# Ouvrir les données de la recherche ?

Anthony Pecqueux, Juliette Galonnier, Stefan Le Courant, Camille Noûs

**Tracés. Revue de Sciences humaines**

**#19 | 2019**
**Les sciences humaines et sociales au travail (ii): Que faire des données de la recherche ?**

# Ouvrir les données de la recherche ?

**Juliette Galonnier, Stefan Le Courant, Anthony Pecqueux et Camille Noûs**

**Édition électronique**
URL : http://journals.openedition.org/traces/10588
ISSN : 1963-1812

**Éditeur**
ENS Éditions

**Édition imprimée**
Date de publication : 31 décembre 2019
Pagination : 17-33
ISBN : 979-10-362-0227-8
ISSN : 1763-0061

Ce document vous est offert par Centre national de la recherche scientifique (CNRS)

**Référence électronique**
Juliette Galonnier, Stefan Le Courant, Anthony Pecqueux et Camille Noûs, « Ouvrir les données de la recherche ? », *Tracés. Revue de Sciences humaines* [En ligne], #19 | 2019, mis en ligne le 22 juillet 2020, consulté le 24 août 2020. URL : http://journals.openedition.org/traces/10588

ÉDITORIAL

# Ouvrir les données de la recherche ?

JULIETTE GALONNIER
STEFAN LE COURANT
ANTHONY PECQUEUX
CAMILLE NOÛS

Impossible de passer à côté de l'attention suscitée au cours des dernières années par ce qu'il est devenu commun d'appeler *les données de la recherche*. Le foisonnement de colloques, séminaires et formations qui leur sont consacrés dans l'ensemble du monde académique en témoigne[1]. Cette prolifération d'événements s'inscrit dans un contexte où les initiatives institutionnelles, les controverses scientifiques et les évolutions législatives concernant différents aspects des données de la recherche se multiplient. Cet engouement s'insère plus largement dans le courant des politiques scientifiques actuelles dites de la *science ouverte* (*open science*) qui jusqu'à récemment concernaient principalement l'accès libre aux publications (*open access*) et s'étendent depuis peu aux données de la recherche (*open data*). Le plan national pour la science ouverte, dévoilé le 4 juillet 2018 par Frédérique Vidal, ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation, comporte ainsi un axe intitulé « Structurer et ouvrir les données de la recherche ». Il prévoit de « rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics ; créer la fonction d'administrateur des données et le réseau associé au sein des établissements ; créer les conditions et promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs »[2]. Il est ainsi désormais attendu qu'à l'échelle de chaque projet, les données de la recherche soient gérées selon un ensemble de bonnes

1 Il serait trop long d'en proposer une liste exhaustive ici, tant les institutions se sont emparées de ces thématiques. Sans même parler des activités autour des *big data*, les événements fluctuent entre discussions scientifiques sur les données de la recherche et formations des différents personnels au nouveau paysage légal et institutionnel.

2 Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, *Plan national pour la science ouverte*, 4 juillet 2018, p. 6-7.

pratiques regroupées sous l'acronyme FAIR : des données « faciles à trouver, accessibles, interopérables et réutilisables ». Lors des Journées nationales de la science ouverte qui se sont tenues les 18 et 19 novembre 2019, Antoine Petit, président-directeur général du CNRS, a présenté la feuille de route du CNRS pour la science ouverte, laquelle prévoit des données « FAIRisées » et la promotion d'outils pour l'analyse et la fouille des textes de données[3]. Au niveau européen, les financements sont conditionnés à la mise en place d'un plan de gestion des données (PGD) ou *data management plan* (DMP). En France, l'Agence nationale de la recherche (ANR) demande également l'élaboration d'un PGD pour les projets financés depuis 2019. Il devra être transmis « dans les 6 mois qui suivent le démarrage scientifique du projet »[4].

## Un moment « données »

On le voit, cette nouvelle attention aux données s'inscrit dans un contexte d'évolution de l'économie générale de la science, marqué notamment par la promotion du financement de la recherche par projets, qui s'est accéléré depuis les années 1980, conduisant à de profondes transformations des métiers de la recherche (Hubert et Louvel, 2012 ; Aust, 2014 ; Giry et Schutz, 2017). Dans le nouveau champ institutionnel de l'enseignement supérieur et de la recherche, les « données » se voient attribuer une place de plus en plus centrale : de l'énergie leur est consacrée et des moyens sont alloués pour leur archivage (et ce alors même que des pans entiers du monde académique souffrent d'un manque criant de ressources). Un nouveau vocabulaire a également vu le jour. Outre la multiplication des acronymes, on parle désormais de *banques de données,* de *FAIRisation des données*, de *(digital) data curation* (conservation des données numériques), de *data set* (jeu de données) ou encore de *data-driven science* (science axée sur les données).

Ce mouvement de l'*open data*, dont Anne-Laure Stérin retrace les développements juridiques dans son article pour ce numéro, est par ailleurs adossé à des débats plus épineux sur la reproductibilité et la vérifiabilité de la science, avec l'idée que les chercheurs et chercheuses devraient fournir leurs jeux de données en complément de leurs publications pour que

3 Voir [URL : https://webcast.in2p3.fr/container/journees-nationales-de-la-science-ouverte-fr], consulté le 23 novembre 2019.

