L'accès des chercheurs aux données d'enquête à propos de handicap dans le contexte français : enjeux légaux et conditions pratiques

Sébastien Oliveau, Lorraine Adam

L'accès des chercheurs aux données d'enquête à propos de handicap dans le contexte français : enjeux légaux et conditions pratiques

Sébastien Oliveau
Progedo, CNRS, UAR 2506
Université d'Aix-Marseille, CNRS, Sciences Po Aix, UMR 7064 Mesopolhis

Lorraine Adam
Progedo, CNRS, UAR 2506

Introduction

Le handicap, sujet de recherche en soi, est aussi une question transversale à de nombreux thèmes étudiés dans le cadre des études de population aujourd'hui, que l'on pense bien sûr au vieillissement ou aux transports, mais aussi à l'économie ou aux divisions sociales (sur ce sujet, voir par exemple la récente thèse de Célia Bouchet, 2022). Les données statistiques sur le sujet sont assez variées, et la DREES[1] a récemment proposé un recensement détaillé des sources disponibles (Bellamy et Ricroch, 2023) indiquant aussi les questions en lien avec le sujet présentes dans chaque enquête ou base de données, ainsi que les liens vers les données (https://drees.solidarites-sante.gouv.fr/ressources-et-methodes/les-donnees-statistiques-sur-le-handicap-et-lautonomie). On constate alors rapidement la variété des sources, mais aussi la variété des diffuseurs.

A titre d'exemple, en octobre 2023, la recherche sur le mot clef « handicap » dans le catalogue du dispositif Quetelet-Progedo-Diffusion (accessible à l'adresse https://data.progedo.fr) renvoie 36 séries, 231 études (jeux de données) et 1283 variables. La requête sur le terme « autonomie » propose 19 séries, 99 études et 176 variables, issus d'une quinzaine de producteurs : l'INSEE et la DREES bien entendu, mais aussi le CEREMA, le CEREQ ou la DGAFP, par exemple.

Notre papier se propose de revenir sur cette variété de sources et de diffuseurs en essayant d'expliquer leur nombre. En effet, si les données produites sur le sujet relèvent a priori des données dites de santé, dont les accès sont réglementés par la loi, cette réglementation s'ajoute à celle sur la protection des données personnelles. C'est pourquoi il nous a semblé pertinent de revenir sur les cadres légaux d'accès à ces données, mais aussi sur les conditions pratiques de leur utilisation.

[1] « Direction de la recherche, des études de l'évaluation et des statistiques » du ministère des solidarités et de la santé (https://drees.solidarites-sante.gouv.fr/).

I) Matériaux, données, données de recherche, données de santé : ouverture et protection

a. La difficile question de la définition

Comme nous le rappelle Agnès Robin en introduction de son ouvrage consacré aux droits des données de la recherche (Robin, 2022), la définition de ce qu'est une donnée pose problème, en tout cas question. D'un point de vue juridique, mais plus largement du point de vue scientifique, il n'existe pas de véritable consensus concernant la définition d'une donnée. En revenant dans le temps, on voit que le sens de ce terme a fortement évolué. Pendant longtemps, c'est-à-dire du 18ème jusqu'au milieu du 20ème siècle, la donnée était avant tout un nombre ou un élément de description. C'est avec l'avènement de l'informatique, et avec son rôle grandissant au sein des sociétés, que le sens de donnée, dans son acception commune actuelle, s'est imposée. La donnée décrit désormais l'élément de base d'une connaissance formalisée. Néanmoins, le terme reste « peu explicite » (Stérin et Noûs, 2019) et si l'on veut être correct, l'utilisation du mot donnée devrait en fait toujours être accompagné d'un ou plusieurs adjectifs qui en précisent le sens : donnée statistique, donnée informatique, donnée textuelle, etc.

D'autre part, la professionnalisation (on pourrait dire l'industrialisation) des processus de recherche depuis une vingtaine d'année, accompagné par le développement de la numérisation et de l'informatisation de la science, a conduit la donnée a joué un rôle toujours plus grand. Il est aujourd'hui difficile de penser une science sans donnée. Pourtant, ce n'est pas la donnée qui définit la science, celle-ci reposant plutôt sur la pensée et la méthode. Néanmoins, force est de constater le rôle croissance pris par la donnée, à tel point que chacun souhaite s'inclure dans ce que l'on pourrait qualifier de mode. Nous rejoignons quant à nous Galonnier *et al.* (2019) quand elles rappellent que « 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. ».

