ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

Propositions de recommandations pour la mise à disposition des données d'enquêtes quantitatives en sciences sociales portant sur des personnes

Note présentée au Conseil scientifique du 1er avril 2021

Henri Leridon, référent intégrité scientifique

Les propositions qui suivent sont destinées à mettre en application les textes et recommandations sur les impératifs de la « science ouverte » relatifs aux données, plus particulièrement les données issues d'enquêtes quantitatives en sciences sociales auprès de personnes physiques.

A/ Dispositions réglementaires et recommandations

La référence juridique essentielle est la ***loi n° 2016-1321 du 7 octobre 2016 pour une République numérique***, qui indique en son article 30 :

> « *Dès lors que les données issues d'une activité de recherche* financée au moins pour moitié par des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne *ne sont pas protégées par un droit spécifique* ou une réglementation particulière et qu'elles *ont été rendues publiques par le chercheur*, l'établissement ou l'organisme de recherche, *leur réutilisation est libre.* »

Le principe d'une ouverture des données est donc clairement énoncé, mais ses modalités ne sont pas précisées. En indiquant que l'accès est libre dès lors « qu'elles [les données] ont été rendues publiques par le chercheur » la loi laisse toute latitude au chercheur ou à l'établissement **pour décider du moment où les données deviennent publiques.**

Une autre contrainte est à considérer, résultant du ***Règlement général pour la protection des données (RGPD, règlement européen du 27 avril 2016***). Celui-ci donne aux personnes dont des données personnelles ont été collectées des droits importants sur ces données : droit de connaitre les finalités du traitement, de contrôler le transfert éventuel des données à des tiers, d'accéder aux données enregistrées afin d'en vérifier l'exactitude, de les modifier et même de les retirer de la base et d'arrêter leur prise en compte dans les traitements[1].

Le règlement prévoit aussi que des données ne peuvent, en principe, pas être utilisées à d'autres fins que celles pour lesquelles elles ont été collectées. On pourrait donc considérer que la mise à disposition de données d'enquêtes à des équipes différentes de celles qui ont réalisé la collecte, avec de nouveaux objectifs, n'est pas possible, du moins sans solliciter l'accord exprès des personnes concernées L'article 5 prévoit cependant que :

> « Les données à caractère personnel doivent être : […] collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités *; le traitement ultérieur* à des fins archivistiques dans

[1] Ce qui permet de s'interroger sur la prétention de l'Etat à se déclarer seul « propriétaire » des données collectées par les chercheurs : peut-on être propriétaire d'un bien que quelqu'un peut à tout moment vous demander de modifier ou de détruire ?

l'intérêt public, *à des fins de recherche scientifique* ou historique ou à des fins statistiques *n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales* (limitation des finalités); »

L'article 89 auquel l'article précédent renvoie précise :

> *« Le traitement* à des fins archivistiques dans l'intérêt public, *à des fins de recherche scientifique* ou historique, ou à des fins *statistiques est soumis*, conformément au présent règlement*, à des garanties appropriées pour les droits et libertés de la personne concernée. Ces garanties garantissent la mise en place de mesures techniques et organisationnelles, en particulier pour assurer le respect du principe de minimisation des données. Ces mesures peuvent comprendre la pseudonymisation, dans la mesure où ces finalités peuvent être atteintes de cette manière*. *Chaque fois que ces finalités peuvent être atteintes par un traitement ultérieur ne permettant pas ou plus l'identification des personnes concernées, il convient de procéder de cette manière ».*

Le *Plan national pour la science ouverte (PNSO)* de juillet 2018, et le *Comité pour la science ouverte* (*CoSO*) qui en est issu ont commencé à énoncer des directives, tout en invitant les chercheurs à en fixer eux-mêmes les modalités, par discipline et type de données. Pour protéger les intérêts des collecteurs de données, la notion « **d'embargo** » (pour les données d'enquêtes) est apparue : non comme une possibilité de dérogation laxiste, mais comme une mesure *limitative* empêchant les collecteurs de données de différer indéfiniment la mise à disposition publique de leurs données.