4 ANR, « L'ANR met en place un plan de gestion des données pour les projets financés dès 2019 », 5 septembre 2019, [URL : https://anr.fr/fr/actualites-de-lanr/details/news/lanr-met-en-place-un-plan-de-gestion-des-donnees-pour-les-projets-finances-des-2019/], consulté le 18 octobre 2019.

leur recherche soit « transparente » et « réplicable » par d'autres (Desquilbet *et al.*, 2019). Cela vaut particulièrement pour les travaux quantitatifs et statistiques, pour lesquels les données et les opérations de code ou de codage sont d'ores et déjà exigées par certaines revues. En juillet 2019, l'École des hautes études commerciales de Paris (HEC) et l'université d'Orléans, avec l'appui du CNRS, ont par exemple lancé conjointement le service de certification Cascad (Certification Agency for Scientific Code and Data), un processus d'évaluation à destination des revues scientifiques permettant d'apprécier la reproductibilité des résultats d'un article s'appuyant sur des données quantitatives. La note ou *rating* de reproductibilité attribuée par ce service de certification allégerait ainsi le travail d'évaluation des revues[5]. Si ce mouvement touche principalement les sciences dites *exactes* et dans une certaine mesure l'économie, des incursions dans le champ des sciences humaines et sociales, y compris pour les travaux qualitatifs, ont également été repérées, comme le relate Sophie Duchesne dans ce numéro. Certaines revues demandent ainsi aux auteur-e-s de joindre les données sur lesquelles s'appuient les articles qu'elles publient ; c'est le cas de l'*American Journal of Political Science* (AJPS) depuis mars 2019 (Leighley, 2019)[6].

Le slogan englobant d'*open science*, pour désirable qu'il puisse paraître, nous invite à questionner le lien qui unirait *open access* et *open data*. En effet, si l'enjeu principal est à chaque fois de chercher à renouer les liens distendus entre sciences et sociétés, ce serait pour autant sur des plans bien différents : celui de l'accessibilité des publications et de leur éventuelle appropriation par les citoyen-ne-s dans le cas de l'*open acess* ; et celui de l'évaluation et du contrôle de la fiabilité du travail des chercheurs et chercheuses, rendus possibles avec l'*open data*[7]. En somme, la période actuelle se caractérise

5 Christophe Pérignon, Kamel Gadouche, Christophe Hurlin, Roxane Silberman et Eric Debonnel, 2019, « Certify reproducibility with confidential data », *Science*, vol. 365, nº 6449, p. 127-128. Voir aussi « Pour une certification de la reproductibilité des études économiques », *Les Echos*, 5 juillet 2018, [URL : https://www.lesechos.fr/idees-debats/cercle/pour-une-certification-de-la-reproductibilite-des-etudes-economiques-134135], consulté le 19 octobre 2019.

6 Le premier article qualitatif agrémenté de ses *replication data* publié par *AJPS* est paru en avril 2019 (Carnegie Allison et Carson Austin, 2019, « The disclosure dilemma : nuclear intelligence and international organizations », *American Journal of Political Science*, vol. 63, nº 2, p. 269-285). Il s'accompagne d'un corpus de *replication material* hébergé par le Harvard Dataverse Network ; ce corpus comprend notamment les documents d'archive utilisés ainsi que des extraits d'entretiens, lesquels sont anonymisés et – notons-le – très peu contextualisés.

7 La question de la mise à disposition des données s'inscrit en effet dans un contexte de méfiance, avec la multiplication d'articles publiés puis retirés après des reproductions d'expériences jugées non concluantes. Le champ de la biologie française a notamment été marqué par une série de signalements anonymes (ou de dénonciations selon les points de vue) sur le site PubPeer : des pairs ont pointé l'existence de résultats manipulés par leurs collègues. Dans ces différentes affaires, à chaque fois la divulgation des données a été érigée comme la solution. Voir David

par l'accélération d'un mouvement de fond à l'œuvre depuis au moins une vingtaine d'années, avec la création d'outils favorisant la science ouverte (par exemple les archives ouvertes d'articles en dépôt comme arXiv depuis 1991, ou HAL depuis 2001) et une succession de débats scientifiques suscités par l'archivage et la reproductibilité – d'abord apparus au Royaume-Uni et aux États-Unis comme le rappelle Sophie Duchesne dans son texte.

Notons toutefois que l'intérêt pour les données ne se réduit pas à la réplicabilité. En France, une nouvelle revue, *Sources : revue interdisciplinaire sur les matériaux et leurs usages dans les études africaines*, fait de la valorisation des matériaux collectés par les chercheurs et les chercheuses le cœur de son projet scientifique : le « retour aux sources du terrain » qu'elle propose de mettre en œuvre implique d'accompagner chaque article des matériaux utilisés et d'une réflexion sur leur contexte de production, afin de mettre en lumière le « va-et-vient entre interprétation et données »[8]. D'*AJPS* à *Sources*, on constate donc à quel point l'enjeu de mise à disposition des données de la recherche peut traduire des conceptions distinctes de la recherche et du travail des revues (Damerdji *et al.*, 2018).

Pour répondre aux demandes d'ouverture et d'accessibilité, les institutions d'enseignement et de recherche se sont par ailleurs dotées de nouvelles solutions techniques permettant de gérer et diffuser les données. La très grande infrastructure de recherche (TGIR)[9] Progedo (Production et gestion des données) créée en 2014 porte principalement sur les données d'enquêtes quantitatives et de statistiques publiques. Elle dispose désormais de nombreux relais régionaux de diffusion, au sein des plateformes universitaires de données (PUD). En 2019, on recensait douze PUD, dont la plupart ont ouvert au cours des cinq dernières années[10]. En outre, de nombreux entrepôts institutionnels de données ont récemment été créés pour héberger les données des chercheurs et chercheuses, comme le service Nakala mis en place par la TGIR Huma-Num qui vise à faciliter le tournant numérique des sciences humaines et sociales. Le contexte est à l'accélération. L'EHESS a par exemple inauguré en juin 2019 Didómena, son propre

Larousserie, « La biologie française minée par des manquements à l'intégrité scientifique », *Le Monde*, 23 octobre 2018, [URL : https://www.lemonde.fr/sciences/article/2018/10/23/la-biologie-francaise-minee-par-l-inconduite-scientifique_5373162_1650684.html], consulté le 21 décembre 2019.

8 Voir [URL : http://imaf.cnrs.fr/IMG/pdf/edito_sources_fr.pdf], consulté le 20 novembre 2019.

9 Outre les infrastructures de recherche (IR) qui relèvent des différents opérateurs de recherche, les très grandes infrastructures de recherche (TGIR), telles Huma-Num ou Progedo, relèvent d'une stratégie gouvernementale.