Il semble de ce point de vue utile de distinguer ce qui constitue des « matériaux » de la recherche (ce que l'on récolte : objets ou informations) et les « données » qui sont le résultat d'une médiation par le chercheur entre ces matériaux à qui il donne du sens et des éléments qu'il fournit à ses pairs et à la société. C'est le sens étymologique du mot « donnée », issu du latin *data*, terme que l'on retrouve d'ailleurs en anglais et en allemand. Cela permet de rappeler tout ce que la donnée doit à la construction humaine. Elle n'est pas donnée en soi, elle est donnée par l'intermédiaire d'une personne.

On reprendra à notre compte le titre de l'ouvrage de Gitelman (2013) « les données brutes sont un oxymore »[2], au sens où elles ont toujours été préparées.

Ces réflexions qui concernent les données de la recherche sont valables de manière plus générale. Les données de la recherche se caractérisent en effet d'abord par leur origine : la médiation, d'un scientifique ou autre. En outre, les chercheurs mobilisent aussi (notamment en sciences sociales), et de plus en plus, des données qu'ils n'ont pas produites eux-mêmes, mais qui, une fois mobilisées et médiatisées, peuvent devenir des données de la recherche. D'ailleurs, il est intéressant de se questionner sur l'intérêt de définir des données de la recherche. La pertinence de définir une donnée comme issue de la recherche, hormis pour poser les contours des droits qui s'appliquent[3], nous semble être la valeur (véracité) qu'on lui attribue.

b. Le cadre légal

Après avoir rappelé le champ de ce que peut-être une donnée, il est nécessaire de comprendre dans quel cadre juridique on évolue, car c'est lui qui va déterminer la manière dont on doit et dont on peut s'en saisir.

Le cadre de la donnée statistique publique est ancien, puisqu'il est institué par la loi de finances du 27 avril 1946 qui acte la création de l'INSEE et par la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques (qui acte aussi la création du comité du secret statistique dont nous reparlerons plus bas). Mais c'est seulement avec la loi relative à l'informatique, aux fichiers et aux libertés (loi n° 78-17 du 6 janvier 1978) que le législateur aborde les données de manière plus large[4]. En outre, ce cadre juridique est renforcé par le règlement de l'union européenne n°2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (connu sous l'acronyme RGPD). Celui-ci est entré en application le 25 mai 2018.

Cet ensemble juridique permet de définir ce qu'il est possible de faire ou non à partir de données personnelles[5], en précisant notamment les exceptions qui autorisent les chercheurs à un meilleur accès à ces données. Dans ce cadre, la question des données de santé renvoie vers le code de la santé publique et particulièrement l'article L1460-1 et suivants, relatifs à la mise à disposition des données de santé. Ceux-ci définissent notamment la nécessaire protection de ces données mais aussi l'usage qui peut être fait de ces données spécifiques qui traite de la santé humaine.

---

[2] « *Raw data is an oxymoron* »

[3] Une des caractéristiques juridiques des données de la recherche est que les droits sont partagés entre leurs auteurs (les chercheurs) et leur producteur (l'institution qui l'emploi).

[4] C'est à cette occasion qu'est créée la Commission Nationale de l'Informatique et des Libertés (CNIL - https://cnil.fr/fr).

[5] Une donnée à caractère personnelle est « toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement » (https://cnil.fr/fr/cnil-direct/question/une-donnee-caractere-personnel-cest-quoi)

Enfin, un cadre juridique en partie concurrent a vu le jour en 2016, à travers la loi dite « pour une république numérique », (la loi n° 2016-1321 du 7 octobre 2016) qui impose l'ouverture des données. C'est donc dans ce contexte que les chercheurs évoluent aujourd'hui : entre ouverture (obligatoire) des données et (nécessaire) protection de la donnée personnelle, protection encore renforcée lorsqu'il s'agit de données de santé.

II) Les données à propos du handicap : une simple donnée de santé ?

