Mettre à disposition des données quanti dans le respect du cadre juridique relatif à la vie privée

Valentin Brunel, Alina Danciu, Marion Lehmans

▶ **To cite this version:**

Valentin Brunel, Alina Danciu, Marion Lehmans. Mettre à disposition des données quanti dans le respect du cadre juridique relatif à la vie privée. Colloque francophone sur les sondages, Oct 2021, Bruxelles, Belgique. hal-03443036

HAL Id: hal-03443036
https://hal.science/hal-03443036v1

Submitted on 23 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

METTRE A DISPOSITION DES DONNEES QUANTI DANS LE RESPECT DU CADRE JURIDIQUE RELATIF A LA VIE PRIVEE - PRATIQUES D'ANONYMISATION EN SHS

Valentin Brunel & Alina Danciu & Marion Lehmans

[1] *Sciences Po, Centre de données socio-politiques (CDSP, CNRS), France*
valentin.brunel@sciencespo.fr, Valentin BRUNEL
[2] *Sciences Po, Centre de données socio-politiques (CDSP, CNRS), France,*
alina.danciu@sciencespo.fr, Alina DANCIU
[3] *Sciences Po, Secrétariat Général, Déléguée à la protection des données, France,*
marion.lehmans@sciencespo.fr, Marion LEHMANS

Depuis 2006, le Centre de données socio-politiques (CDSP, UMS Sciences Po - CNRS) répond aux besoins de la communauté scientifique en matière de production et d'accès aux données ainsi qu'en termes de soutien au développement des méthodes pour collecter et traiter des données. Les données du CDSP sont actuellement diffusées sur le portail français des données en sciences humaines et sociales, Quetelet PROGEDO Diffusion, et moissonnées dans le catalogue européen des données, le CESSDA Data Catalogue.

Le CDSP documente et diffuse des bases de données, comme par exemple les résultats d'élections, mais aussi des enquêtes académiques de sociologie et science politique. Ces enquêtes ont été réalisées soit par sondage auprès d'un échantillon représentatif de la population soit, dans le cas des enquêtes qualitatives, effectuées à partir d'entretiens, d'observations, etc. auprès d'un échantillon réduit de personnes.

Le CDSP met en œuvre les principes du FAIR Data dans l'ensemble de ses pratiques. Les données gérées, extrêmement variées, ont conduit les équipes d'ingénieurs à établir et respecter des procédures formalisées. L'exploration et l'accès aux données sont facilités par l'utilisation de DOI (Digital Object Identifier), le respect de normes comme DDI (Data Documentation Initiative) ou TEI (Text Encoding Initiative). Le CDSP a mis en place également toute une série d'outils d'exploration d'outils comme un serveur Nesstar et travaille actuellement sur un prototype de base de questions respectant le standard de documentation DDI-L, adapté au traitement des données longitudinales ou des enquêtes comparatives.

Le travail du CDSP est donc d'assurer à partir des jeux de données, leur transformation garantissant à la fois :

- la conservation de l'intégrité et de l'intérêt des données pour permettre leur utilisation secondaire,
- la protection des données à caractère personnel des personnes interrogées, dans le respect de l'article 8 de la Charte des droits fondamentaux de l'Union européenne (2000/C 364/01),
- l'ouverture des données exigée par la Loi pour une République Numérique 2016-1321.

Au cœur de ces injonctions contradictoires entre ouverture des données et respect de la vie privée, le travail d'anonymisation prend tout son sens. Peut-on fixer des critères ou indicateurs quantitatifs d'anonymisation, permettant de garantir à la fois l'utilisation secondaire des données par la communauté de recherche et l'anonymat des répondants ?

1. Les questions posées par le cadre juridique relatif à la vie privée

Les données quantitatives tirées d'enquêtes en sciences humaines et sociales correspondent à des catégories particulières de données à caractère personnel au sens de l'article 9 du RGPD : elles peuvent être très précises et comporter de nombreuses informations sensibles sur les personnes. Or, leur compilation, l'interconnexion d'outils de gestion et le croisement de bases de données conduisent ou peuvent conduire à l'identification des personnes concernées, que des données nominatives telles que le nom, le prénom ou l'adresse email figurent directement dans la collecte de données ou pas.