10 À Lille, Caen, Nanterre, Strasbourg, Rennes, Nantes, Dijon, Poitiers, Lyon, Grenoble, Aix-Marseille, Toulouse. [URL : http://www.progedo.fr/promouvoir/plates-formes-universitaires-de-donnees/], consulté le 20 novembre 2019.

entrepôt de données de la recherche ; l'IRD a ouvert DataSuds en septembre 2019 et Sciences Po DataSpire à l'automne 2019[11]. La multiplication de ces infrastructures techniques vient matérialiser de façon très concrète ce moment « données » et va, sans nul doute, orienter et contraindre les activités de recherche à venir (Jarrige, Le Courant et Paloque-Bergès, 2018).

Prenant acte de la prolifération existante, ce numéro propose de marquer un temps d'arrêt pour porter un regard réflexif sur les pratiques des acteurs et actrices des métiers de la recherche au sens large : quelles sont ces fameuses données avec lesquelles nous travaillons ? Qu'est-il attendu que nous en fassions exactement ? Que voudrions-nous en faire et qu'en avons-nous fait par le passé ? Comme nous le verrons, les réponses à ces questions varient fortement d'une époque ou d'un pays à l'autre, d'une discipline à l'autre, d'un établissement à l'autre, et d'un chercheur ou d'une chercheuse à l'autre. En effet, l'une des particularités de ce moment « données », ce sont aussi les controverses et les dilemmes qu'il suscite, ainsi que les disparités qu'il ne va pas manquer d'occasionner dans la mesure où il est aussi fortement conditionné à la question des financements (ou à leur absence).

## Données personnelles et données sensibles

Le mouvement de mise à disposition des données de la recherche se déploie en parallèle de l'évolution de la législation en matière de protection des données personnelles. Entré en vigueur le 25 mai 2018 à l'échelle de l'Union européenne, le règlement général sur la protection des données (RGPD) encadre désormais strictement le traitement des données personnelles, soit « toute information se rapportant à une personne physique identifiée ou identifiable »[12]. Dans la perspective du RGPD, le traitement des données personnelles n'est licite qu'à certaines conditions[13]. Le RGPD établit en outre que la mise en conformité avec ses principes relève de la responsabilité

11 En France, ces plateformes et entrepôts sont recensés sur Cat Odipor (catalogue pour une optimisation du partage et de l'interopérabilité des données de recherche).

12 Cela inclut les nom, prénom, adresse, numéro de téléphone, numéro de sécurité sociale, données biométriques, ADN, voix, image, et plusieurs éléments spécifiques se rapportant à l'identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale d'une personne.

13 Ces exceptions sont les suivantes : si « la personne concernée a consenti au traitement de ses données à caractère personnel pour une ou plusieurs finalités spécifiques » ; si le traitement est nécessaire « à l'exécution d'un contrat », « au respect d'une obligation légale », « à la sauvegarde des intérêts vitaux de la personne concernée ou d'une autre personne physique », « à l'exécution d'une mission d'intérêt public ou relevant de l'exercice de l'autorité publique » et « aux fins des intérêts légitimes poursuivis par le responsable du traitement ou par un tiers, à moins que ne

des organisations et des institutions (avec un risque de sanctions administratives), une tâche qui incombe désormais aux délégué-e-s à la protection des données (DPO pour *data protection officer*) nommé-e-s dans chaque institution. Les implications pour la recherche sont nombreuses. Si auparavant une demande était censée être déposée à la Commission nationale de l'informatique et des libertés (CNIL) avant de mettre en place un protocole de recherche (une procédure dont très peu de chercheurs et chercheuses avaient connaissance et à laquelle quasiment aucun-e ne se pliait, hormis pour les grandes enquêtes statistiques), ce sont désormais les institutions qui incitent fortement leurs personnels à déposer une demande de traitement de données auprès de leur DPO avant de commencer une recherche. Les répercussions de ces changements sur la conduite de la recherche sont longuement abordées dans l'entretien que nous avons mené avec le service des enquêtes de l'Institut national d'études démographiques (INED).

Plus largement, les recommandations du RGPD sur la gestion des données entrent parfois en contradiction avec les principes de l'*open data*. Même si le traitement de données « à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique, ou à des fins statistiques » bénéficie de dérogations (article 89), les principes de « minimisation » des données et de « limitation de la conservation » doivent être pris en compte dans l'archivage. Comme le rappellent les membres du service des enquêtes de l'INED, cela implique un certain nombre de précautions lorsqu'il s'agit précisément d'ouvrir et de partager ces données (vérification méticuleuse du respect de l'anonymat et de l'impossibilité d'identifier les personnes, restriction de l'accès aux données à la seule communauté scientifique). En pratique, et c'est là un des paradoxes du mouvement actuel, la mise à disposition peut donc s'accompagner d'une perte de la richesse des données, proportionnelle à l'ampleur de leur ouverture. Se dessine ainsi une tension entre d'un côté l'exigence de transparence et de reproductibilité d'une science ouverte au public et de l'autre un impératif de protection des enquêté-e-s et de leur vie privée.