Le chercheur qui travaille sur le handicap et mobilise, ou produit, des données sera donc confronté à cet ensemble de textes juridiques. Nous pouvons rappeler ici que le handicap est considéré comme relatif à la santé et que par conséquent, l'ensemble des textes présentés précédemment s'appliquent. Si le cadre légal a été clarifié, il convient néanmoins aujourd'hui de rappeler aussi comment ces données s'articulent dans le cadre des catégories en usage dans la recherche.

a. La place des données sur le handicap dans les données

On peut qualifier les données de cinq façons différentes, en fonction de leur structure ou de leur objet. Le premier niveau est celui de ce que l'on nommera les fichiers standards, ou données ouvertes (*public files* ou *open data*). Ces données ne concernent pas les humains, ou alors de manière parfaitement anonyme (données personnelles suffisamment agrégées pour ne plus pouvoir identifier de sous ensemble par exemple).

Il existe ensuite un second niveau, communément appelé données pseudonymisées (souvent rencontrées dans la littérature de langue anglaise sous le terme de *scientific use files*). Ces données sont qualifiées d'indirectement identifiantes. Seules, elles restent anonymes, mais recoupées avec d'autre base de données, elles pourraient potentiellement amenées à une identification. Paul Ohm (2009) rappelle qu'une stricte anonymisation est impossible sauf au prix d'une perte d'information qui rend les données inutiles[6]. Ces fichiers constituent donc une sorte d'entre-deux qui sont réservés à la recherche et aux producteurs de données. En France, elles sont définies par le comité du secret statistique[7] comme « fichier production recherche »[8].

[6] Pour reprendre son expression : « les données peuvent être utiles ou parfaitement anonymes, mais pas les deux », traduction libre de « *Data can be either useful or perfectly anonymous but never both* ».
[7] Le comité du secret statistique (https://www.comite-du-secret.fr/) est une institution dont les compétences ont été fixées par la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. C'est elle qui régule l'accès aux données individuelles collectées par voie d'enquête statistique ou transmises au service statistique public.
[8] Les fichiers FPR sont également soumis au secret statistique, mais bénéficient d'une procédure d'accès simplifiée.

Viennent ensuite les données confidentielles (*secure use files*), qui sont directement identifiantes. Cela ne signifie pas que l'identité des personnes est indiquée en clair, mais que les informations contenues dans les fichiers permettent d'identifier les personnes. Le papier de référence sur la question est celui de Latanya Sweeney[9] (2000b) qui montre que grâce à trois informations contenues dans les fichiers du recensement américain (code postal, date de naissance et genre), il est possible d'identifier 87% des personnes[10].

Au sein des données confidentielles, le droit français, et avec lui les chercheurs, va encore distinguer les données dites sensibles[11], qui nécessitent une attention et un traitement particuliers. Il s'agit de données traitant par exemple de la religion, des orientations politiques, des revenus ou encore de la santé. Comme nous l'avons mentionné plus haut, les données personnelles de santé constituent enfin une 5ème catégorie de données, régies par des textes législatifs spécifiques.

Le règlement européen sur la protection des données personnelles (RGPD) procède à une définition large des données de santé. De ce point de vue, les données sur le handicap sont bien des données de santé et sont donc régies par les mêmes règles. Pour mémoire, « constitue un handicap toute limitation d'activité ou restriction de participation à la vie en société subie dans son environnement par une personne en raison d'une altération substantielle, durable ou définitive d'une ou plusieurs fonctions physiques, sensorielles, mentales, cognitives ou psychiques, d'un polyhandicap ou d'un trouble de santé invalidant » (art. L. 114 du code de l'action sociale et des familles).

b. Quelques exemples pour essayer de voir plus clair

Une fois le cadre posé, il peut être intéressant de se pencher sur quelques exemples pour essayer d'illustrer la variété des données dont les chercheurs peuvent disposer pour travailler.

Le premier exemple sera celui des données de l'enquête européenne SHARE (https://www.share-eric.eu). Anonymisées, les données SHARE sont des données ouvertes. Ce sont certes des données de santé, mais elles ne constituent pas des données personnelles, au sens où l'on ne peut pas identifier les personnes interrogées. Les données de l'enquête CARE (https://data.progedo.fr/series/adisp/enquete-capacites-aides-et-ressources-des-seniors-care), produites par la Drees sont des données pseudonymisées. Leur accès est donc régi par le comité du secret statistique. L'enquête RPCH (« Remontées individuelles Prestation Compensation du Handicap »), produite aussi par la Drees est quant à elle une enquête considérée comme comportant des données confidentielles. Son accès est

[9] On renverra aussi à son autre working paper intitulé « *Simple Demographics Often Identify People Uniquely* » (Sweeney 2000a)

[10] Les données de l'article ont été reprises par Golle (2006) qui n'a identifié que 61% des personnes (cela reste très important). A cette occasion l'auteur a testé le recensement de 2000 et a identifié 63% des personnes.

