Pratiques d'anonymisation et pseudonymisation des données diffusées par le Centre de données socio-politiques (CDSP)

Valentin Brunel, Paul Colin, Alina Danciu, Quentin Gallis, Guillaume Garcia

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

Pratiques d'anonymisation et pseudonymisation des données diffusées par le Centre de données socio-politiques (CDSP)

Auteurs : Valentin BRUNEL, Paul COLLIN, Alina DANCIU, Quentin GALLIS, Guillaume GARCIA
Validation : Nawale Lamrini, Nicolas SAUGER
Date : 15 avril, Version 1.0

Le Centre de données socio-politiques, Unité d'Appui à la Recherche (UAR) de Sciences Po et du CNRS, documente et diffuse des **données d'enquêtes en sciences sociales, réalisées selon des méthodes qualitatives ou quantitatives**, et accompagnées d'une documentation suffisamment riche pour en permettre la réutilisation.
Les données diffusées par le CDSP sont diffusées soit en open data, pour les résultats d'élections, soit en accès restreint à la communauté scientifique (chercheur.e.s, étudiant.e.s, ingénieur.e.s…) et avec un contrôle manuel pour les données d'enquêtes pseudonymisées. La licence CC BY 4.0 est utilisée.
Le CDSP respecte la norme OAIS pour la gestion des données qu'il a en charge et a défini des étapes dans leur gestion : dépôt, data management (vérification, anonymisation, pseudonymisation, ajout de métadonnées), diffusion. Le traitement des données se fait en respectant les standards internationaux en vigueur (Data Documentation Initiative - DDI) et nos processus - et données - suivent les principes FAIR (des DOI sont par exemple attribués aux jeux de données, etc).
Une première section de ce document est consacrée au cadre législatif de la diffusion des données, ainsi que les types d'accès aux données du CDSP. Sont abordées ensuite nos procédures d'anonymisation et pseudonymisation de données. Des annexes donnant des exemples et un glossaire viennent finir ce document.

1) Le Cadre législatif de la diffusion des données

Selon la Directive 2013/37/UE, dite PSI (Public Sector Information), et la loi n°78-753 du 17 juillet 1978 (modifiée), dite CADA (Commission d'Accès aux Documents Administratifs), les

données issues d'une activité de recherche financée au moins pour moitié par des dotations d'un Établissement Public, de l'Etat, des collectivités, des subventions d'agences de financement nationales ou par l'Union européenne sont assimilées à des "documents administratifs".[1]
Dès lors qu'elles sont considérées comme achevées (notamment après une première publication volontaire, comme la publication scientifique) et si elles n'entrent pas dans le cadre des exceptions légales (cf. cas particuliers ci-dessous), elles doivent être :

- communicables à toute personne qui en fait la demande,
- librement réutilisables.

La Loi n° 2016-1321 du 7 octobre 2016 pour une République Numérique va plus loin et modifie le code des relations entre le public et l'administration pour aller vers un principe d'ouverture et de diffusion spontanée des données sur Internet (open data), et de libre réutilisation sans condition et à toutes fins (y compris à des fins commerciales et y compris pour les établissements publics).

La réutilisation des informations publiques est soumise à la condition que ces dernières ne soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources (auteurs et date de dernière mise à jour) soient mentionnées.

S'il y a une obligation à publier certaines données (concernant des émissions de substances dans l'environnement, par exemple), d'autres sont publiables sous conditions. Dans ce cas rentrent les données personnelles, définies comme toutes informations identifiant directement ou indirectement des personnes physiques ou ensembles de données qui, recoupées, peuvent permettre d'identifier des personnes). L'accord exprès des personnes avant le recueil, puis l'anonymisation des données avant publication ou à la fin du projet, et enfin l'accès des personnes à leur données durant le projet, sont obligatoires (cf. RGPD, succédant en 2016 à la Loi Informatique et Liberté).[2]

Les données tirées d'enquêtes en sciences humaines et sociales correspondent pour une part importante d'entre elles 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 rendre possible 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 non.

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)[3]
>
> *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.*

---

[1] Source : https://data.ird.fr/cadre-juridique/
[2] Source : https://data.ird.fr/cadre-juridique/
[3] Pour aller plus loin : https://pod.univ-lille.fr/ethique-et-protection-des-donnees-en-recherche/video/16591-s17-anonymisation-pseudonymisation/

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, ainsi qu'à des fins statistiques. Il dispose 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[4] : 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 ?

"*Exemple d'inférence : si un jeu de données supposément anonyme contient des informations sur le montant des impôts de personnes ayant répondu à un questionnaire, que tous les hommes ayant entre 20 et 25 ans qui ont répondu sont non imposables, il sera possible de déduire, si on sait que M. X, homme âgé de 24 ans, a répondu au questionnaire, que ce dernier est non imposable*."

