SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

CFS 2020:

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

Valentin Brunel (Sciences Po/CNRS)
Alina Danciu (Responsable Pôle Documentation-Diffusion CDSP (Sciences Po/CNRS)
Marion Lehmans (DPO Sciences Po/ Présidente SupDPO)

6 - 8 octobre 2021 - Université Libre de Bruxelles

Introduction

Les données de la science de plus en plus soumises à deux injonctions qui semblent contradictoires :

- ouverture, autant que possible
- protection, impérative, de la vie privée des répondants.

Dans quelle mesure l'anonymisation peut-elle être garantie de manière formelle, tout en s'assurant que les données diffusées soient suffisamment détaillées pour une utilisation secondaire par la communauté des chercheurs ?

Plan de la présentation

- Cadre juridique
- Recommandations des Délégués à la protection des données
    - Pseudonymiser ou anonymiser ?
- Le Centre de données socio-politiques (CDSP) et la mise à disposition de données
- Nos pratiques en termes d'anonymisation
- Retour d'expérience sur les données ELIPSS

1 - Cadre juridique (1/2)

Protection des données

Règlement (UE) 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, et abrogeant la directive 95/46/CE (règlement général sur la protection des données).

Avis WP-2016 du G29 sur les techniques d'anonymisation

Recommandations de la CNIL en matière d’anonymisation et pseudonymisation

1 - Cadre juridique (2/2)

Science Ouverte : Diffusion sans entrave des publications et des données de la recherche : L'objectif est de faire sortir la recherche financée sur fonds publics du cadre confiné des bases de données fermées. Elle réduit les efforts dupliqués dans la collecte, la création, le transfert et la réutilisation du matériel scientifique.

Loi pour une République Numérique (2016)

Initiative d'Helsinki sur le multilinguisme dans la communication savante (2019)

Plan S : « À partir de 2021, toutes les publications savantes sur les résultats de la recherche financée par des subventions publiques ou privées accordées par des conseils de recherche et des organismes de financement nationaux, régionaux et internationaux devront être publiées dans des revues ou sur des plateformes en accès ouvert, ou immédiatement disponibles sans embargo dans des archives ouvertes. »

2 - Les données particulières (1/2)

Données sensibles : Informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.

> Le règlement européen interdit de recueillir ou d’utiliser ces données, sauf, notamment, si leur utilisation est justifiée par l'intérêt public et autorisé par les autorités compétentes.

2 - Les données particulières dans la recherche SHS (2/2)

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.

3 - Recommandations de DPO à la communauté scientifique (1/2)

- Connaître les responsabilités de chacun
- Assurer la transparence auprès des personnes concernées par les traitements de données à caractère personnel
- Rédiger un protocole d'anonymisation/pseudonymisation selon les grandes phases d'archivage du projet de recherche (courant, intermédiaire, pérenne) et en fonction des réutilisation de données envisagées. Inclure les modalités de dépôt et d'accès au jeu de données sur un ou plusieurs entrepôts de données de la recherche dans la réflexion préalable
- Se former : Faire la différence entre l'anonymisation et la pseudonymisation et connaître et pratiquer les modes opératoires de chacune

3.1 - Pseudonymiser ou anonymiser ? (1/2)

Anonymisation	Pseudonymisation
Jeu de données non soumis au RGPD	Jeu de données soumis au RGPD
Jeu de données plus ou moins altéré	Jeu de données préservé (pas d'altération du détail des données)
Compteur, générateurs de nombres aléatoires, hachage, chiffrement	Ajout de bruit, agrégation, k-anonymat, l-diversité, t-proximité, confidentialité différentielle, etc.

- **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 ?

3.1 - Pseudonymiser ou anonymiser ? (2/2)

L’anonymisation dépend de :

- 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 (données personnelles VS données sensibles).

4 - Recommandations de DPO à la communauté scientifique (2/2)

- Se faire accompagner
- Evaluer les champs socio-démographiques (nombre, finesse) et utiliser les techniques adéquates permettant de supprimer les croisements rares
    - Réfléchir à mettre en place des standards disciplinaires concernant l'agrégation des champs socio-démographiques (politique institutionnelle / nationale / internationale)
- Utiliser des outils permettant d’évaluer le degré de pseudonymisation ou d’anonymisation des jeux de données

5 - Données et méthode

- Le CDSP s'inscrit dans le mouvement des centres de données en sciences sociales qui ont été créés en Amérique du Nord et en Europe depuis les années 1960.
- 350 enquêtes et jeu de données en science politique et SHS, dont plus de 80 enquêtes du panel probabiliste ELIPSS
- Ces enquêtes posent des problématiques particulières en matière d'anonymisation
- Problématiques traitées par les data managers en prenant en compte l'aspect longitudinal des données

L’anonymisation au CDSP

- Centre de données depuis 2005
- Evolution des outils et des compétences , respectant les principes FAIR (Findable, Accessible, Interoperable, Reusable)*
- Injonctions contradictoires : ouverture des données et protection de la vie privée des enquêté.e.s
- Deux cas : les données produites au CDSP, les données d’autres chercheurs et institutions.