Dans son article, Marwan Mohammed s'inquiète quant à lui du risque de bureaucratisation et de neutralisation de la recherche portant sur des sujets dits *sensibles* (criminalité, secret industriel, etc.), avec l'allongement, voire le refus, des demandes d'autorisation pour certaines recherches, notamment qualitatives. Cette évolution est selon lui préoccupante en ce qu'elle implique des contraintes administratives supplémentaires pour

prévalent les intérêts ou les libertés et droits fondamentaux de la personne concernée […], notamment lorsque la personne concernée est un enfant ».

les chercheurs et les chercheuses sans véritablement s'accompagner d'une réflexion sur la protection de ces dernier-e-s, face notamment aux incursions de la justice pénale ou des services policiers qui s'intéresseraient à leurs données de recherche[14]. En effet, refuser de mettre à disposition ses données ne signifie pas forcément que l'on souhaite dissimuler ses méthodes de travail ou tricher sur ses résultats : la rétention de données traduit aussi parfois une volonté de protéger les enquêté-e-s (notamment lorsque des pratiques illégales ou contestables sont mises au jour) mais aussi soi-même en tant que chercheur ou chercheuse (afin d'éviter de faire face à des poursuites ou de perdre l'accès au terrain et la confiance de ses enquêté-e-s). Cette tension est d'autant plus vive que le monde académique ne peut pas prétendre à la protection de ses sources en cas d'enquête policière ou judiciaire, contrairement au journalisme – par exemple – qui dispose de la possibilité d'invoquer le secret des sources. À ce sujet, Félix Tréguer décrit dans son article les risques d'exposition des chercheurs et chercheuses et de leurs interlocuteurs et interlocutrices d'enquête à ce qu'il qualifie de *surveillance d'État*. Cette dernière est rendue possible par l'évolution législative étendant les pouvoirs des services de renseignements et par l'usage généralisé et peu contrôlé que font de nombreuses institutions d'enseignement et de recherche des outils numériques proposés par des grands groupes tels Google ou Microsoft, qui ont contracté des partenariats avec les autorités. Félix Tréguer propose alors aux lecteurs et lectrices de *Tracés* plusieurs solutions techniques pour sécuriser leurs matériaux numériques et leurs ordinateurs.

Pour résumer, la période récente se caractérise par trois mouvements concomitants : l'accélération du mouvement de la science ouverte, adossé aux exigences de transparence et de réplicabilité ; l'évolution des cadres législatifs sur les données personnelles ; et l'extension du domaine du renseignement. Ces trois mouvements trouvent un point de jonction sur la question des données de la recherche, qui cristallisent des enjeux à la fois politiques, scientifiques, infrastructurels, éthiques et législatifs : *open science*, réplicabilité, développement de nouvelles solutions techniques, protection des données personnelles, surveillance de la recherche, etc. Tous ces éléments dessinent les contours du moment « données » que propose d'explorer ce hors-série du cycle « Les sciences humaines et sociales au travail ».

14 Sur ce sujet, outre les cas mentionnés par Marwan Mohammed et Félix Tréguer dans leurs articles respectifs pour ce numéro, voir le cas de Marie-Ève Maillé, contrainte par la justice canadienne à révéler les noms des participants à son enquête sur les effets d'un projet éolien. Radio Canada, « Une chercheuse forcée par la justice de révéler l'identité de ses sources », 31 octobre 2016, [URL : https://ici.radio-canada.ca/nouvelle/811463/source-identite-uqam-chercheuse-eolienne-entreprise-ordonnance-scientifiques-canadiens], consulté le 23 novembre 2019.

Comme souvent, ce qui se présente comme porteur de nouveaux enjeux n'est pas si nouveau. Christelle Rabier rappelle ainsi dans son article que les chercheurs et les chercheuses ont toujours conservé des traces de leur travail, avec des pratiques d'archivage plus ou moins sophistiquées et sur des supports matériels très variés (fiches, dossiers, photographies, carnets de notes, disquettes, cartes à jouer, clés USB, *cloud*…). Le moment actuel s'inscrit donc dans le prolongement de pratiques archivistiques déjà profondément ancrées dans nos métiers (Daston, 2017). Il s'en démarque néanmoins par les évolutions techniques et infrastructurelles des dernières décennies (notamment l'étiolement de l'empire du papier en faveur du développement du numérique), ainsi que par l'injonction institutionnelle généralisée à la gestion des données de la recherche. Ce moment singulier qui érige les données en attendus centraux de la recherche (les livrables ou *work packages* des projets décrits par exemple dans l'article de Delphine Cavallo) nécessite toutefois de prendre du recul, pour dénaturaliser le terme et interroger ces données qu'il est devenu si impérieux de gérer et de partager.

## Définir les données ?

L'effervescence qui caractérise le moment actuel suscite en effet des interrogations. On y parle des *données* comme si elles constituaient une évidence. On leur adjoint des préfixes (*méta-*, *para-*) et des verbes (*FAIRiser*, *mettre en banque*, *mettre à disposition*). Un même terme est par ailleurs employé pour désigner des réalités très différentes, qui ne relèvent pas des mêmes enjeux : *données* de la recherche, *données* personnelles, *données* sensibles. L'un des objectifs de ce numéro est donc également de réfléchir à ce que sont ces fameuses données.

Sans prétendre à autre chose que de dresser quelques grandes pistes généalogiques, il est intéressant de retracer l'évolution du mot. Dans un texte intitulé « Data before the fact », Daniel Rosenberg (2013) fait remonter l'apparition dans les traités scientifiques du terme *data* vers 1730, au moment de la traduction vers l'anglais des textes en latin de Bacon et Newton. Le succès est rapide et contemporain de l'avènement de la science moderne, mais il est intéressant de pointer avec Rosenberg le sens initial du terme, qui renvoie, d'une part, aux principes allant de soi (*taken for granted*), qui précèdent tout argument et, d'autre part (en une acception théologique), à ce qui vient de Dieu et n'est donc pas questionnable. Or, cette idée d'immuabilité véhiculée par le terme *données* ne traduit qu'imparfaitement la réalité de la démarche scientifique. Howard Becker s'intéresse

à la question en 1952 dans un article sur les liens entre « science, culture et société » (Becker, 1952, p. 278). Il y reprend notamment une phrase de son éditeur H. E. Jensen (1950, p. 9) : « C'est un malheureux hasard de l'histoire que le mot *datum* (du latin *dare,* "donner") plutôt que *captum* (du latin *capere*, "prendre") en soit venu à incarner l'unité de base du travail scientifique. Car la science ne traite pas de "ce qui a été donné" par la nature au scientifique mais de "ce qui a été pris" ou sélectionné de la nature par le scientifique, en accord avec ses objectifs »[15].