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'a été rendu public à notre connaissance jusqu'à présent. Les variables[5] en sciences sociales sont par nature plus sensibles et moins uniformisées qu'en mathématiques ou en astronomie, par exemple. Par ailleurs, dans des disciplines comme les mathématiques ou l'astronomie, on ne traite pas, en règle générale, de données personnelles.

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 différents, tels que :

- la précision des variables ;
- le nombre de variables disponibles dans une enquête ;
- la répartition et le nombre 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.

[4] https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data
[5] Voir définition glossaire

2) Accès aux données du CDSP

Les données documentées et diffusées par le CDSP sont disponibles sur data.sciencespo.fr[6], selon des modalités d'accès diverses. Les métadonnées sont par ailleurs moissonnées sur les plateformes de PROGEDO et du CESSDA.

- Données anonymisées : diffusées en accès libre.
  NB : certains résultats électoraux sont diffusés par le CDSP sur la plateforme data.gouv.fr
- Données pseudonymisées : diffusées en accès restreint, avec un contrôle manuel des accès par les ingénieurs du CDSP, à la communauté de recherche et d'enseignement uniquement.

Ces deux catégories correspondent respectivement aux **Accès A et B** du tableau suivant (un mapping avec les accès CESSDA a été réalisé) :

Accès	Type accès données	Procédure de demande	Type utilisateur	Mapping CESSDA
A	Accessible à tous, pour des projets de recherche et enseignement au sein de la communauté scientifique ou des projets hors communauté scientifique (ex : journalistes, chargés de marketing, etc.)	avec enregistrement, en acceptant les CGU : https://data.sciencespo.fr/misc/cond_jur/ToU.pdf	TOUS	Open Registration and click of terms of use)
B	Accessible uniquement pour la communauté scientifique, pour des projets de recherche et d'enseignement (thésards, étudiants en master et licence inclus)	avec enregistrement, en acceptant les CGU: https://data.sciencespo.fr/misc/cond_jur/ToU.pdf	chercheur.e.s, doctorant.e.s, étudiant.e.s, ingénieur.e.s, chargé.e.s d'études appartenant à une institution de recherche ou enseignement supérieur	Safeguarded/accoutable Secure download and signed user contract/license (registration or application)

[6] Exemple de notice d'enquête : https://data.sciencespo.fr/dataset.xhtml?persistentId=doi:10.21410/7E4/DWHY6T

3) Les principes généraux et grandes étapes de l'anonymisation au CDSP

Dans cette section, nous allons présenter brièvement nos techniques et outils d'anonymisation.

1) La pseudonymisation

Quelle que soit l'enquête diffusée, le CDSP procède a minima à une pseudonymisation des données. Cette première étape, indispensable, correspond à la suppression des variables directement identifiantes (nom et prénom, adresse, etc.) si disponibles, et à leur remplacement par des identifiants numériques.

Cette première étape peut aussi prendre la forme d'un recodage ou d'une suppression des variables les plus potentiellement identifiantes, comme par exemple les communes d'habitation ou encore l'âge précis. La plupart du temps, ce recodage ou cette suppression suivent un objectif d'équilibre entre la volonté de perdre le moins possible la précision des variables et le fait de protéger l'anonymat des répondants.

Pour prendre un exemple d'une pratique assez répandue dans les jeux de données quantitatifs du CDSP, l'âge, si disponible en clair, pourra faire l'objet d'un recodage pour ses valeurs extrêmes uniquement. En revanche, si le nombre de répondants est trop faible, un recodage en tranches d'âge sera plutôt favorisé.

Voici les variables traditionnellement surveillées dans le cadre de l'opération de pseudonymisation (liste non exhaustive) : *sexe, âge, profession, commune, région, département, taille de l'unité urbaine, catégorie de l'aire urbaine, situation maritale, revenu, diplôme, nombre d'habitants dans le foyer, ethnicité et origine*.

Pour chacune de ces variables, nous veillons à limiter le nombre de catégories avec très peu d'effectif (moins de cinq), voire à les supprimer ou les recoder. Quelques croisements communs (âge et diplôme, genre et PCS, par exemple) permettent aussi de nous assurer que les individus ne se retrouvent pas trop rapidement isolés dans une seule catégorie.

2) Les freins mis à l'utilisation et à l'appariement des données

Une autre étape importante concernant la protection des données personnelles est la procédure de demande des données et les divers obstacles qui y sont posés à une utilisation abusive des données diffusées. Ainsi, comme précisé plus haut (voir partie 2), les données sont disponibles uniquement aux personnes justifiant de l'affiliation à une structure de la communauté scientifique.