Les établissements de recherche minimisent les risques de violation de données - notamment le risque d'atteinte à la vie privée des personnes concernées - par la mise en place de mesures techniques et organisationnelles assurant la sauvegarde des droits fondamentaux et des intérêts de la personne concernée. Parmi celles-ci figurent la pseudonymisation et l'anonymisation.

> Définitions (source : site de la CNIL)
>
> *L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et ce de manière irréversible.*
>
> *La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires.*

L'article 89 du RGPD clarifie les garanties et dérogations applicables au traitement à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques. Il y est stipulé qu'il convient de procéder à la pseudonymisation ultérieure des traitements de données réalisés à ces fins dès que cela est possible, afin de garantir la protection des droits et libertés des personnes concernées.

Le groupe de travail européen "Article 29" sur la protection des données a publié dès 2014 des lignes directrices sur les techniques d'anonymisation. qui ont naturellement été repris par la CNIL : Ils conseillent de procéder au cas-par-cas, en respectant les principes suivants :

- **L'individualisation** : est-il toujours possible d'isoler un individu ?
- **La corrélation** : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ?
- **L'inférence** : peut-on déduire de l'information sur un individu ?

Dans le domaine des sciences humaines et sociales, les indicateurs d'anonymisation/pseudonymisation qui permettraient d'assurer objectivement cette minimisation des risques, restent encore difficiles à évaluer. Aucun standard ou préconisation valable pour tout jeu de données, n'ont été rendus publics à notre connaissance jusqu'à présent. Les variables en sciences sociales sont par nature plus sensibles et moins uniformisées qu'en mathématiques ou en astronomie par exemple.

De par la nature et la structuration des données et variables traitées, l'anonymisation de données quantitatives ne dispose pour le moment d'aucun standard précis, car elle dépend de nombreux facteurs concurrents, tels que :

- la précision des variables ;
- le nombre de variables disponibles dans une enquête ;
- la répartition des individus enquêtés ;
- l'éloignement dans le temps de l'enquête ;
- la sensibilité des données récoltées.

Dans les services d'appui à la recherche, l'anonymisation se traduit donc par un équilibre entre degré de finesse et pertes des données (ou variables) dans les fichiers diffusés à la communauté scientifique.

2. Nos pratiques et quelques outils pour les tester

Alors que l'anonymisation reste difficile à borner et réaliser de manière automatique, des outils ont été mis en place dans d'autres centres de données afin d'aider à sa mise en œuvre. Ces outils ont été testés au sein du CDSP afin de vérifier la compatibilité de nos pratiques et de ces instruments.

3.1 Nos pratiques

L'anonymisation au sein du CDSP procède surtout par généralisation, jamais par randomisation, de façon à ne pas modifier l'exactitude des données. Cette généralisation se fait habituellement sur les variables clefs d'identification directe, soit des variables socio-démographiques. Les ingénieurs utilisent actuellement des logiciels de traitement statistique comme R, STATA, SPSS ou SAS pour traiter les données et s'assurer de leur anonymisation par des opérations de recodage des variables ou bien de suppression de certaines informations.

Prenons le cas spécifique du panel probabiliste ELIPSS, dont les données sont produites et diffusées par le CDSP. Les données issues des appareils de collecte font l'objet d'une post-production importante visant à regrouper certaines catégories de réponse, et y agréger des données issues d'autres sources (fichiers de pondération, informations sur la région de résidence). Un deuxième travail au cours de la documentation plus précise des données vise à contrôler les catégories où les effectifs sont très faibles et vérifier qu'elles ne puissent conduire à l'identification des répondants. Ensuite, pour garantir la confidentialité des données de ce panel, les ingénieurs veillent à ce qu'il soit impossible d'apparier toutes les données individuelles issues du panel entre elles. Seul le module signalétique de l'enquête annuelle ELIPSS a été apparié de manière systématique à chaque fichier d'enquête. Toute demande d'appariement provenant de plusieurs enquêtes du panel est strictement encadrée par des procédures spécifiques.

