10.13144/lil-1471.

Gitelman, Lisa, éditeur. *« Raw Data » Is an Oxymoron*. MIT Press, 2013.

National Science Board. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. National Science Foundation, 2005, https://www.nsf.gov/pubs/2005/nsb0540/.

Oliveau, Sébastien, et al. « Aix-Marseille University SSH data platforms: Skills to support research in social sciences and humanities (SSH) in the Mediterranean ». *Egypte/Monde arabe*, vol. n° 22, n° 2, décembre 2020, p. 95‑105. *www-cairn-info.lama.univ-amu.fr*, https://doi.org/10.4000/ema.13176.

Rentier, Bernard. *Science ouverte, le défi de la transparence*. Académie Royale de Belgique, 2018.

Robin, Agnès. *Droit des données de la recherche. Science ouverte, innovation, données publiques*. Larcier, 2022.

Silberman, Roxane. *Les Sciences sociales et leurs données : rapport au ministre de l'éducation nationale et de la technologie*. 1999, http://www.ladocumentationfrancaise.fr/rapports-publics/004000935/index.shtml.

Stérin, Anne-Laure, et Camille Noûs. « Ouverture des données de la recherche : les mutations juridiques récentes ». *Tracés. Revue de Sciences humaines*, n° #19, #19, décembre 2019, p. 37‑50. *journals.openedition.org*, https://doi.org/10.4000/traces.10603.

Wilkinson, Mark D., et al. « The FAIR Guiding Principles for scientific data management and stewardship ». *Scientific Data*, vol. 3, mars 2016, p. 160018.