De plus, les jeux de données diffusés dans le cadre d'enquêtes longitudinales, tels que ceux produits avec le panel ELIPSS[7], le sont de telle sorte que les appariements sont rendus difficiles entre les données par l'absence d'une variable de passage commune. Le CDSP peut réaliser, si besoin, des appariements manuels sur demande, mais il est impossible pour une personne extérieure de croiser directement les données de différentes enquêtes entre elles. Cela permet d'éviter l'accumulation d'informations sur une seule personne, qui permettrait de l'identifier. On rend ainsi plus ardue l'identification, tout en disposant de moins de données sensibles sur les personnes.

Enfin, tout appariement avec des données externes est rendu impossible par l'absence de table de passage (c'est-à-dire d'une variable commune permettant de joindre ces données), dans le cadre des données ELIPSS comme des autres données diffusées par le CDSP. Cela permet d'éviter encore une fois de mettre à disposition des utilisateurs une quantité d'informations qui, recoupées, permettraient d'identifier des individus.

3) Dans certains cas, l'anonymisation

La pseudonymisation associée aux freins mis à l'utilisation "*sauvage*" de données concerne l'ensemble des jeux de données diffusés au CDSP. Cependant, dans certains cas, il nous a paru utile d'y ajouter une étape supplémentaire.

Pour les jeux de données dits pédagogiques, librement téléchargeables, le respect de la réglementation et la protection des données des répondants nous imposent une anonymisation stricte des réponses aux enquêtes. Pour cela, nous avons utilisé des outils automatiques d'anonymisation fournis par UK Data.

Ces outils ont l'avantage, par rapport à nos méthodes traditionnelles, de garantir statistiquement que l'anonymisation est complète et que la réidentification des personnes est rendue plus complexe, avec notamment des indicateurs chiffrés. La réidentification n'étant jamais formellement impossible (même si dans la pratique ces cas sont très rares), les indicateurs se présentent sous forme de risques.

A cet égard, le package R "sdcMicro"[8], développé par UK Data, permet de procéder de manière intuitive à des recodages de variables "*problématiques*", de repérer quels sont les individus les plus à risque en termes d'identification potentielle, de supprimer les valeurs les plus identifiantes, ainsi que de procéder à un k-anonymat en prenant en compte certaines

[7] https://cdsp.sciences-po.fr/fr/projets/panel-elipss/
[8] Un papier sur l'utilisation de scdMicro est disponible ici : http://www.tdp.cat/issues/tdp.a004a08.pdf

variables. Plus largement, le package permet aussi de garantir un i-anonymat, ou encore de procéder à des post-randomisations.

L'utilisation de ce package dans le cadre des jeux de données pédagogiques (voir Annexe) permet de garantir l'impossibilité formelle d'identifier au sens du RGPD un répondant sur la base de ses réponses à l'enquête donnée. Ainsi, le respect d'un k-anonymat de 2 garantit qu'il est impossible dans ces jeux de données d'isoler un individu sur la base des variables potentiellement identifiantes. Cette méthode suppose les étapes suivantes : définir quelles sont les variables potentiellement identifiantes, les ranger dans l'ordre de dangerosité, identifier quels recodages sont pertinents ou non, et quel degré de suppression de variables devient réellement handicapant pour la réutilisation des données. Même si l'outil simplifie le processus d'anonymisation, il reste un nombre très important de choix à effectuer par l'équipe documentation, qui demandent une véritable connaissance du management des données.

APPENDICE 1 : TROIS EXEMPLES D'ANONYMISATION

1) Enquêtes ELIPSS

L'anonymisation des enquêtes ELIPSS est un processus long et complexe, qui met en œuvre les savoir-faire de plusieurs équipes au sein du CDSP, et engage l'ensemble du laboratoire. En effet, les données étant produites au sein du CDSP, il s'agit à la fois d'assurer la séparation des différents types des données (données de contact (nom, prénom, adresse, etc.) et données d'enquête), la conservation et la protection de ces données sur des espaces différenciés, et leur bonne transformation en vue d'une diffusion conforme à la réglementation.

Nous suivrons pour cela le chemin d'une enquête ayant été diffusée dans le cadre des jeux de données pédagogiques, afin de présenter l'ensemble des procédures visant à la protection de l'anonymat des répondants.

a) Préparation des données et post-production

La première étape concerne l'ensemble des procédures visant à construire la base de données qui fera l'objet d'une documentation. Ainsi, il faut transformer les données brutes issues du logiciel d'enquête en une base pouvant être diffusée à l'équipe chargée de la documentation.

