ANONYMISATION/ PSEUDONYMISATION

DÉFINITION

L'anonymisation et la pseudonymisation des données de recherche sont des opérations visant à modifier des sets de données en supprimant ou transformant les données personnelles afin d'empêcher la possibilité d'identifier les personnes qui ont fait l'objet de la recherche. Ces deux opérations visent donc les données dites sensibles, concernant l'identité, la santé, les pratiques culturelles, les opinions politiques/religieuses ou les appartenances sociales des personnes. Elles s'inscrivent dans le cadre des lois sur la protection des données personnelles (LPD / RGPD).

- Dans le premier cas – l'anonymisation – il s'agit d'un processus de transformation des données irréversible. Toutes les données identifiantes (noms, âge, domicile…) ne sont pas enregistrées et, donc, ne font pas partie du set de données. Il n'est alors impossible, même pour les chercheur.e.s qui ont mené la recherche, d'associer une personne précise à un ensemble de données.
- Dans le second cas – la pseudonymisation (ou codage) – les données directement identifiantes sont regroupées dans des fichiers spécifiques, séparés des données exploitées par les chercheur.e.s où elles ont été remplacées par des données indirectement identifiantes. L'anonymat des personnes est donc garanti dans le jeu de données mais leur réidentification est possible car chaque enquêté.e s'est vu attribuer un code alphanumérique permettant, si besoin, de retrouver son identité à partir d'une table de correspondances.

Le lecteur ou la lectrice doit prêter attention au fait que l'on trouve parfois un autre couple de notion pour exprimer la même différence :

- Anonymisation définitive (désignée ci-dessus par le terme « anonymisation »)
- Anonymisation temporaire ou réversible (désignée ci-dessus par le terme « pseudonymisation »)

ANONYMISATION

« L'anonymisation est un traitement de données personnelles qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute réidentification de la personne, par quelque moyen que ce soit » (Commission nationale de l'informatique et des libertés [CNIL], 2022). C'est une opération irréversible dont il faut mesurer les conséquences (nécessaire perte d'informations). N'étant plus identifiantes, les données anonymisées ne sont donc pas soumises aux lois sur la protection des données (LPD / RGPD). Elles peuvent donc être partagées et réutilisées sans restriction et être conservées sans limite de durée à condition que les responsables du traitement préservent, dans le temps, le caractère anonyme des données produites.

Pour anonymiser un set de données, on peut par exemple procéder à :

- La randomisation : modification des attributs dans un jeu de données (en permutant les dates de naissance des individus par exemple).
- La généralisation : regroupement par classes des attributs dans un jeu de données (en remplaçant les dates de naissance par des classes d'âge par exemple).

L'objectif dans les deux cas consiste à brouiller les pistes concernant les données sociogéographiques afin de rendre impossible toute réidentification des individus par corrélation.

PSEUDONYMISATION (CODAGE)

« La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données à une personne physique identifiée sans information supplémentaire » (CNIL, 2022). L'opération consiste donc à remplacer les données directement identifiantes (nom, prénoms, etc.) d'un jeu de données par des données indirectement identifiantes (un code alphanumérique par exemple). C'est donc une opération réversible puisque les informations supprimées dans le set de données sont regroupées dans un document distinct (table de correspondances) pouvant être consulté dans le but de réidentifier les données. Puisqu'il est toujours possible de réidentifier les participant.e.s à l'enquête, les sets de données pseudonymisés (ou codés) sont considérées comme des données personnelles et sont donc soumis à la loi sur la protection des données (LPD / RGPD), notamment en ce qui concerne la durée de conservation et la possibilité pour les personnes concernées d'exercer leurs droits. Le partage et la réutilisation des sets de données pseudonymisées sont soumis à autorisation (notamment l'accès au fichier contenant les données réidentifiantes). Les responsables du traitement préservent, dans le temps, l'accès régulé au fichier de réidentification (table de correspondances).

Pour pseudonymiser un set de données, on procède donc à des opérations des codages :

- En remplaçant des informations identifiantes (nom, prénom) par des nombres aléatoires.
- En chiffrant une partie des données (qui ne sont alors plus lisibles en l'absence de la clef de chiffrement).
- En substituant les données par des informations plus générales (remplacer la commune par le canton par exemple), floues ou fausses (mais sans conséquences pour l'analyse, comme l'attribution d'un pseudonyme par exemple).