[11] Les données sensibles sont « les informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale. » (https://cnil.fr/fr/cnil-direct/question/1823 sur la base de l'article 9 du RGPD)

également régi par le comité du secret statistique et le chercheur pourra y accéder via le centre d'accès sécurisé aux données (CASD, voir infra). Les données de l'enquête « Trajectoires et origines, enquête sur la diversité des populations en France » (https://data.progedo.fr/studies/doi/10.13144/lil-0494) constitue des données sensibles, puisqu'elle porte notamment sur les origines et la religion des personnes. On y trouve plusieurs questions liées au handicap. Enfin, les données de l'assurance maladie, comme la consommation de soins, constituent et de santé et entrent donc dans le cadre de la plateforme des données de santé[12], dite *Health Data Hub* présentée ci-dessous.

Toutes comportent des informations concernant la santé et le handicap, mais selon leur degré d'anonymisation, des restrictions différentes s'appliquent.

III) L'accès aux données

De cette variété de situation pour définir les données découle assez logiquement des cadres différents pour y accéder. Il n'y a pas, par conséquent, de point d'accès unique pour obtenir des données. On peut évidemment le regretter et souhaiter que l'on aille vers un guichet unique de la donnée[13]. Néanmoins, il ne faut pas négliger les contraintes techniques d'une part, et les conditions pratiques d'autre part. Il reste alors plus efficace aujourd'hui d'avoir plusieurs guichets. On en distingue quatre principaux que l'on se propose de présenter ci-après.

a. La logique de Quetelet-Progedo-Diffusion

Le premier guichet est celui proposé par l'infrastructure de recherche Progedo. Connu sous le nom de Quetelet-Progedo-Diffusion (https://data.progedo.fr), ce portail a connu plusieurs avatars depuis sa création dans les années 2000 (Oliveau 2023). Regroupant les données diffusées par l'Adisp (voir infra) et par l'Ined (Institut national d'études démographiques https://www.ined.fr/), on y trouve notamment des données sur le handicap. Il s'agit en grande majorité de données issues de la statistique publique[14], ainsi que de données d'enquête de chercheurs. Les fichiers diffusés par Quetelet-Progedo-Diffusion sont régis par l'avis du comité du secret statistique du 14 décembre 2018 relatif à l'accès aux

[12] Le code de la santé publique, dans son article L1462-1 (https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000038886833) précise le rôle de la « plateforme de données de santé ».

[13] C'est l'objectif à terme du portail des données en santé « FReSH » : https://www.ouvrirlascience.fr/portail-des-etudes-individuelles-en-sante/

[14] Pour mémoire, selon l'article 1 de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques : « Le service statistique public comprend l'Institut national de la statistique et des études économiques et les services statistiques ministériels. Les statistiques publiques regroupent l'ensemble des productions issues : des enquêtes statistiques dont la liste est arrêtée chaque année par un arrêté du ministre chargé de l'économie ; de l'exploitation, à des fins d'information générale, de données collectées par des administrations, des organismes publics ou des organismes privés chargés d'une mission de service public. La conception, la production et la diffusion des statistiques publiques sont effectuées en toute indépendance professionnelle. »

fichiers de production recherche (FPR)[15]. Cela signifie concrètement que pour demander des fichiers les chercheurs doivent préalablement obtenir l'aval du comité du secret statistique. La procédure est gérée par Progedo via son service des Archives de Données Issues de la Statistiques Publique (http://www.progedo-adisp.fr/).

Les fichiers proposés sont ce que l'on appelle des « fichiers production recherche » (FPR). Il s'agit de versions pseudonymisées des fichiers confidentiels originaux. A ce titre, et comme précisé précédemment, ils ne peuvent pas être laissé en libre accès. C'est pourquoi les chercheurs doivent s'engager à n'utiliser ces fichiers que dans le cadre de leurs recherches, à ne pas les transmettre à qui que ce soit (chaque utilisateur doit s'inscrire et s'engager de la même façon), et enfin à les détruire à la fin de leur utilisation. Des sanctions pénales sont d'ailleurs prévues (amendes et emprisonnement) en cas de mésusage avéré, comme l'indique le document que les chercheurs doivent signer avant de télécharger les données.