Parmi les variations sur le terme *données/data*, il y a celle, proche, proposée par Bruno Latour (1993, p. 188) selon qui « décidément, on ne devrait jamais parler de "données", mais "d'obtenues" ». L'enjeu n'est pas tant de changer de terme (ce qui aurait sans doute les mêmes effets de naturalisation – on ne se souviendrait plus à force que le nouveau terme vient *d'obtenir*…) que d'en questionner les effets. Plusieurs travaux récents en *Science and Technology Studies* se sont attelés à cette tâche. Dans la postface qu'il a signée pour le volume *Raw Data Is an Oxymoron* (Gitelman, 2013), Geoffrey Bowker (2013) rappelle ainsi que les données ne sont jamais « *raw* » (brutes) mais toujours « *cooked* » (préparées) : elles n'*existent* pas par elles-mêmes mais sont générées, selon des modalités dont les études historiques montrent la variété et la sophistication (Aronova *et al.*, 2017). Jérôme Denis (2018, p. 44) propose quant à lui de questionner l'invisibilité du travail des données, laquelle résulterait de « la production et de la circulation même des données dont la solidité et la valeur semblent étroitement liées à leur capacité à faire oublier le fait même qu'elles ont été travaillées ». Le tournant postmoderne en sciences sociales a aussi contribué à repenser les modes de fabrique des savoirs en montrant que le terrain et les archives ne sont pas un simple moment de collecte de données qui seraient interprétées ultérieurement mais le lieu même, dans l'interaction, de production de la connaissance (Tedlock, 1991). Dès lors, il devient très difficile, voire périlleux, de chercher à séparer, d'une part, ce qui relèverait de l'expérience personnelle de terrain et, d'autre part, ce qui constituerait des données qu'il serait possible de mettre à disposition. Cette position, qui s'est construite au sein de l'anthropologie à partir de séjours prolongés d'observation voire de participation à des activités, peut s'étendre à l'ensemble des pratiques scientifiques, comme le montrent les travaux en études des sciences.

15 Notre traduction de : « It is an unfortunate accident of history that the term *datum* (Latin, "to give") rather than *captum* (Latin, to "take") should have come to symbolize the unit-phenomenon in science. For science deals, not with "that which has been given" by nature to the scientist, but with "that which has been taken" or selected from nature by the scientist in accordance with his purpose ».

Dans son article pour ce numéro, Alexandra Ortiz Caria revient en détail sur le processus de « fabrication » d'un corpus en analyse de conversation d'inspiration ethnométhodologique : elle met en évidence différentes strates de données, depuis les données primaires produites au moyen de dispositifs de captation audiovisuelle jusqu'aux données secondaires qui résultent de la transcription écrite de ces captations. Ces dernières font l'objet de discussions intenses entre membres de la discipline lors de *data sessions*. Ce processus de fabrication collective des données et des savoirs susceptibles d'en résulter vient rappeler que les données ne sont jamais stabilisées et peuvent toujours être soumises à de nombreuses modifications. L'ensemble du processus est tout autant théorique que méthodologique, puisque transcrire une séquence conversationnelle est déjà une opération analytique.

Ces réflexions sur le caractère construit voire analytique des données et de leurs modes de saisie et traduction ne sont pas sans rappeler une querelle philosophique majeure du début du XX^e^ siècle, qui concerne ce que le philosophe américain Wilfrid Sellars a qualifié de *mythe du donné*. Cette querelle a par exemple opposé Bertrand Russell aux pragmatistes John Dewey et George H. Mead (Garetta, 2010), sur la question de savoir s'il pourrait y avoir un fondement au réel (une donnée première en quelque sorte, quelque chose d'intangible et d'initialement « donné ») au-delà de la perception sensible. On saisit alors, notamment avec Jocelyn Benoist (2012), combien cette recherche d'un fondement, et plus largement le geste de séparer l'esprit et le monde, est le produit de l'épistémologie moderne. Une manière de sortir des querelles entre naturel et construit serait de se tourner vers l'approche antidualiste promue par les pragmatistes. Pour Dewey (1993), le schème général de l'enquête caractérise tout autant des activités spécifiques (comme l'activité scientifique bien entendu, à laquelle il consacre plusieurs chapitres) que celles de la vie ordinaire, à partir du principe continuiste selon lequel ces plans n'ont pas à être pensés isolément : l'activité scientifique n'est en quelque sorte qu'une extension de l'attitude que nous déployons au quotidien. Une telle perspective aide à saisir en quoi extraire des données et les mettre à part dans un plan de gestion (ou une infrastructure numérique, un article scientifique dédié – *data paper* –, etc.) revient notamment à rompre ce lien : isoler la thématique des données conduit à extraire un élément de la continuité du monde social, laquelle se maintient en dépit des activités d'un-e sociologue, anthropologue ou historien-ne qui viendrait en prélever quelques fragments.

Ces considérations ne sont pas seulement techniques ou érudites ; elles visent également une efficacité pratique. Là en effet où les débats actuels tendent à naturaliser le terme (donc à ne pas en débattre, à faire comme s'il

était un fondement ou un principe allant de soi), ces rappels (notamment que le donné peut être un mythe) permettent de garder en mémoire l'absence d'évidence des données, alors même qu'elles sont précisément censées faire preuve en soi (authentifier une démarche scientifique, la répliquer, ou permettre de débusquer une supercherie). Ces prises de recul par rapport à la naturalisation du terme permettent de questionner sa place concrète dans les jeux de langage de la recherche. Le plus souvent, un-e anthropologue fait du terrain, elle ou il ne « recueille » pas des « données » ; de même, un-e historien-ne consulte des archives, construit un corpus ; et l'on peut s'interroger sur ce que seraient des données pour un-e philosophe. Dire cela ne revient pas à balayer ce moment « données », mais invite au contraire à le repenser à partir du constat qu'il modifie considérablement la conception du travail scientifique et de ses différents métiers, ainsi que la politique de l'enquête.