La procédure de post-production passe par trois étapes[9] : le nettoyage de la base brute tirée du logiciel d'enquête, la génération d'une table de passage permettant l'appariement à d'autres enquêtes en vue de l'enrichissement de la base de données, et enfin la création des fichiers de diffusion.

[9] *Document interne réservé à l'équipe "Production" du CDSP.*

Lors de la première étape, un certain nombre de variables tirées du logiciel d'enquête sont supprimées, et le fichier est stocké sur un espace distinct et réservé à quelques personnes. La génération de la table de passage (étape 2) se fait dans un second temps et permet un suivi efficace des individus d'une enquête à l'autre, afin de bien réaliser les appariements.

Ces appariements constituent la troisième étape de la procédure de postproduction. La table de passage et la variable d'identification servent à relier les individus ayant répondu à l'enquête à leurs réponses sur toute une série de variables tirées de l'Enquête annuelle. Ces variables – au nombre d'une centaine environ – permettent de connaître les principales caractéristiques socio-démographiques des répondants, et sont appariées à celles de l'enquête principale. Enfin, des pondérations sont calculées sur la base des répondants à l'enquête et ajoutées au jeu de données.

Celui-ci est alors complet et déposé sur un espace de stockage distinct. Les fichiers intermédiaires sont supprimés.

b) Recodages et pseudonymisation

Une fois la base de données post-produite obtenue, l'équipe en charge de la documentation s'occupe de la rendre diffusable à la communauté scientifique. Pour cela, un certain nombre de variables doivent encore être supprimées, et des recodages peuvent être effectués.

Parmi les variables supprimées, on peut compter les variables ayant trait aux conditions de passation de l'enquête (paradonnées), comme le type d'appareil de réponse, les temps de passation, etc. D'autres variables habituellement supprimées ou recodées sont les variables textuelles ou comportant des verbatims, potentiellement identifiants. Ensuite, l'équipe en charge de la documentation vérifie que les modalités des différentes variables sont cohérentes et qu'il ne reste pas de variable potentiellement identifiante. Lors de cette vérification, il est possible que de nouveaux recodages soient effectués.

A l'étape précédente, les données ont été recodées suivant un certain nombre de critères, comme le nombre de modalités par variable, ou encore les effectifs par modalité. De même, les données de l'enquête annuelle ont fait l'objet d'un appariement sur plusieurs enquêtes annuelles, afin de récupérer les informations socio-démographiques les plus récentes pour tous les répondants à l'enquête.

Le fichier dans son ensemble fait l'objet d'une relecture approfondie, variable par variable, avant d'être diffusé. À la fois les métadonnées (libellé de la variable, des modalités de réponse, filtres, questions, instructions enquêteur, etc.) et les données (tris à plat, à minima) sont passées en revue.

c) Anonymisation : le cas des bases de données pédagogiques

La base de données est maintenant prête à être diffusée sur les plateformes de diffusion du CDSP. Cependant, une étape supplémentaire est nécessaire afin de garantir l'anonymat complet des panélistes si l'on veut la diffuser plus largement.

En effet, le CDSP a récemment produit des "*bases de données pédagogiques"*, dont l'accès est moins restrictif que les enquêtes ELIPSS "*traditionnelles"*. Ces bases de données ont pour vocation d'être utilisées pour l'enseignement et l'auto-formation, et sont donc volontairement amputées d'une partie conséquente de leurs données et documentation habituellement présentes dans les DIP (*Dissemination Information Package*) du CDSP.

La procédure de création des bases de données pédagogiques consiste à ne retenir qu'une cinquantaine de variables choisies pour leur intérêt pédagogique. Certaines variables sont recodées ou modifiées, pour les besoins de l'anonymat et de la simplicité d'utilisation. En conséquence, les données peuvent différer de la base originelle. De plus, la documentation est réduite à un simple tableur contenant les intitulés de questions, les labels de réponse, ainsi que d'autres informations nécessaires (univers, notes…).

Par ces étapes, les bases de données pédagogiques nécessitent beaucoup moins d'efforts et de temps pour être prises en main, et se prêtent mieux à une utilisation dans le cadre de cours, ou de formation à la manipulation de données. En revanche, cela les rend inaptes à la recherche, en raison du manque de certaines informations et à la transformation d'autres.

En raison du but de diffusion plus large de ces bases de données, il est nécessaire de les anonymiser complètement. En conséquence, le package sdcMicro a été utilisé avec R pour estimer les possibilités de réidentification permises par certaines variables démographiques. Les variables retenues sont l'âge, le sexe, le niveau de diplôme, et la Tranche d'Unité Urbaine.