Néanmoins, dans certains cas exceptionnels, un enjeu important soulevé par les ingénieurs du CDSP est le fait qu'en croisant un nombre suffisant de variables, il reste parfois encore possible d'isoler un individu, notamment dans le cas des jeux de données comportant moins de deux mille personnes. Dans ce cas, le CDSP procède à des recodages afin d'agréger

certaines modalités de variables problématiques. A cet égard, la variable la plus souvent anonymisée est l'âge précis, ainsi que les variables liées à la commune d'habitation du répondant. Mais ce travail dépend aussi du jeu de données précis : il peut arriver dans certains cas que des variables plus originales soient problématiques, auquel cas il faut à leur tour les anonymiser (un bon exemple peut-être la profession, parfois trop précise).

Un enjeu important est celui de la perte d'informations demandées et intéressantes pour la communauté de recherche. Qu'il s'agisse de la couverture géographique ou de l'âge, certaines données socio-démographiques permettent d'obtenir des résultats de recherche plus précis et complets. Par exemple, le recodage de la date de naissance en catégories conduit nécessairement à une perte de finesse dans les traitements statistiques, ce qui peut nuire à leur significativité. Pour y remédier, le CDSP effectue par exemple des recodages uniquement des valeurs extrêmes de l'âge. Enfin, avec le temps, les ingénieurs ont aussi adopté une approche plus souple de l'anonymisation : un croisement de suffisamment de variables donnera toujours lieu à l'isolement d'une ou deux personnes. L'important est de savoir si les variables croisées permettent ou non l'anonymisation. Ainsi, plutôt que sur le recodage de données, les ingénieurs du CDSP s'interrogent désormais sur le type de données diffusées et le risque d'identification dû à certaines variables.

3.2 Les outils

Afin de vérifier la conformité de nos pratiques avec les recommandations et les usages de la communauté internationale, nous avons décidé de tester deux outils mis en place par le centre UK Data, basé à Londres. Les équipes de ce laboratoire ont produit un logiciel et un script basé sur R permettant d'opérer de manière automatique un certain nombre de vérifications portant sur les jeux de données à diffuser.

QAMD : QAMyData

Ce premier outil a pour vocation de réaliser un bilan initial sur l'état du jeu de données et les risques potentiels de sa diffusion. Les vérifications portent sur quatre points :

- vérification de l'état des fichiers,
- vérification des métadonnées (au moins celles présentes au sein du fichier de données),
- vérification de l'intégrité des données (doublons, fautes de frappe, données manquantes),
- risques liés à l'identification potentielle.

Si tous ces éléments sont importants dans le cadre d'une démarche de diffusion de jeux de données, seul le dernier est directement lié à notre sujet. La vérification avec QAMD porte sur deux cas : les variables comportant des modalités avec moins d'observations qu'une limite donnée (dans le logiciel par défaut, cette limite est d'une observation), et les variables comportant des mots ou des caractères considérés comme "bloquants", comme par exemple les arobases. En cela, cet outil semble conforme aux pratiques du CDSP, qu'il permet d'accélérer dans une certaine mesure en facilitant l'identification de variables problématiques.

sdcMicro

Ce deuxième outil se présente comme un package fonctionnant sur R permettant de réaliser des recodages, procéder à l'anonymisation d'un jeu de données et évaluer le risque que ce dernier présente.

L'outil a une interface divisée en quatre panneaux : le premier est consacré au jeu de données et aux recodages généraux, comme à la déclaration des types de variables (numériques, de strates, NA…). Le deuxième concerne l'anonymisation en tant que telle, et permet de la mettre en œuvre en réalisant une série de tris ou d'opérations de floutage ou de randomisation sur les données. Le troisième produit une analyse en termes de risques et d'utilité quant au recodage. Enfin, le dernier permet d'exporter les données recodées et le script de leur recodage.

L'outil fonctionne par la création de problèmes d'anonymisation, basés sur la prise en compte d'une ou plusieurs variables potentiellement identifiantes. Ces variables peuvent ensuite faire l'objet d'une suppression, d'un recodage, ou de procédures statistiques plus poussées. Si elles sont contenues, il peut s'agir par exemple d'ajouter du bruit ou de les réagréger. En cela, elles sont plus proches des techniques de randomisation que des techniques d'agrégation favorisées au CDSP.

sdcMicro pourrait donc se présenter comme une solution omnipotente en termes d'anonymisation. La variété des procédures proposées et le niveau de risque évalué sont intéressants et méritent d'être creusés.