Et de fait, l'un des enjeux de ce numéro (et du cycle dans lequel il s'inscrit sur les sciences humaines et sociales au travail) est de favoriser l'expression de la variété de définitions et de pratiques liées aux données de la recherche, et de rappeler que cette expression n'est pas le seul fait des chercheurs et chercheuses mais de toutes celles et ceux qui travaillent au quotidien avec les données. Des archivistes et documentalistes, des juristes, des ingénieur-e-s et technicien-n-es de la recherche donnent à chaque fois une définition riche et dense de ce terme – montrant ainsi qu'il n'est pas réductible à la simplification qu'on peut parfois redouter. L'article de Séverine Janssen présente à ce titre un projet expérimental visant à fabriquer un corpus sonore de la ville de Bruxelles. Il met en évidence un autre modèle de production et d'indexation des données. En rompant avec l'impératif de sélection « entre le valable et le trivial, entre le vrai et le faux », cette base de données sonore prend le parti de tout archiver et de tout mettre à disposition : la collecte sauvage, effectuée aussi bien par des membres du projet que par des citoyen-ne-s ordinaires, produit une masse d'enregistrements éclectiques, dont la sérialité et le brouhaha sont offerts à l'imagination de tou-te-s sans présumer des usages futurs qui pourront en être faits. La donnée, qui s'accompagne d'un appareillage minime et ne fait pas l'objet de luttes de qualification ou d'opérations de standardisation, est alors laissée à l'appréciation de ses utilisateurs.

Mais dans le monde de la recherche, celles et ceux qui sont aujourd'hui chargé-e-s de penser l'archivage et la mise à disposition (non seulement des résultats mais aussi des processus d'enquêtes) se confrontent, dans la pratique, à la question de savoir ce qu'est une donnée. Il s'agit pour elles et eux de définir l'ensemble des éléments nécessaires pour qu'une recherche soit appropriable et exploitable par d'autres. Dans son article, Anne-Laure

Stérin montre la complexité de l'analyse juridique nécessaire pour déterminer le degré d'ouverture possible pour différents types de données. Elle rappelle qu'« on n'ouvre pas les données de la recherche comme on ouvrirait un robinet d'eau ». Les données ne sont pas FAIR a priori, mais doivent être retravaillées. Dans l'entretien que l'équipe de l'INED nous a accordé apparaît en outre l'idée qu'une donnée ne fonctionne pas seule, qu'elle doit être accompagnée d'un appareillage conséquent afin de garantir son interprétation et sa réutilisation. Pour les enquêtes par questionnaire, l'exploitation des données nécessite par exemple toute une documentation – le libellé de la question, le nombre de répondant-e-s, la liste des variables supprimées et conservées, les manuels de formation et les consignes données aux enquêteurs, etc. – permettant de saisir le sens de résultats d'enquêtes. De son côté, dans le second entretien du numéro, l'équipe de beQuali – banque d'archivage d'enquêtes qualitatives[16] – a fait le choix de ne pas utiliser le terme *données* en raison de sa polysémie. Ils préfèrent parler de *matériaux collectés sur le terrain* : c'est l'enquête dans sa globalité qui constitue plutôt l'unité de travail et qu'il s'agit de resituer et de contextualiser en vue de l'archivage. Cela passe par la production d'une enquête sur l'enquête qui présente le contexte de réalisation de la recherche, son cadre théorique, la méthodologie déployée, etc. Tout cela montre bien que pour celles et ceux qui les côtoient au quotidien, les données ne se laissent pas simplement définir.

## Les économies des données

Les débats, controverses, hésitations, tâtonnements et ajustements actuels semblent indiquer qu'un changement de paradigme a bel et bien lieu et que nous nous situons à un moment charnière de l'évolution des politiques de la recherche. Il est dès lors indispensable de s'interroger sur les conséquences que peut avoir l'exigence de partager ses données quant à la manière dont nous exerçons nos métiers et concevons, documentons, archivons et appareillons les activités de recherche. Les textes de ce hors-série proposent quelques pistes de réponse.

Tout d'abord, cette injonction à fournir des données – dans le cadre d'enquêtes financées sur fonds publics, ou pour permettre de contrôler la validité d'une publication – ne va-t-elle pas produire une normalisation de la recherche et une standardisation de ses méthodes d'enquête ? Cela peut en effet conduire à penser la recherche à l'envers, en prévoyant en amont ce

16 Voir [URL : https://bequali.fr/fr/], consulté le 20 novembre 2019.

que seront les données qu'il sera possible de verser pour alimenter les plateformes et se conformer aux exigences du FAIR. Cette anticipation n'encourage pas forcément l'originalité des méthodes ou la prise de risque. Le partage des données nécessite un appareillage qui varie suivant les disciplines et les méthodes d'enquête utilisées. On peut s'interroger sur la possibilité de transposer les exigences d'explicitation des conditions de collecte à des travaux qui s'appuient sur de longues enquêtes de terrain et sur la pratique ethnographique, laquelle repose tant sur des observations de (ou participations à des) situations que sur l'intersubjectivité entre la chercheuse ou le chercheur et ses interlocutrices ou interlocuteurs. Pour fonder le sérieux d'une recherche, il serait ainsi bien plus simple de fournir la transcription d'une série d'entretiens, les résultats bruts d'une enquête statistique que des cahiers de terrain dans lesquels des notes auront été prises à la hâte, et avec des abréviations souvent idiosyncrasiques rendant toute copie (par scan ou photo) inutile : il faudrait en outre se payer le luxe de retranscrire des pages et des pages de ces carnets… Les infrastructures qui sont actuellement en train d'être mises en place ne risquent-elles pas, par leur matérialité, par le type de classement qu'elles sous-tendent de participer de cette uniformisation de la recherche ? On peut également s'interroger sur le destin des données primaires audiovisuelles produites en analyse de conversation d'inspiration ethnométhodologique qui, comme le rappelle Alexandra Ortiz Caria, doivent être soumises à de nombreux traitements d'anonymisation avant d'être partageables.

