L’anonymisation et la pseudonymisation : comment les mettre en œuvre pour les données de recherche ?

Pour accéder à la ressource : https://doranum.fr/aspects-juridiques-ethiques/lanonymisation-et-la-pseudonymisation_10_13143_sjqq-hc40/

Date de publication : 07/10/2024

Sommaire

Qu’est-ce que l’anonymisation et la pseudonymisation ? 2
Dans quels cas utiliser la pseudonymisation ou l’anonymisation ? 3
Comment pseudonymiser ou anonymiser des données personnelles ? 3
1. La pseudonymisation 3
1.1. Le compteur 4
1.2. Le générateur de nombres aléatoires 4
1.3. Le chiffrement à clé secrète 5
1.4. La fonction de hachage 5
1.5. Exemples de logiciels de pseudonymisation 5
2. L’anonymisation 6
2.1. La randomisation 6
2.2. La généralisation 7
2.3. Exemples de logiciels d’anonymisation 7
Lexique 9
Autres ressources DoRANum 11
Crédits 12

Depuis la **loi pour une République numérique** (2016), les données de recherche et les codes sources des logiciels achevés, financés au moins pour moitié par des fonds publics et rendus publics, font l'objet d'un **principe d'ouverture par défaut sauf exceptions**.
Les données à caractère personnel* font partie de ces exceptions.
Pour ces données, le chercheur peut être amené à prendre des mesures de pseudonymisation ou d'anonymisation afin de **protéger les personnes identifiées ou susceptibles de l'être** dans son travail de recherche. Ces mesures sont à prendre dès la phase de traitement des données afin de pouvoir, par la suite, les déposer dans un entrepôt et les partager.

À noter que pour les données en santé, des règles supplémentaires sont à respecter : avis du comité de protection des personnes pour les recherches impliquant la personne humaine (RIPH), autorisation de la CNIL, …
Pour en savoir plus, vous pouvez vous référer à l'article de la CNIL paru le 22 mai 2024 : Recherches dans le cadre de la santé : quelles sont les formalités ?

*****renvoi au lexique (p.9)

Qu'est-ce que l'anonymisation et la pseudonymisation ?

L'anonymisation et la pseudonymisation sont deux **traitements** différents qui visent à assurer la **protection des données à caractère personnel**.

- **L'anonymisation** est l'action de retirer toutes les informations d'identification d'une personne d'un ensemble de données. L'anonymisation est une mesure **irréversible**. Les données anonymisées ne sont plus des données personnelles. Par conséquent, elles **ne relèvent plus du RGPD** (règlement général sur la protection des données).
- **La pseudonymisation** est l'action de remplacer les informations d'identification d'une personne (nom, prénom, etc.) par des données **indirectement identifiantes*** (alias, numéro de classement, etc). La pseudonymisation est une mesure **réversible**. C'est pour cela que les données

pseudonymisées demeurent des données personnelles. Leur traitement*** reste soumis aux obligations du RGPD**.

En résumé, la différence entre l'anonymisation et la pseudonymisation tient dans le **caractère réversible ou non de l'identification d'une personne**.

Dans quels cas utiliser la pseudonymisation ou l'anonymisation ?

Pour les deux traitements des données, la CNIL (Commission nationale de l'informatique et des libertés) recommande aux chercheurs de rédiger un **protocole** en concertation avec un délégué à la protection des données (DPD).

- La pseudonymisation est préconisée lorsque le chercheur a besoin d'avoir des **informations exactes au niveau individuel** sans pour autant avoir accès aux données directement identifiantes.

- L'anonymisation étant une « technique destructrice d'information », la CNIL conseille au chercheur de l'utiliser lorsqu'il dispose d'une **grande quantité de données,** pour une **utilisation précise** et avec une méthode qui **préserve les propriétés du jeu de données**.

Si le chercheur choisit la pseudonymisation, l'ouverture de ses données sera plus restreinte que s'il choisissait l'anonymisation. En effet, la pseudonymisation étant un traitement moins protecteur, certaines données ne seront pas partageables.
Le chercheur peut opter pour l'anonymisation afin de partager de manière ouverte ses jeux de données.

Comment pseudonymiser ou anonymiser des données personnelles ?