Le guide *« Usage et gouvernance des données »* publié en 2019 par le CoSO précise **:**

> (p. 2) « Le PNSO prend en compte ces limites en évoquant non seulement les secrets encadrés par la loi, mais aussi le fait que l'obligation d'ouverture doive être « encadrée par les bonnes pratiques définies par chaque communauté scientifique, par exemple pour définir des durées d'embargo». [2]
>
> *(p. 5) « Préconisation B – Décliner la science ouverte par discipline*
>
> « Dans un premier temps, il faudrait élaborer des guides de la science ouverte, présentant les motivations, les principes et le cadre juridique, plus adaptés à chaque discipline (ou au moins, pour commencer, grand ensemble de disciplines ; ou encore à de grands types de données). »

Dans son « *Plan Données de la recherche du CNRS* » (plaquette de novembre 2020) le CNRS, organisme multidisciplinaire, indique que :

« Le Plan données de la recherche est avant tout piloté par les besoins des scientifiques et prendra en compte les contextes disciplinaires » (art 2, 4è §) ;

et prévoit :

« la possibilité de mettre en place des périodes propriétaires ou d'embargo 'raisonnables' sur les données (par exemple six mois à deux ans) prenant en compte les pratiques disciplinaires. » (art 4, 5è §)

[2] Ainsi les astronomes ont convenu qu'un embargo de 6 mois à un an était raisonnable compte tenu du coût d'acquisition de leurs données, selon Marin Dacos, conseiller « science ouverte » auprès du DGRI *(Colloque OFIS* du 4 avril 2019*)*, position confirmée par des chercheurs de la discipline. A noter que l'embargo s'entend après « pré-traitement des données » pour les rendre intelligibles. Pour certains programmes il peut être convenu que les données seront ouvertes immédiatement.

C’est dans ce contexte d’une double injonction à « ouvrir les données » et à « protéger les données personnelles collectées » que le présent document vise à préciser les conditions d’accès aux données d’enquêtes en sciences sociales, portant sur des personnes physiques.

B/ La notion de « données »

La loi fait référence aux « données issues d’une activité de recherche », sans plus de précision. C’est sans doute pour donner au texte une grande généralité. Il faut cependant constater que la collecte de données dans les enquêtes en sciences sociales n’est pas de même nature que celle de données purement factuelles comme une température, une localisation géographique, un dosage de substance dans l’air, les aliments ou l’organisme, voire une date de naissance… Toutes ces mesures sont sujettes à erreur, bien sûr, et leur traitement doit en tenir compte, mais la donnée mesurée a un caractère objectif que n’ont pas les réponses à un questionnement en sciences sociales. Les variables issues de ces dernières sont donc d’interprétation moins directe, et leur mise à disposition suppose que leur signification soit explicitée par les auteurs de l’enquête[3].

Il en résulte qu’une « donnée » ne peut ici exister sans un accompagnement : dictionnaire des valeurs possibles, population concernée (en raison des filtres), pondération nécessitée par le plan de collecte ou le redressement ex-post, rapprochement nécessaire avec d’autres variables… *L’exploitation comme la mise à disposition d’une telle donnée ne peut donc se faire qu’après préparation de ce corpus.* Cette étape est chronophage et demande des moyens sans lesquels la mise à disposition ne peut être envisagée.

Par ailleurs, s’agissant d’enquêtes auprès de personnes physiques, la collecte, le traitement et la mise à disposition des données collectées *« sont protégées par un droit spécifique ou une réglementation particulière »,* en l’occurrence le RGPD. Le plus souvent, si la collecte a été nominative, les données seront anonymisées, mais elles peuvent rester « indirectement nominatives » (ou « pseudo anonymisées ») et soumises à des procédures d’exploitation spécifiques, que nous ne décrirons pas ici.

La collecte doit être suivie du placement des données dans des *entrepôts spécialisés*, capables de mettre les données à disposition en respectant ces contraintes. Ces structures doivent impérativement disposer des moyens nécessaires. Une fois dûment documentée, la base devient « exploitable » indépendamment des règles d’accès définies ci-après. L’entrepôt doit aussi indiquer comment la source doit être citée, et la référence à la première publication des auteurs de l’enquête dès que celleci est disponible.

C/ L’embargo sur les données d’enquêtes

Pour tenir compte de l’importance du travail d’élaboration d’une enquête, de collecte des données, des opérations de contrôle et de préparation des données, de la responsabilité des « collecteurs » (chercheurs et ITA) dans la qualité et la compréhensibilité des données qui sont ouvertes au public, *il est proposé que chaque établissement de recherche ou université fixe des règles dans les limites qui suivent.*

[3] Dans le rapport de la mission Bothorel « Pour une politique publique de la donnée » (novembre 2020), la définition suivante de la donnée est rappelée (issue d’un arrêté du 22 décembre 1981) : « représentation d’une information sous une forme conventionnelle destinée à faciliter son traitement ».