Cet outil demande cependant avant tout recodage d'identifier quelles variables ont besoin d'être recodées pour éliminer un risque de levée de l'anonymat. Ensuite, les différentes techniques d'anonymisation ne sont pas beaucoup expliquées. Enfin, l'évaluation du risque n'est pas très claire et pourrait faire l'objet d'une documentation plus poussée.

Conclusions provisoires concernant nos pratiques et les outils disponibles

Les outils mis en place au sein du centre UK Data semblent a priori conformes à nos pratiques d'anonymisation. Ils se présentent comme des moyens de les automatiser, au moins dans leurs aspects les plus mécaniques (variables disposant de peu d'observations, recodages facilités). Cependant, ces outils ne permettent pas de manière complètement automatique d'anonymiser des jeux de données. Ils peuvent, sous certaines conditions, accélérer le travail d'anonymisation, voire le contrôler plus facilement, mais ne sauraient se substituer à une expertise métier ni à une connaissance fine du jeu de données à anonymiser.

Cela renvoie à la nature même de l'anonymisation, qui, afin d'être mesurée et de permettre le plus grand nombre de réutilisations intéressantes - dans le respect du RGPD - ne peut être automatique. En effet, seule une connaissance fine des jeux de données permet de sélectionner les variables à regrouper voire à supprimer, sans perte d'informations trop importante. C'est sans doute cet aspect très "cas par cas" de l'anonymisation qui empêche de produire et diffuser des règles universelles la concernant.

Nos tests conduisent à tirer notamment les conclusions suivantes :

- Certains outils sont des extensions d'outils statistiques existants : l'accès à ces outils d'anonymisation est donc conditionné à la maîtrise préalable d'outils type R,
- Aucun outil ne peut se passer d'une expertise technique de traitements de données ni d'une compréhension des enquêtes. Concernant cette compréhension des enquêtes, nous aurions au CDSP un certain nombre de recommandations pratiques, absentes jusqu'ici des guides d'anonymisation existants : effectuer un traitement variable par

variable (en connaissant le jeu de données) ; éviter l'anonymisation sur les variables d'intérêt ; réaliser quelques croisements évidents pour contrôler les possibilités d'identification les plus banales ; ne pas systématiquement tailler dans les données sans analyse préalable, car certaines variables ne sont pas directement identifiantes,

- La formation académique mériterait d'être enrichie de formations techniques de documentation et de mise à disposition de données.

Conclusion

Les techniques d'anonymisation sont donc à développer pour notamment favoriser et familiariser au respect du RGPD, préalable incontournable à la mise à disposition des données. Cette communication entend aussi montrer que l'anonymisation ne peut être faite par une procédure automatique, qu'elle dépend du jeu de données et de l'appréciation du *data manager* ou du chercheur.

La mise à disposition ayant vocation à permettre la reproductibilité des résultats de la recherche et à alimenter la recherche, il faut continuer de s'interroger sur l'impact de la diffusion de données ainsi anonymisées. La recherche peut-elle être reproduite à l'identique dans ces conditions ? Les finalités de recherche ultérieures sont-elles alors impactées ?

La décennie à venir sera celle de la transformation des usages et pratiques de mise à disposition. Le monde du soutien à la recherche doit encore construire ses règles et recommandations. Les travaux de la présente note ont vocation à y contribuer.

Ressources

https://arx.deidentifier.org/downloads/

https://cran.r-project.org/web/packages/sdcMicro/index.html

https://www.ukdataservice.ac.uk/about-us/our-rd/qamydata.aspx

Bibliographie

Elliot, M, Mackey, E, O'Hara, K & Tudor, C 2016, *The Anonymisation Decision Making Framework*. UKAN, Manchester

Groupe de travail "Article 29" de la CNIL, *Avis 05/2014 sur les Techniques d'anonymisation, 2014*

, https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf

Cristina Magder, *Evaluating risk in data: the rules and the tool (sdcMicro)*, Online Demo, Centre de Données Socio-Politiques, Sciences Po Paris 21 October 2019

Cristina Magder, *What Elements to Assess in Data and How to Run QAMyData*, Online Demo, Centre de Données Socio-Politiques, Sciences Po Paris 21 October 2019

Freya De Schamphelaere, *Directive (EU) 2019/1024 on open data and the re-use of public sector information. Opportunities for data archives*