Après une première lecture des réidentifications possibles, ces variables ont été récodées en catégories plus larges (hormis le sexe). Par exemple, les classes d'âges ont été ramenées, dans la base de données pédagogiques "*Petev, Ivaylo, 2020, "Styles de vie et Environnement (2017)"*, [10] de 7 à 5 modalités, de façon à retirer certaines classes d'âges dont les effectifs étaient faibles, et donc susceptibles de permettre de identifications.

Par la suite, l'outil de gestion du k-anonymat a été utilisé pour supprimer les observations qui violaient un k-anonymat fixé à 2. Cela signifie que lorsque des configurations de modalités (valeurs identiques pour l'ensemble de variables sélectionnées) apparaissaient moins de trois fois, une de ces valeurs était supprimée (transformée en NA) pour empêcher la réidentification. Ainsi, si seulement deux personnes étaient des femmes de plus de 75 ans titulaires d'un doctorat et vivant en milieu rural, l'une des valeurs d'âge, de diplôme ou de lieu de résidence serait pour chacune transformée en NA. Le choix était établi en fonction de l'importance de la variable (qui peut être déterminée au sein du logiciel) et du jeu de données dans son ensemble (afin de supprimer le moins d'observations possibles).

L'ensemble des étapes d'anonymisation est résumé dans un script qui est exporté depuis l'interface de sdcMicro. Voici reproduit celui qui a été utilisé pour créer une base de données pédagogique à partir de l'enquête *Sven* (Petev, Ivaylo, 2020, "Styles de vie et Environnement (2017)", https://doi.org/10.21410/7E4/KPFUQV) :

---

[10] *Petev, Ivaylo, 2020, "Styles de vie et Environnement (2017)", https://doi.org/10.21410/7E4/KPFUQV, data.sciencespo, V4*

```
# created using sdcMicro 5.6.1
library(sdcMicro)

obj <- NULL
if (!exists("dataset_sven")) {
  stop('object "dataset_sven" is missing; make sure it exists.`', call. = FALSE)
}
obj$inputdata <- readMicrodata(path="dataset_sven", type="rdf", convertCharToFac=FALSE,
drop_all_missings=FALSE)
inputdataB <- obj$inputdata

## Convert a numeric variable to factor (each distinct value becomes a factor level)
inputdata <- varToFactor(obj=inputdata, var=c("habitat_TUU2014","demo_sexe","demo_diplome","demo_age"))
## Set up sdcMicro object
sdcObj <- createSdcObj(dat=inputdata,
            keyVars=c("habitat_TUU2014","demo_sexe","demo_diplome","demo_age"),
            numVars=NULL,
            weightVar=NULL,
            hhId=NULL,
            strataVar=NULL,
            pramVars=NULL,
            excludeVars=NULL,
            seed=0,
            randomizeRecords=FALSE,
            alpha=c(1))

## Store name of uploaded file
opts <- get.sdcMicroObj(sdcObj, type="options")
opts$filename <- "dataset_sven"
sdcObj <- set.sdcMicroObj(sdcObj, type="options", input=list(opts))


## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="habitat_TUU2014", before=c("1","2","3"), after=c("1"),
addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="habitat_TUU2014", before=c("4","5","6"), after=c("2"),
addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="habitat_TUU2014", before=c("7"), after=c("3"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="habitat_TUU2014", before=c("8"), after=c("5"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_diplome", before=c("1","4","5","6","7"), after=c("1"),
addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_diplome", before=c("8","9"), after=c("2"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_diplome", before=c("10"), after=c("3"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_diplome", before=c("11"), after=c("4"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_age", before=c("1","2"), after=c("1"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_age", before=c("3"), after=c("2"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_age", before=c("4"), after=c("3"), addNA=FALSE)
```

```
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_age", before=c("5"), after=c("4"), addNA=FALSE)
## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="demo_age", before=c("6","7"), after=c("5"), addNA=FALSE)
## Local suppression to obtain k-anonymity
sdcObj <- kAnon(sdcObj, importance=c(3,1,2,4), combs=NULL, k=c(3))
```

2) Enquêtes déposées au CDSP par la communauté scientifique

Dans le cadre de ses missions d'UAR, le CDSP documente et diffuse, comme précisé auparavant, les enquêtes qu'il produit, mais aussi des enquêtes quantitatives et qualitatives déposées par la communauté scientifique nationale et internationale.

Enquêtes quantitatives

Ces enquêtes font l'objet de vérifications et recodages par les membres de l'équipe. Des tris à plat et des tris croisés sont réalisés pour vérifier la cohérence des données, l'absence des valeurs aberrantes, mais aussi et surtout si certaines catégories de réponse n'ont pas des effectifs trop faibles. En dessous de cinq personnes, nous réalisons des recodages. Au cas où certaines variables comme la Commune sont déposées, nous les supprimons, de commun accord avec l'équipe déposante, de la base de données diffusée.