Schéma pour l'ouverture des données d'enquête

Fin de collecte	*(Dépôt des données)*	Base techniquement "exploitable"	*(1ère publication et/ou 'data paper')*	Fin de l'embargo (base ouverte)
< --------- **maxi 2 ans*** -------- >		< période d'embargo		>
< ---------- **maxi 3 ans*** ---------- >				

** Sauf circonstance exceptionnelle*

1- ***La fin de la collecte est définie comme le recueil des dernières informations individuelles auprès des répondants dans le processus normal de la collecte, indépendamment d'opérations accessoires (comme des contrôles) qui pourraient avoir lieu ensuite ;***

2- ***Les données d'enquête collectées doivent être placées dans un entrepôt de données avec toute la documentation permettant de les exploiter. Sauf difficulté exceptionnelle, la base doit être techniquement exploitable au plus tard deux ans après la fin de la collecte. Ce dépôt peut se faire sous embargo, l'exploitation des données étant réservée pendant l'embargo aux concepteurs de l'enquête (cf. point 4) ;***

3- ***La période entre la fin de la collecte et l'existence d'une base techniquement exploitable doit permettre de constituer les fichiers, d'entreposer les données, de les contrôler, de construire des variables dérivées, de mettre au point les pondérations, et de documenter toutes les étapes de la collecte et du traitement des données ;***

4- ***Lorsque la base de données est techniquement exploitable, elle doit être ouverte au plus tard dans les trois ans suivant la fin de la collecte, sauf circonstances exceptionnelles ; la durée de l'embargo est fixée par le chef d'établissement dans la limite indiquée, après consultation des responsables de l'enquête. L'ouverture, qui signifie la mise à disposition des données à tous les demandeurs, doit se faire dans le respect des règles du RGPD ;***

5- ***La période d'embargo doit permettre aux responsables de l'enquête de préparer une première publication, qui peut être un « data paper » présentant l'enquête et devant être cité par les tous les usagers des données, ou un article incluant les premiers résultats ;***

6- ***Les circonstances exceptionnelles mentionnées aux articles 2 et 4 peuvent résulter soit de circonstances imprévisibles (par exemple le non-respect des délais par l'organisme en charge de l'entreposage), ou de difficultés intrinsèques au type de collecte ou de données en cause (comme les données de cohortes ou panels nécessitant un retour sur les données déjà collectées…) ;***

7- ***L'embargo défini au paragraphe 4 bénéficie aux concepteurs et réalisateurs de l'enquête ; La liste des chercheurs et équipes de recherche protégés par l'embargo doit être établie avant la fin de la période de collecte des données ; elle peut être élargie, avec l'accord des responsables de l'enquête ;***

8- ***Les mêmes règles s'appliquent à des vagues de collecte ultérieures, ou à l'élaboration de variables dérivées si elles sont entrées dans la base de données initiale, ce qui est fortement encouragé ; dans ce dernier cas le délai d'embargo ne peut dépasser un an à partir du moment où la donnée a été documentée et intégrée à la base de données ;***

9- ***Toute publication utilisant les données de l'enquête, y compris celles intervenant pendant la période d'embargo, doit préciser le lieu d'entreposage des données, les conditions de leur accessibilité et respecter les règles de citation de la base utilisée.***

C/ L'exigence de réplicabilité

La publication d'un article utilisant des données d'enquête pose la question de la mise à disposition des données dans un contexte particulier : celui de la soumission, puis de la publication d'un article. Au stade de la soumission, les revues tendent à exiger des auteurs qu'ils joignent à l'article les données utilisées et tous les éléments utiles à la compréhension et au contrôle de leur démarche[4]. Utile et nécessaire au travail des referees, cette exigence place *de facto* des données entre les mains de la revue.

Or l'envoi des données à la revue se heurte à deux obstacles. Le premier est celui de la propriété : en France, pour les travaux financés sur fonds publics, c'est l'Etat qui s'estime propriétaire des données collectées. En dernier ressort, c'est donc lui qui devrait donner son accord (au travers de l'établissement concerné), et non l'auteur de l'article. En second lieu, se pose la question de l'embargo. Si l'article est soumis durant la durée convenue de l'embargo, l'auteur n'est pas censé mettre les données à disposition.