Encart 1 : Du rapport « Silberman » à Quetelet-Progedo-Diffusion

En 1999, le rapport « les Sciences sociales et leurs données » remis au ministre de l'éducation nationale et de la technologie proposait la création d'un institut pour la diffusion des données en sciences sociales (Silberman 1999). En 2001, le Comité de Concertation pour les Données en Sciences Humaines et Sociales (CCDSHS) est créé par décret (décret n° 2001-139 du 12 février 2001[16]). Le CDDSHS a pour mission de définir une politique des données pour les sciences sociales. Le 1er juillet 2001, le CNRS en partenariat avec l'EHESS, l'INED et l'université de Caen crée le « centre Quetelet » (Unité Mixte de Service T2419 « Centre d'archivage et de diffusion des données en sciences humaines et sociales - Quetelet »), qui participe à la création en 2005 du réseau Quetelet (sous forme d'un Groupement d'Intérêt Scientifique). Le réseau regroupe le service d'enquêtes de l'INED, le Centre de Données Sociopolitiques de Sciences Po (CDSP) (héritier de la Banque de données sociopolitiques établie à l'IEP de Grenoble) et l'équipe des Archives de Données Issues de la Statistique Publique (ADISP)[17] (Caporali et al., 2015).

Une simplification du dispositif s'opère graduellement au cours de la décennie 2010. La feuille de route des infrastructures de recherche française (2008)[18] prévoit 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

[15] L'avis, dont les annexes sont régulièrement modifiées, est disponible depuis cette page : https://www.comite-du-secret.fr/procedure-fr/fichiers-de-production-et-de-recherche-fpr/
[16] Le décret n° 2015-1469 du 13 novembre 2015 supprime le CCDSHS.
[17] L'ADISP hérite du fonds de données élaboré au sein du Laboratoire d'Analyse Secondaire et de Méthodes Appliquées à la Sociologie (LASMAS). Le LASMAS est devenu le Centre Maurice Halbwachs en 2004
[18] Ministère de l'enseignement supérieur et de la recherche, 2008, p. XIII

Quetelet-PROGEDO. En 2018, Quetelet PROGEDO se transforme en Unité de Service et de Recherche et se nomme simplement « PROGEDO ». Elle remplit le rôle autrefois dévolu au CCDSHS, à savoir la définition d'une politique des données pour les sciences humaines et sociales, ainsi que la mission du réseau Quetelet (diffusion des données) à travers le portail « Quetelet-PROGEDO-Diffusion ». Elle a en outre la charge de l'archivage de données quantitatives via l'équipe de l'ADISP[19].

b. Le Health Data Hub : un accès dédié aux données de santé

Le Health Data Hub (HDH) se veut un guichet unique pour les données personnelles de santé françaises. Le HDH sert aussi de secrétariat au Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé (CESREES : https://www.health-data-hub.fr/cesrees). C'est la loi n° 2019-774 du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé (https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000038821260) qui instaure l'ensemble de ces dispositifs par l'ajout de l'article L. 1462-1 au code de la santé publique.

Les données individuelles de santé non anonymes étant strictement protégées, leur accès se fait sous le contrôle d'une institution dédiée, le CEESRES, qui expose en ligne la procédure : https://www.health-data-hub.fr/starter-kit. La procédure se fait en ligne et est constituée d'un dossier explicitant les objectifs de l'étude, son intérêt, son protocole, ses besoins en données, ainsi que diverses pièces administratives. Le CEESRES rend ensuite un avis sur lequel la Commission Nationale de l'Informatique et des Libertés[20] s'appuie pour donner accès ou non aux données.

c. France Cohortes : une nouvelle infrastructure pour les cohortes de santé

Les cohortes médicales constituent un ensemble de données riches d'enseignement pour la recherche. Onéreuses à mettre en place et à suivre, les cohortes sont l'œuvre de collectifs œuvrant au sein de grandes institutions de recherche, comme l'INSERM par exemple. Construites et administrées par des chercheurs au sein de leurs laboratoires, leur accès est jusqu'à aujourd'hui dispersé. L'objectif de l'Infrastructure France Cohortes (https://francecohortes.org/) est d'offrir un accès unifié et si possible simplifié à ces cohortes et d'accompagner les chercheurs. Créé en 2021, France cohortes regroupe aujourd'hui 13 cohortes (soit environ 600.000 personnes suivies).