Le texte de Christelle Rabier montre de son côté que, bien que les pratiques d'archivage soient aussi anciennes que la recherche, le virage actuel concerne avant tout les destinataires de ces archives. Pour reprendre ses termes, l'archive n'est plus uniquement cette « dimension ordinaire du travail intellectuel » fournissant « à la fois la trace, la source et le cadre de l'activité savante », mais se transforme, petit à petit, en une vitrine chargée de prouver et garantir la solidité d'une recherche. Ce risque de normalisation induit par le passage d'un archivage pour soi à une archive pour les autres, est susceptible de n'occasionner qu'un déplacement entre la scène et les coulisses de la recherche. Sommé-e-s de produire des données, des chercheurs et chercheuses ne seraient-elles pas tenté-e-s de contenir les risques de cette exposition en produisant un artefact d'enquête – partagé et diffusé – qui permettrait de conserver cachés les maladresses, les hésitations et autres cafouillages, ou simplement la cuisine interne ? Dans de telles conditions, quel statut accorder à ces données ouvertes ? Et comment les réutiliser ? C'est ce qu'interroge Sophie Duchesne dans son article lorsqu'elle souligne les contradictions inhérentes à l'archivage : outre un intérêt pédagogique certain

pour les étudiant-e-s en sciences sociales, le travail de mise à disposition d'enquêtes qualitatives n'aboutit au final qu'à des formes de réutilisation limitées. En outre, la pratique de la réanalyse d'enquêtes passées, au demeurant passionnante, se satisfait selon elle difficilement des opérations de standardisation auxquelles sont soumises les données lorsqu'elles sont mises en banque.

Bien qu'une grande part des chercheurs et des chercheuses ignore les injonctions actuelles des politiques de la science ouverte, les réformes légales et injonctions financières récentes vont les pousser à devoir désormais tenir compte de l'économie des données dans leurs travaux : calculer les coûts, les risques et les bénéfices de la mise à disposition ; trancher entre ce qu'il est possible de divulguer et ce qu'il est préférable de dissimuler. Comme le rappellent les chercheurs de l'INED, « donner ses données » implique de les préparer, ce qui représente un coût matériel et temporel considérable. Or ce travail supplémentaire ne s'accompagne pas toujours de la reconnaissance escomptée tandis que la mise à disposition des coulisses de leurs recherches les expose à des critiques sur leur manière de travailler. L'entretien mené avec l'équipe de beQuali revient également sur les réticences de certain-e-s chercheurs et chercheuses à ouvrir la boîte noire de leur travail. Se pose aussi la question de jusqu'où aller dans la transparence : outre les données en elles-mêmes, s'agit-il également de partager les différentes strates d'interprétations, de classification, de codage qui ont contribué à leur analyse ?

La mise en place de plateformes et de grandes infrastructures, l'exigence à disposer à l'avance ou dans les six mois d'un plan de gestion des données pour acquérir un financement invitent également à saisir l'expression *économie de la donnée* sous son acception la plus triviale. Le moment « données » implique nécessairement des enjeux financiers qu'il ne faudrait pas occulter. Dans son article, Delphine Cavallo retrace les difficultés méthodologiques, techniques et infrastructurelles qui entourent la rédaction de plans de gestion de données dans les projets européens qu'elle a eu l'occasion d'accompagner. Elle montre que l'opérationnalisation des grands principes du FAIR est loin d'aller de soi dans la pratique, notamment car la capacité à gérer des données en disposant des infrastructures suffisantes crée de nouveaux déséquilibres et rapports de force entre partenaires. L'archivage et la mise à disposition des données ont notamment un coût que les partenaires extra-européens ne peuvent pas toujours supporter, ce qui conduit à un renforcement des inégalités entre systèmes nationaux de recherche.

En outre, dans le contexte de pénurie de budget et de postes de titulaires que connaît le paysage académique français, la réflexion sur les aspects économiques de la mise à disposition des données nous conduit à chercher au détriment de qui et de quoi se font ces nouvelles dépenses. Toute une

gamme de nouvelles tâches apparaît, qui ne s'accompagnent pas nécessairement de la création de postes de titulaires dans les laboratoires ou structures mutualisées comme les Maisons des sciences de l'homme : les opérations de traitement de données se font le plus souvent dans le cadre de contrats à durée limitée, de missions, voire de stages. Répondre aux exigences légales et contractuelles autour des données passe généralement, pour le moment, par une extension de la logique de la précarité et le recours à des contrats courts chargés uniquement de gérer les données. Cela revient en outre à nier la spécificité des métiers et compétences engagés dans le travail des données : la rappeler est au centre des articles de Delphine Cavallo et d'Anne-Laure Stérin, ainsi que des deux entretiens avec beQuali et l'INED. Il ne faudrait pas non plus que ces tâches, encore peu valorisées – comme le soulignent les membres de beQuali qui se qualifient eux-mêmes de « travailleurs de l'ombre » – soient dévolues à celles et ceux, masterant-es, doctorant-es ou post-doctorant-es, qui occupent les places les plus fragiles du monde académique. En cela, ce moment « données » reproduirait, voire risquerait de renforcer les hiérarchies du monde de la recherche qui placent en bas de l'échelle les personnes qui travaillent à la collecte de données et réserve les élaborations théoriques aux postes les plus élevés.