A cet égard on peut souligner une grande ambiguïté. S'il s'agit de faciliter le travail des *referees* pour qu'ils puissent s'assurer que les données utilisées existent vraiment et soutiennent la démonstration proposée, on peut convenir que les données soient transmises à la revue à cette fin et restent ensuite confidentielles. Mais si la revue entend les mettre à la disposition de tout lecteur qui en ferait la demande, il peut y avoir un véritable conflit avec les règles d'embargo, et surtout avec celles du RGPD.

Ajoutons aussi que si la revue devenait ainsi, de facto, pourvoyeuse de données, elle ne serait sûrement pas en mesure de donner un accès « éclairé » à ces données, comme doivent le faire les entrepôts dédiés à cette activité (voir le point D/ ci-après). En résumé :

10- ***Les revues peuvent demander à voir les données utilisées dans une proposition d'article aux seules fins de contrôle de la validité de l'article soumis, et doivent renvoyer leurs lecteurs aux entrepôts existants, en demandant aux auteurs d'apporter***

[4] On n'évoquera pas ici le comportement prédateur de certaines revues, qui entendent s'approprier les données ainsi collectées et les revendre sous forme de « métadonnées ». Notons que l'on est ainsi passé d'une situation de confiance *a priori* dans la démarche des auteurs à une méfiance généralisée. On exige aussi, parfois, que n'importe quel lecteur puisse répliquer exactement la démarche de l'auteur, ce qui constitue aussi une défiance envers les *referees* qui sont supposés contrôler la plausibilité des résultats de l'article.

toutes précisions utiles sur les lieux d'entreposage de leurs données et les conditions d'accès.

11- En aucun cas une revue ne saurait être considérée comme une pourvoyeuse de données, même à la demande de lecteurs à des fins de réplication.

D/ La responsabilité du chercheur/collecteur dans l'exploitation de ses données par d'autres.

Comme il a été rappelé plus haut, le PNSO dit que les « chercheurs seront invités à déposer les données dans des entrepôts de données certifiés, dont la gouvernance et les règles de propriété intellectuelle seront conformes aux bonnes pratiques. »

Il est de l'intérêt de tous, y compris du collecteur des données, que celles-ci soient utilisées à bon escient et de façon appropriée. C'est la raison pour laquelle nous avons affirmé plus haut *qu'une donnée sans son mode d'emploi précis ne doit pas être considérée comme une donnée au sens de la loi de 2016.* En corollaire, l'obligation de dépôt devrait être assortie de l'obligation de documentation, en prévoyant les moyens permettant d'y parvenir. Pour ce faire :

12- Un Plan de gestion des données doit être élaboré avant le début de la collecte, incluant les modalités d'entreposage des données collectées.

Dans le domaine des sciences sociales, il existe déjà des entrepôts bien constitués (comme **QuételetProgedo**), dont la documentation repose sur le travail des chercheurs-collecteurs et sur celui des responsables de l'entreposage, et qui garantissent ainsi un accès éclairé aux données d'enquêtes. Toutefois, ces procédures de mise à disposition sont principalement orientées vers un public de professionnels (chercheurs) et ne sont pas forcément adaptées ou suffisantes pour un public plus large. Pour ce dernier, une pédagogie plus complète serait sans doute nécessaire : il n'est pas toujours simple de comprendre, par exemple, que telle variable n'a pas de signification en elle-même, qu'elle doit obligatoirement être associée à d'autres variables ; et les notions de pondération ou de redressement ne sont pas évidentes pour tout le monde. On peut, par exemple, encourager la mise à disposition d'une base allégée et totalement anonymisée.

13- En toute hypothèse, les chercheurs et les collecteurs doivent impérativement documenter toutes les étapes d'élaboration, de collecte, de contrôle, de recodage, de stockage de leurs données**,** afin d'être à même de répondre à toute question sur ces phases de leur activité, pendant le processus de publication et après.

14- Il résulte des diverses contraintes mentionnées dans les paragraphes précédents que des moyens importants devraient être alloués aux entrepôts et à ceux qui préparent les données à entreposer, pour qu'ils puissent assurer convenablement leur mission de mise à disposition des données à toutes les catégories de demandeurs.