L'objectif dans les trois cas consiste à brouiller les pistes concernant les données sociogéographiques afin de rendre impossible toute réidentification des individus sans accès aux tables de correspondances. Il est donc essentiel que les fichiers contenant les informations de réidentification ne soient accessibles qu'aux personnes autorisées et dans des conditions préalablement spécifiées.

Pour connaître la liste des identifiants directs (à partir desquels une personne peut être immédiatement identifiée) et des identifiants indirects (qui peuvent compromettre la confidentialité des données s'ils sont reliés à d'autres sources de données), voir la directive publiée par le Réseau portage (2020).

EXEMPLE DE PSEUDONYMISATION DES ENTRETIENS

Texte d'origine :

« L'année passée j'ai suivi un couple originaire d'Afghanistan. Le mari avait un cancer de Hodgkins, c'était dur. Ils avaient déjà deux enfants et elle était enceinte de jumeaux. Elle a saigné en cours de grossesse et a dû être hospitalisée. Lui était en plein traitement. Affaibli mais quand même au centre pour réfugiés. Il n'arrivait pas à garder vraiment les enfants en l'absence de sa femme. Il dormait beaucoup. Les enfants avaient 3 ans et 18 mois. Elle, elle stressait beaucoup, car il lui envoyait des SMS et disait que cela n'allait pas. Il se sentait aussi malade avec la chimio. Un jour, on a retrouvé les enfants au centre commercial Ikea en face du centre pour réfugiés … ils avaient traversé la grande route tous seuls. Le service social … » (exemple factice - situations à complexité similaire dans les données)

Texte pseudonymisé :

« [Récemment], j'ai suivi un couple originaire du [Moyen Orient]. Le mari avait une [maladie chronique], c'était dur. Ils avaient déjà deux enfants et elle était enceinte de jumeaux. Elle a [eu des complications] en cours de grossesse et a dû être hospitalisée. Lui était […] affaibli mais quand même au centre pour réfugiés. Il n'arrivait pas à garder vraiment les enfants en l'absence de sa femme. Il dormait beaucoup. Les enfants avaient [moins de 4 ans]. Elle, elle stressait beaucoup, car il lui envoyait des SMS et disait que cela n'allait pas. […]. Un jour, [les enfants ont fugué et on les a retrouvés après quelques heures] … ils avaient traversé [une] grande route toute seuls. Le service social … ».

Source : Perrenoud (2021)

RÉFÉRENCES

Commission nationale de l'informatique et des libertés (2019). L'anonymisation des données, un traitement clé pour l'open data. https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data

Commission nationale de l'informatique et des libertés (2020). L'anonymisation de données personnelles. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

Commission nationale de l'informatique et des libertés (2022). Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

Groupe de travail « article 29 » (2014). Avis 05/2014 sur les Techniques d'anonymisation. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_fr.pdf

Perrenoud, P. (2021). Rapport concernant les procédures et processus utilisés pour le partage de données qualitatives sur un DATA repository – Projet Open Data 2020 HES-SO. https://arodes.hes-so.ch/record/9174?ln=fr

Piquette, S. (2021). Anonymisation/pseudonymisation dans les projets de recherche [tutoriel]. Cycle de conférence sur l'éthique de la recherche et la protection des données personnelles. https://pod.univ-lille.fr/video/16591-s17-anonymisation-pseudonymisation/

Réseau portage (2020). Directives sur la dépersonnalisation des données. https://zenodo.org/record/4047176#.Ys5mUTfP02x

Stam, A. & Kleiner, B. (2020). Data anonymization: legal, ethical, and strategic considerations. FORS Guide No. 11, Version 1.0. Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2020-00011

Date de mise à jour	Qui	Changements et commentaire	Version
01.06.2022	Laurent Amiotte-Suchet, HESAV	Création de la guideline	V1
14.02.2023	Jean-Gabriel Piguet, HES-SO VS Tania Zuber-Dutoit , HEIG-VD	Relecture	V1