Par ailleurs, pour les variables filtrées, nous regardons s'il y a des anomalies au niveau du filtre, et si les valeurs manquantes sont correctement renseignées. Il nous arrive d'avoir par exemple des valeurs SYSMISS. Au cas où le recodage ne peut pas être réalisé par nos soins, nous revenons vers l'équipe déposante pour obtenir des précisions. Dans tous les cas, nous informons les déposants des traitements réalisés sur les données et les syntaxes sont sauvegardées.

D'autres recodages peuvent être réalisés, selon des circonstances particulières. Par exemple, le baromètre du Défenseur des droits possède une variable *"Indice de masse corporelle".* Une de ses modalités de réponse était "*Anorexie*", terme médical que nous avons préféré recoder en "*Maigreur*", pour plus de neutralité.

Nous vérifions également les réponses aux questions ouvertes, si elles sont diffusées. Par exemple, pour les réponses concernant la profession, si celle-ci est une profession rare, comme celle de Santonnier (environ 150 existent en France), nous sommes très vigilants quant à la diffusion d'une telle réponse.

Enquête qualitatives

Les enquêtes qualitatives sont traitées selon des modalités relativement similaires à celles mises en œuvre pour les enquêtes quantitatives. Une partie des éléments exposés ci-dessous sont décrits dans la publication suivante :

FROMONT, Emilie, Selma BENDJABALLAH, Guillaume GARCIA, Sarah CADOREL, Emeline JUILLARD, and Emilie GROSHENS. "Anonymat et confidentialité des données : l'expérience de beQuali." In La diffusion

numérique des données en SHS. Aix en Provence: Presses universitaires de Provence, 2018.https://hal-sciencespo.archives-ouvertes.fr/hal-02873570

Certaines spécificités des protocoles mis en place pour les enquêtes qualitatives sont liées aux quatre facteurs suivants :

1) La nature et la complexité des jeux de données qualitatives

Les corpus d'enquêtes étant parfois complexes – constitués de plusieurs dizaines, voire centaines de documents différents –, les informations personnelles et sensibles peuvent se retrouver dans de nombreux fichiers différents, qu'il s'agisse de textes ou de tableaux, ou parfois des photographies. Le travail d'anonymisation doit être réalisé en priorité sur les "données", c'est-à-dire les matériaux collectés sur les terrains d'enquête (essentiellement des transcriptions d'entretiens et des notes d'observations), qui concentrent la plus grande quantité d'informations sur les enquêtés. Il doit aussi être fait, de manière secondaire, sur les autres documents constitutifs des corpus (mentions de prises de contact avec certaines personnes, groupes ou institutions, précisions sur des lieux, présence d'extraits de verbatim des entretiens dans des documents d'analyse, etc.). Ces différentes informations pouvant se recouper, il convient d'être particulièrement vigilant pour veiller à la cohérence globale de l'anonymisation.

2) La dimension proprement "qualitative" des données personnelles et sensibles

Les transcriptions d'entretiens et notes d'observation contiennent souvent des récits approfondis ou des descriptions fines des personnes (parcours biographiques, opinions, activités, etc.). La complexité des développements sous lesquels les informations personnelles et sensibles sont présentées complexifie d'autant le travail d'anonymisation. La difficulté de ce travail est renforcée par le fait que les mêmes informations peuvent être répétées, parfois sous des formes sensiblement différentes, à différents endroits d'un même texte, ou dans plusieurs textes distincts, ce qui nécessite un effort de mise en cohérence. Cette difficulté est redoublée par le fait qu'aux informations personnelles et sensibles, telles que définies par la loi, s'ajoutent parfois d'autres éléments susceptibles de porter préjudice aux enquêtés, qui nécessitent dès lors une couche supplémentaire d'intervention – notamment des propos ou des actions susceptibles de tomber sous le coup de la loi (par exemple, des propos ou des actions racistes ou xénophobes), ou des informations relevant davantage de l'éthique de l'enquête (interviennent ici n'importe quels types d'informations susceptibles de mettre les enquêtés en défaut dans leurs milieux sociaux d'appartenance, en cas de rupture de confidentialité).
Les enquêtes dites "*ethnographiques*", c'est-à-dire menées sur de petits milieux sociaux circonscrits, et marquées par un degré important - bien que variable - d'interconnaissance des personnes enquêtées, ainsi que par une immersion plus ou moins profonde des enquêteurs dans ces mêmes milieux, posent des problèmes additionnels d'anonymisation. Le fait que différents individus enquêtés puissent se connaître et se mentionner les uns les autres, dans les entretiens ou les notes d'observations, suppose une vigilance accrue. Par ailleurs, le fait que les enquêteurs puissent s'exposer, en tant que personne, dans divers documents (notes d'observation, transcriptions d'entretiens, journal d'enquête, etc.) doit être intégré dans les stratégies d'anonymisation.