Enfin, il semble bien que l'économie des données doive, dans le monde de la recherche aussi, prendre un sens encore plus trivial. Ici encore, les données semblent pouvoir devenir un bien marchandable, comme l'évoque en conclusion de son article Anne-Laure Stérin. À titre d'exemple, en novembre 2019, le groupe éditorial Elsevier – qui regroupe un grand nombre de revues scientifiques et dont les pratiques commerciales ont déjà fait l'objet de nombreuses controverses dans le monde académique – a proposé une expérimentation aux universités néerlandaises : construire un modèle de revue entièrement en *open access* en échange des (méta)données des scientifiques et de leurs institutions[17]. Si le groupe Elsevier s'intéresse aux (méta)données, c'est sans doute qu'il est possible de les monnayer. Au-delà des politiques scientifiques, les groupes privés s'emparent du lien, qui ne paraissait pas immédiat de prime abord, entre *open science*, *open access* et *open data*. Quand les données deviennent la compensation de la gratuité des publications, cela redéfinit pourtant le partage. Il n'est plus question d'ouvrir les données à une communauté scientifique mais de les mettre au service des intérêts d'un groupe privé.

17 Voir [URL : https://www.scienceguide.nl/2019/11/elsevier-biedt-100-open-access-in-ruil-voor-metadata/], consulté le 23 novembre 2019.

La prolifération d'événements autour des données de la recherche ne se fait donc pas sans débats et controverses. Malgré les efforts de réflexivité et la formulation de critiques, les infrastructures de données s'installent et les pratiques qu'elles induisent s'implantent. Les nombreuses questions et inquiétudes que suscite ce moment « données » expliquent sans doute la forme de certaines contributions réunies dans ce numéro (des prises de position, des mises en garde) tant nous manquons encore de recul pour analyser de façon systématique les évolutions occasionnées par cette considération nouvelle pour les données. Espérons que ce numéro sera suivi de nombreux prolongements, notamment pour mettre à l'épreuve l'hypothèse selon laquelle ce moment « données » n'est pas circonscrit au monde de la recherche mais concerne plus largement la question de la conservation des traces des activités humaines, laquelle implique la constitution d'une (potentiellement) infinie mémoire numérique en réseau. Enfin (ou hélas) la bibliothèque de Babel imaginée par Jorge Luis Borgès ?

## Bibliographie


Aronova Elena, Oertzen Christine (von) et Sepkoski David éd., 2017, numéro thématique « Data histories », *Osiris*, vol. 32, n° 1.

Aust Jérôme, 2014, « Financer la recherche sur projet. Figures historiques d'un dispositif de gouvernement », *Genèses*, vol. 94, n° 94, p. 2-6.

Becker Howard, 1952, « Science, culture, and society », *Philosophy of Science*, vol. 19, n° 4, p. 273-287.

Benoist Jocelyn, 2012, « Mythe du donné, mythe de la pensée », *Les études philosophiques*, vol. 103, n° 4, p. 515-531.

Bowker Geoffrey C., 2013, « Data flakes : an afterword to *Raw Data Is an Oxymoron* », *Raw Data Is an Oxymoron*, G. Lisa éd., Cambridge, The MIT Press, p. 167-171.

Damerdji Amina, Hayat Samuel, La Valle Natalia, Pecqueux Anthony et Rabier Christelle, 2018, « Éditorial. Le savoir-faire des revues », *Tracés. Revue de Sciences humaines*, hors-série, p. 11-24.

Daston Lorraine éd., 2017, *Science in the Archives : Pasts, Presents, Futures*, Chicago, The University of Chicago Press.

Denis Jérôme, 2018, *Le travail invisible des données*, Paris, Presses des Mines.

Desquilbet Loïc, Granger Sabrina, Hejblum Boris, Legrand Arnaud, Pernot Pascal *et al.*, 2019, *Vers une recherche reproductible. Faire évoluer ses pratiques* [en ligne], Unité régionale de formation à l'information scientifique et technique de Bordeaux, [URL : https://hal.archives-ouvertes.fr/hal-02144142v3/document], consulté le 20 novembre 2019.

Dewey John, 1993, *Logique. La théorie de l'enquête*, Paris, Presses universitaires de France.

Garreta Guillaume, 2011, « Le donné est-il un mythe ? Données sensibles et données de l'enquête », *Les données de l'enquête*, B. Michel, L. Sandra et O. Barbara éd., Lille, Presses universitaires du Septentrion, p. 37-52.

Giry Johan et Schultz Émilien, 2017, « L'ANR en ph(r)ase critique. Figures et déterminants de la critique d'un dispositif de financement », *Zilsel*, vol. 2, nº 2, p. 63-96.

Gitelman Lisa éd., 2013, *Raw Data Is an Oxymoron*, Cambridge, The MIT Press.

Hubert Matthieu et Louvel Séverine, 2012, « Le financement sur projet : quelles conséquences sur le travail des chercheurs ? », *Mouvements*, nº 71, p. 13-24.

Jarrige François, Le Courant Stefan et Paloque-Bergès Camille, 2018, « Éditorial. Infrastructures, techniques et politiques », *Tracés. Revue de Sciences humaines*, nº 35, p. 7-26.

Jensen H.E., 1950, « Editorial note », *Through Values to Social Interpretation*, H. Becker éd., Durham, Duke University Press, p. 7-11.

Latour Bruno, 1993, « Le "pédofil" de Boa-Vista : montage photo-philosophique », *La clef de Berlin. Petites leçons de sociologie des sciences*, Paris, La Découverte, p. 171-225.

Leighley Jan, 2019, « Celebrating verification, replication, and qualitative research methods at the AJPS » [en ligne], [URL : https://ajps.org/2019/03/20/celebrating-verification-replication-and-qualitative-research-methods-at-the-ajps%ef%bb%bf/], consulté le 20 novembre 2019.

Rosenberg Daniel, 2013, « Data before the fact », *Raw Data Is an Oxymoron*, G. Lisa éd., Cambridge, The MIT Press, p. 15-40.

Tedlock Barbara, 1991, « From participant observation to the observation of participation : the emergence of narrative ethnography », *Journal of Anthropological Research*, vol. 47, nº 1, p. 69-94.