[19] La mission de PROGEDO est bien plus large, puisqu'elle comprend aussi la prise en charge des volets français des grandes enquêtes internationales et plus globalement le soutien aux enquêtes en sciences sociales et la valorisation de l'ensemble de ces corpus de données.
[20] Voir note 4

d. Le centre d'accès sécurisé aux données, la solution la plus aboutie

Le Centre d'Accès Sécurisé aux Données (CASD) est aujourd'hui la solution technique la plus aboutie pour donner accès à des données confidentielles tout en préservant la sécurité des données. Il s'agit d'un dispositif informatique constitué d'une « bulle » à laquelle le chercheur accède à distance, via un terminal sécurisé (la « SD Box »). Le CASD permet d'accéder à toutes les données possibles, après accord du producteur et éventuellement des gardiens administratifs (comité du secret statistique ou CESREES par exemple). Après cet accord, le chercheur suit un enrôlement de deux heures à distance, puis se rend dans les locaux du CASD pour une prise d'empreintes digitales. Celles-ci sont enregistrées sur une carte à puce qui sera nécessaire à l'identifications sur le boîtier sécurisé (SD-Box).

La SD-Box doit être situé dans un établissement de recherche, au sein d'une pièce dédiée, et protégée des regards indiscrets. A ce prix (et celui de la location du matériel[21]), le chercheur pourra accéder aux données confidentielles depuis son bureau et éventuellement les croiser avec des données issues d'autres sources. Les informations qui sortent de la bulle à destination du chercheur (textes, illustrations et données retravaillées) sont toutes vérifiées par un personnel du CASD chargé de vérifier la conformité légale (respect du secret statistique).

On comprend que cette solution ouvre des pistes de recherches intéressantes, qui n'étaient pas possibles auparavant (et ne le sont d'ailleurs toujours pas en dehors de ce dispositif) comme, par exemple, l'appariement de données de revenus et de santé.

Encart 2 : la démarche auprès du CASD

La démarche pour accéder à des données via le CASD se fait en trois temps.

1. Il faut d'abord définir son projet de recherche et identifier les sources de données souhaitées, en contactant si besoin le(s) producteur(s) de données.

2. Un dossier d'autorisation est ensuite à déposer en ligne sur le site du Confidential Data Access Portal (CDAP - https://cdap.casd.eu/). Il sera alors étudié par le comité du secret statistique (cf. infra).

3. Une fois les autorisations obtenues, un enrôlement a lieu et un contrat est passé avec le CASD.

Conclusion

La France dispose de cadres légaux contraignants, mais aussi de solutions innovantes, pour accéder aux données. A ce titre, la tension entre la protection des personnes et l'accès des chercheurs semble

[21] Le CASD est un Groupement d'intérêt Public (GIP). Si l'accès aux données publiques reste gratuit, le dispositif d'accès est quant à lui onéreux. Il faut compter environ 4.000€ par an, tarif variant selon la puissance de calcul demandé, le nombre de projets et de chercheurs impliqués. Il est en partie subventionné par Progedo.

résolue de façon bien équilibrée. Néanmoins, la variété du cadre législatif, les définitions parfois concurrentes des données, et la multiplicité des acteurs en présence, peuvent dérouter les chercheurs.

Il existe un grand nombre de ressources en ligne[22] pour s'informer et pour se former à ces questions d'accès aux données. On pense en particulier au site de la CNIL (https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante) ou à celui des délégués à la protection des données des établissements d'enseignement supérieur et de recherche « SupDPO » (https://supdpo.fr/publications/#GTRecherche).

C'est pour cette raison que les communautés de recherche en SHS ont très tôt mis en place des dispositifs d'accompagnement. La première plateforme universitaire de données (PUD) a vu le jour en 2003 à Lille (Oliveau et al., 2020). Elle a ensuite été suivie par d'autres partout en France. On compte actuellement 17 PUD qui maillent le territoire, au plus près des étudiants et des chercheurs. Elles ont vocation à faire connaître les données et leurs modalités d'accès et à former les usagers à leur utilisation (Oliveau et al., 2020). Les ateliers de la donnée apparus à partir de 2022 à l'initiative de recherche.data.gouv (https://recherche.data.gouv.fr/fr/page/ateliers-de-la-donnee-des-services-generalistes-sur-tout-le-territoire), forment un réseau plus généraliste, qui associe les PUD lorsqu'elles partagent le même site, auquel on peut se référer si l'on a besoin d'être accompagné dans la recherche de données.