Au total, l'anonymisation suppose potentiellement 4 couches d'intervention différentes : les données personnelles et sensibles au sens de la loi, les éléments de la vie privé de manière générale, mais aussi les propos ou descriptions pouvant tomber sous le coup de la loi, voire les informations personnelles, intimes, portant sur les enquêteurs.

3) La dimension hétérogène des corpus traités

Les différences d'une enquête à une autre, ou à l'intérieur d'une même enquête, peuvent être importantes s'agissant des éléments à anonymiser. Par exemple, pour une enquête donnée, certains éléments sensibles (opinions, déclarations ou observations de pratiques) sont particulièrement importants pour la recherche en questions, ce qui peut amener à souhaiter conserver la précision de ces éléments, et implique en conséquence de diminuer la précision des identifiants indirects. Dans d'autres cas, ce sont les informations sociographiques sur les enquêtés, rangées dans la catégorie des identifiants indirects, qui seront conservées de manière privilégiée, ce qui implique de supprimer les propos ou observations de pratiques sensibles. Ces choix peuvent se poser à l'intérieur d'un même jeu de données, du fait de l'hétérogénéité, parfois, des composantes d'une enquête de terrain ; les entretiens ou observations peuvent en effet avoir été menés sur des groupes ou des institutions différents, avec des grilles de questionnement hétérogènes, ce qui explique que les protocoles peuvent varier selon ces différentes composantes.

4) La nature des documents supports des corpus

La nature parfois hétérogène des documents de supports des corpus traités introduit de la complexité dans les pratiques d'anonymisation.

- Première situation : s'agissant des corpus préalablement conservés, en tout ou partie, sur des supports physiques (essentiellement papier), l'anonymisation peut être réalisée selon deux modalités : soit a priori sur les supports papiers, avant leur numérisation (cas 1), soit sur les fichiers issus de la numérisation (cas 2).
- Seconde situation : s'agissant des corpus (ou des parties de corpus) nativement numériques, l'anonymisation est effectuée directement sur les fichiers numériques (cas 2).
- Cas 1 : les informations à anonymiser sont masquées par des post-it directement sur les documents papier (manuscrit et ou tapuscrit). L'information est dans ce cas purement et simplement supprimée.
- Cas 2 : les informations à anonymiser sont remplacées par des hyperonymes (technique de la généralisation), ou via l'utilisation de balises de remplacement qui signalent qu'une partie de texte a été modifiée. Aucune technique conduisant à falsifier l'information (ex : randomisation) n'est utilisée.

  NB : lors des passages entiers de textes (descriptions de propos, d'actions, de scènes) sont supprimés, l'indication de la suppression est fournie aux lecteurs directement dans le texte.

Conclusion

Les différentes pratiques présentées dans le cadre de ce document sont aujourd'hui ce qui a paru le plus propice aux ingénieur.e.s du CDSP afin de protéger l'anonymat des personnes interrogées dans le cadre de recherches en sciences sociales.

Cependant, les débats autour de l'anonymisation et l'intérêt que suscite cette technique font que la technologie évolue énormément, et que de nouveaux outils et pratiques sont sans cesse proposés à la communauté scientifique. Le CDSP se tient à jour des évolutions dans le domaine de la protection des données personnelles. Nous aimerions ainsi, par exemple, nous pencher dans les prochains temps vers les techniques de post-randomisation, aussi appelées de confidentialité différentielle (differential privacy), pour les données quantitatives.

La recherche et la mise à jour constante de nos pratiques, afin d'être le plus en phase avec ce qui se fait de mieux concernant l'anonymisation, se reflétera aussi dans les sessions de formation autour des techniques d'anonymisation que le CDSP a commencé à proposer lors de la semaine DATA-SHS 2021.