Mentionnés pour la première fois en 2016, dans un article de la revue Nature : https://www.nature.com/articles/sdata201618

ELIPSS - Étude Longitudinale par Internet Pour les Sciences Sociales

- Un dispositif de recherche unique en France, mis en place depuis la fin de l'année 2012, afin d'étudier l'évolution des comportements, des situations et des opinions dans la société française.
- Panel probabiliste d'environ 3000 personnes en 2018 qui répondent à des enquêtes conçues par des chercheur.e.s en SHS
- Depuis 2012, une enquête par questionnaire/ mois en moyenne
- Des informations très riches et détaillées concernant les enquêté.e.s
- Des données diffusées à la communauté scientifique

3 - Nos méthodes d’anonymisation

Exemple : les données du panel ELIPSS

1e étape : Post-production (suppression de variables, recodages)

2e étape : Contrôle des catégories de variables au cours de la documentation

3e étape : Interdiction de croiser les jeux de données entre eux

3 - Nos méthodes d’anonymisation

Anonymisation par généralisation

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, ou encore la profession.

Mais ce travail dépend aussi du jeu de données : 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).

3 - Nos méthodes d’anonymisation

- Impossible d'apparier toutes les données individuelles issues du panel entre elles
- Une partie seulement de l'enquête annuelle ELIPSS est appariée de manière systématique à chaque fichier d'enquête
- Ces fichiers sont également assortis de données contextuelles issues du recensement et de leur pondération transversale (voir la documentation des pondérations ELIPSS)
- L'appariement des informations du panel avec des données extérieures (fiscales, santé, etc.) est exclu et impossible

3 - Nos méthodes d’anonymisation

Deux outils proposés par UK Data Archives (libres):

QAMyData

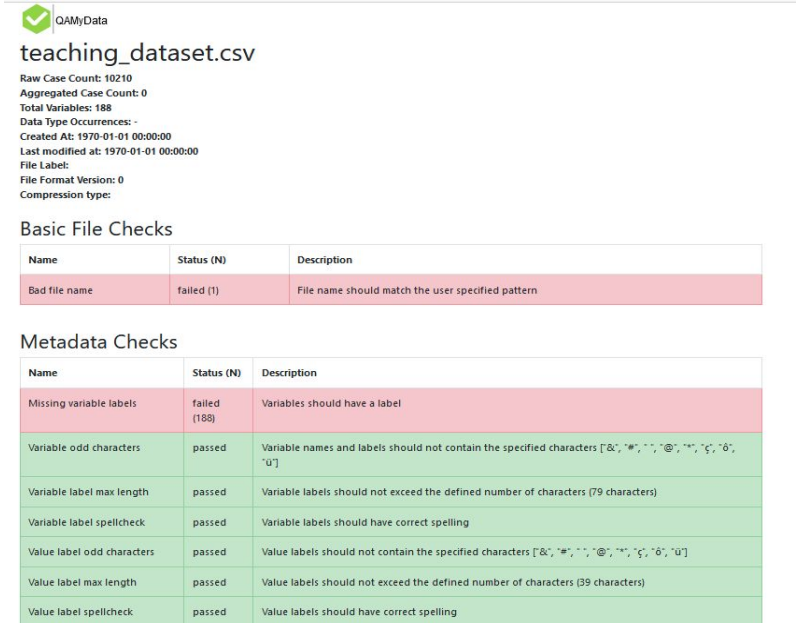
scdMicro

3 - Nos méthodes d’anonymisation

QAMyData

Permet d’effectuer des vérifications sur :

- l’intégrité du fichier
- les métadonnées et leur syntaxe
- les valeurs des variables (NA, doublons…)
- les possibilités d’identification (faibles valeurs, caractères spéciaux)



QAMyData

teaching_dataset.csv

Raw Case Count: 10210
Aggregated Case Count: 0
Total Variables: 188
Data Type Occurrences: -
Created At: 1970-01-01 00:00:00
Last modified at: 1970-01-01 00:00:00
File Label:
File Format Version: 0
Compression type:

Basic File Checks

Name	Status (N)	Description
Bad file name	failed (1)	File name should match the user specified pattern

Metadata Checks

Name	Status (N)	Description
Missing variable labels	failed (188)	Variables should have a label
Variable odd characters	passed	Variable names and labels should not contain the specified characters ["&", "#", " ", "@", "*", "ç", "ô", "ü"]
Variable label max length	passed	Variable labels should not exceed the defined number of characters (79 characters)
Variable label spellcheck	passed	Variable labels should have correct spelling
Value label odd characters	passed	Value labels should not contain the specified characters ["&", "#", " ", "@", "*", "ç", "ô", "ü"]
Value label max length	passed	Value labels should not exceed the defined number of characters (39 characters)
Value label spellcheck	passed	Value labels should have correct spelling

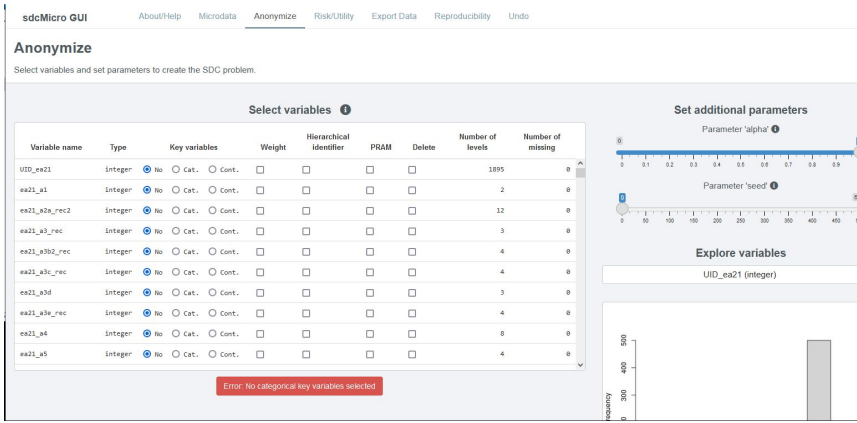
3 - Nos méthodes d’anonymisation

sdcMicro

Package R librement téléchargeable

Permet d’effectuer des modifications d’anonymisation :

- k anonymisation
- post-randomisation
- suppression des individus à haut risque d’identification



3 - Nos méthodes d’anonymisation

- Un exemple avec le jeu de données Enquête annuelle 2021 (ELIPSS)

Choix des variables à problème est déterminant.

Age, Genre, PCS, Région ? Mais aussi Poids, Taille, etc.

Le principe est d’empêcher qu’une personne se retrouve seule avec ces caractéristiques.

3 - Nos méthodes d’anonymisation

Hors le grand nombre de variables identifiantes et les effectifs relativement faibles font que l’anonymisation formelle concernant les jeux de données destinés aux chercheurs n’est pas possible.

Donc deux solutions :

- Un jeu de données “recherche” dont les conditions d’accès sont limitées
- Un jeu de données “pédagogiques”, ouvert librement, aux données plus agrégées

3 - Nos méthodes d’anonymisation

Les données pédagogiques du CDSP :

Une expérimentation pour diffuser librement des données en SHS.

Principe : simplifier les jeux de données à la fois en colonne (moins de variables) et en ligne (recodages afin d’agréger plus les catégories).

Une règle simple : respecter le k-anonymat pour k=2.

Discussion

L’anonymisation formelle demeure impossible en SHS, si l’on souhaite conserver des données relativement précises.

En cause : la faiblesse des échantillons et le grand nombre de variables.

Cependant il est possible d’établir un protocole d’anonymisation / pseudonymisation en plusieurs temps et en fonction des ré-utilisations envisagées (ouverte des données à la communauté scientifique ou au grand public)

Discussion

Outils de UK Data Archive intéressants, permettent d'automatiser

Mais : impossible de les utiliser sans connaître les données. Même alors, l'outil peut comprendre comme variables identifiantes des variables pas toujours très problématiques.

Workshop de CLOSER sur les mêmes questions : nécessité d'un travail individuel sur chaque type de données.

Enfin, des différents degrés de l'anonymisation…

Conclusion

Peut-être une bonne chose qu’il n’y ait pas de réponse facile.

Rôle des ingénieurs.

Importance du temps passé à comprendre les jeux de données.

Des pratiques de diffusion différenciées, en fonction du degré d’anonymisation.

Merci !

valentin.brunel@sciencespo.fr
alina.danciu@sciencespo.fr
marion.lehmans@sciencespo.fr

Special thanks to : Quentin Gallis (quentin.gallis@sciencespo.fr)