On ne peut que conseiller aux chercheurs de prendre contact avec des collègues ayant déjà accédeé aux données visées, qui seront d'un conseil utile[23]. De même, des chercheurs siègent au comité du secret statistique (https://www.comite-du-secret.fr/comite/composition-du-comite-du-secret-statistique) et au CESREES (https://www.health-data-hub.fr/cesrees). Ils peuvent aider à savoir ce qui est attendu pour la construction des dossiers.

On l'aura compris, si les données n'ont certainement jamais été aussi exposées qu'aujourd'hui, elles sont aussi très bien protégées. Leur accès reste parfois difficile, mais les dispositifs de mise à disposition et d'accompagnement se multiplient depuis plusieurs années pour en faciliter l'accès.

[22] On consultera avec intérêt les sites du Comité National Consultatif d'Ethique pour les sciences de la vie et de la santé (https://www.ccne-ethique.fr/fr/themes/recherche-legislation-deontologie/recherche), du Comité d'éthique du CNRS (COMETS https://comite-ethique.cnrs.fr/ + https://hal.science/COMETS) et du Comité d'éthique de l'INSERM (https://www.inserm.fr/ethique/comite-dethique-de-linserm/ + https://www.hal.inserm.fr/CEI/).

[23] Voir la liste des projets rendus possibles par le CASD : https://www.casd.eu/projets-developpes-sur-le-casd/ et le HDH : https://www.health-data-hub.fr/projets-laureats

Bibliographie

BELLAMY, Vanessa, **RICROH** Layla, 2023, « Apport de la statistique publique dans la connaissance du handicap ». In CUDEP, Handicaps et autonomies, actes du 19ème colloque national de démographie.

BOUCHET, Célia, 2022, « Handicap et destinées sociales : une enquête par méthodes mixtes ». Phdthesis, Institut d'études politiques de paris - Sciences Po. https://theses.hal.science/tel-03637654.

CAPORALI Arianna, **MORISSET** Amandine, **LEGLEYE** Stéphane, Camille **RICHOU**, 2015, « La mise à disposition des enquêtes quantitatives en sciences sociales : l'exemple de l'Ined », Population-F, 70 (3):, p. 567‑ 97.

GALONNIER Juliette, **LE COURANT** Stefan**, PECQUEUX** Anthony, Camille **NOÛS**.,2019, « Ouvrir les données de la recherche ? » Tracés. Revue de Sciences humaines19, p. 17‑ 33.

GOLLE,Philippe 2006, « Revisiting the uniqueness of simple demographics in the US population ». In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, 77‑ 80. WPES '06. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1179601.1179615.

OHM Paul, 2009, « Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization ». SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=1450006.

OLIVEAU Sébastien, 2023, « Le partage des données quantitatives en SHS : enjeux scientifiques et éthiques, conditions matérielles ». In *Communication scientifique et science ouverte. Opportunités, tensions et paradoxes.*, Annaïg Mahé, Ingrid Mayeur, Elsa Pourpardin, et Camille Prime-Claverie (éds), Louvain-la-Neuve ,De Boeck, p. 13‑ 25.

OLIVEAU Sébastien, **BLÖSS-WIDMER** Isabelle, **DOIGNON** Yoann, **DE BELSUNCE** Clément**,** 2020, « Aix-Marseille University SSH data platforms: Skills to support research in social sciences and humanities (SSH) in the Mediterranean ». Egypte/Monde arabe, 22 (2), p. 95‑ 105. https://doi.org/10.4000/ema.13176.

ROBIN Agnès, 2022, *Droit des données de la recherche. Science ouverte, innovation, données publiques*, Bruxelles, Larcier, Création Information Communication.

SILBERMAN Roxane 1999 « Les Sciences sociales et leurs données : rapport au ministre de l'éducation nationale et de la technologie ». Rapport public. http://www.ladocumentationfrancaise.fr/rapports-publics/004000935/index.shtml.

STERIN Anne-Laure, **NOÛS** Camille, 2019.,« Ouverture des données de la recherche : les mutations juridiques récentes », Tracés. Revue de Sciences humaines, 19, 37‑ 50. https://doi.org/10.4000/traces.10603.

SWEENEY Latanya, 2000a, « Simple Demographics Often Identify People Uniquely ». *Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh*, 34.

———. 2000b, « Uniqueness of simple demographics in the U.S. Population », *Laboratory for Int'l Data Privacy, Working Paper LIDAP-WP4*. https://cir.nii.ac.jp/crid/1573950400321397120.