GLOSSAIRE

Acronymes :

- DPO : *Data protection officer*, délégué.e à la protection des données. Il/elle conseille et oriente les organismes dans leur gestion des données et peut faire office d'intermédiaire avec la CNIL.
- RGPD : Règlement général de protection des données
- CNIL: Commission nationale de l'informatique et des libertés
- DIP - Dissemination Information Package[11]
- SIP - Submission Information Package

Concepts :

- Données personnelles : définies par la CNIL et le RGPD comme "*toute information se rapportant à une personne physique identifiée ou identifiable*".
- Données sensibles : définies par la CNIL et le RGPD comme "les informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale. Ce sont également les données génétiques, les données biométriques aux fins d'identifier une personne physique de manière unique, les données concernant la santé, la vie sexuelle ou l'orientation sexuelle d'une personne physique."
- Données délicates : Au sein des guides sur l'anonymisation créés dans le cadre du WP 1 du projet UPMET porté par le CDSP, , la notion de "données délicates" renvoie à toutes les données qui pourraient porter préjudice aux enquêté.es. Dans cette acception, cette expression se distingue de la notion de "données sensibles" par le fait qu'elle n'est pas encadrée juridiquement.
- K-anonymat : Le K-anonymat est un standard d'anonymisation qui permet de protéger l'identité d'individus en les noyant dans un groupe de personnes aux mêmes caractéristiques. Suivant cette technique, le K est un chiffre qui représente un nombre d'individus. L'objectif de la méthode est de faire en sorte que, parmi toutes les variables sélectionnées comme identifiantes, il n'y ait jamais moins de K personnes ayant les mêmes caractéristiques. Au CDSP nous travaillons pour le moment avec un K=2, ce qui signifie qu'on ne trouvera jamais de personne unique combinant certains traits identifiants.
- l-diversité : La l diversité est encore un pas supplémentaire par rapport au K-anonymat. Elle permet de s'assurer que les personnes partageant les mêmes attributs ne partagent pas une même propriété sensible. Il s'agit donc de sélectionner une variable dite sensible et s'assurer que chaque groupe homogène de personnes présente au moins plusieurs valeurs sur cette variable, et pas seulement la plus

[11] Source: https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais/

sensible (par exemple il faut un positif au Covid, mais aussi un négatif). Le l de l diversité est donc le nombre de valeurs différentes associées au groupe homogène, concernant la variable sensible.

- Variable : une caractéristique mesurable qui peut prendre différentes valeurs. La taille, l'âge, le revenu, la province ou le pays de naissance, les années d'études et le type de logement sont tous des exemples de variables. Pour en savoir plus : https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214817-fra.htm

- Table de passage : Une table de passage/correspondance est composée de 2 colonnes, par exemple : la clé qui représente la valeur d'entrée à vérifier, et la "valeur" qui représente la valeur associée post-traitement.

ANNEXE 2 : OUTILS ET RESSOURCES SUPPLÉMENTAIRES SUR ANONYMISATION DE DONNÉES

Outils d'anonymisation :
En plus des outils déjà mentionnés dans ce document et utilisés par le CDSP, d'autres outils existent :
Amnesia, outil développé dans le cadre d'Open Aire (données quantitatives et qualitatives)
Cornell Anonymization Toolkit (données quantitatives)
IQDA anonymization tool - Guide d'utilisation disponible ici (données qualitatives)
ARX Data Anonymization Tool - Logiciel permettant l'anonymisation par K-anonymat et l-diversité

Ressources sur les techniques d'anonymisation :
> Avis 05/2014 du groupe de travail « article 29 » du 10 avril 2014 sur les Techniques d'anonymisation

>CNIL Anonymisation des données personnelles

>CESSDA ERIC (2018) Processing personal data

>Finnish Social Science Data Archive Anonymization and Identifiers

>Selma Bendjaballah, Sarah Cadorel, Emilie Fromont, Guillaume Garcia, Emilie Groshens, et al.. Anonymat et confidentialité des données : l'expérience de beQuali : L'expérience et les solutions mises en œuvre par beQuali. La diffusion numérique des données en SHS, Presses universitaires de Provence, 2018, 9791032001790. ffhal-02873570f

>AEPD-EDPS joint paper on 10 misunderstandings related to anonymisation | European Data Protection Supervisor. (n.d.). Retrieved March 11, 2022, from https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en

>Bergeat, M. (n.d.). ANONYMISATION DE DONNÉES INDIVIDUELLES : BIEN CALÉES, BIEN PROTÉGÉES ?
19.http://www.jms-insee.fr/2015/S09_2_ACTE_BERGEAT_JMS2015.PDF

>Feten Ben Fredj. Méthode et outil d'anonymisation des données sensibles. Cryptographie et sécurité [cs.CR]. Conservatoire national des arts et metiers - CNAM; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, 2017. Français. ⟨NNT : 2017 CNAM 1128⟩. ⟨tel-01783967⟩

>Les méthodes perturbatives d'anonymisation de données individuelles : avantages et inconvénients, développements récents et exemples de mise en oeuvre (1-SMS_secret 24 juin 2019.pdf)

Ressources juridiques :
Identifier les données à ouvrir : https://guides.etalab.gouv.fr/pdf/guide-juridique.pdf
Ouverture des données de la recherche :
https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Guide_Juridique_V2.pdf