1. La pseudonymisation

Il existe plusieurs techniques de pseudonymisation. La CNIL recense celles qui reposent sur la création de pseudonymes basiques (compteur et générateur de

nombre aléatoire) et celles qui s'appuient sur des techniques cryptographiques (chiffrement à clé secrète et fonction de hachage).
Attention : les pratiques de substitution, généralisation et floutage sont considérées comme un premier niveau de protection mais ne sont pas des techniques de pseudonymisation au sens du RGPD. Ces pratiques consistent à remplacer des données directement identifiantes par d'autres données fictives. Par exemple, pour les enquêtes scientifiques où il est important de garder une ressemblance socio-culturelle (choix des prénoms, profession, lieu de vie, etc.) tout en protégeant l'identité des personnes.

1.1. Le compteur

Il s'agit de substituer un nombre, généré par un compteur, à une ou plusieurs données directement identifiantes (comme un nom ou une date de naissance). Le compteur débute à un nombre défini et s'incrémente à chaque nouvel enregistrement afin de garantir que chaque pseudonyme soit unique.

Données initiales			Résultat
Nom	Prénom	Date de naissance	Pseudonyme
Dupont	Henri	27/04/1942	254
Grand	Marine	05/09/2002	255
Paris	Mathilda	12/03/2006	256

Cette technique est simple et adaptée aux jeux de données de petite taille, et qui ne sont pas complexes.

1.2. Le générateur de nombres aléatoires

Il s'agit de remplacer les données directement identifiantes par des nombres générés aléatoirement. Le pseudonyme n'indique pas d'information sur l'ordre des données dans un jeu de données.

Données initiales			Résultat
Nom	Prénom	Date de naissance	Pseudonyme
Dupont	Henri	27/04/1942	514
Grand	Marine	05/09/2002	89210
Paris	Mathilda	12/03/2006	3315

1.3. Le chiffrement à clé secrète

Il s'agit de chiffrer les données directement identifiantes pour les rendre incompréhensibles sans la clé secrète permettant de réidentifier les personnes.

Données initiales			Résultat
Nom	Prénom	Date de naissance	Pseudonyme
Dupont	Henri	27/04/1942	e29843178f52fb577986…
Grand	Marine	05/09/2002	806e37f131008057776f…
Paris	Mathilda	12/03/2006	55912722ffa374bce6320…

1.4. La fonction de hachage

Il s'agit d'une fonction mathématique qui produit une liste de caractères et de chiffres pour remplacer les données directement identifiantes. Cette fonction est conçue pour ne pas être inversée. Il est donc impossible de récupérer facilement les données, comme dans le cas du chiffrement. En effet, les mots de passe ne sont pas stockés tels quels mais ils sont remplacés par leur « empreinte ».

Données initiales			Résultat
Nom	Prénom	Date de naissance	Pseudonyme
Dupont	Henri	27/04/1942	eeff86b4a738a90c0dfaa…
Grand	Marine	05/09/2002	906cfc57fdc8e489f4383…
Paris	Mathilda	12/03/2006	ab9ef82e90636a9ba351…

1.5. Exemples de logiciels de pseudonymisation

Pour l'usage d'un logiciel, l'utilisateur doit se tourner vers le Service Protection des Données désigné par son laboratoire et le chargé de la sécurité des systèmes d'information afin d'évaluer la conformité de l'outil avec la règlementation en vigueur.

Etalab (département de la direction interministérielle du numérique, DINUM) a développé **un outil d'intelligence artificielle de pseudonymisation** pour le Conseil d'État. Cet outil est open-source, il peut être librement réutilisé pour d'autres projets de pseudonymisation.

Etalab propose dans son guide "Pseudonymiser des documents grâce à l'IA", des **ressources disponibles pour pseudonymiser**.

2. L'anonymisation

Avant de mettre en œuvre une mesure d'anonymisation, qui a pour conséquence une perte d'information, la CNIL recommande de :

- « **Examiner les catégories de données** à anonymiser […] ;
- **Supprimer les éléments d'identification directe ainsi que les valeurs rares** qui pourraient permettre une réidentification aisée des personnes (par exemple, la connaissance précise de l'âge des individus présents dans un jeu de données peut permettre dans certains cas de réidentifier très facilement les personnes centenaires) ;
- **Distinguer les informations importantes des informations secondaires ou inutiles** (c'est-à-dire supprimables, qu'il est préférable de ne pas collecter du tout en vertu du principe de minimisation***** des données) ;
- **Définir la finesse** idéale et acceptable pour chaque information conservée ;
- **Définir les priorités** (exemple : est-il plus important de conserver une grande finesse sur telle information ou de conserver telle autre information ?) ».

Ce questionnement aide à choisir la technique d'anonymisation la plus pertinente.

Il existe deux catégories de techniques d'anonymisation : la randomisation ou la généralisation.

2.1. La randomisation

Il s'agit de changer les attributs dans un jeu de données (ex : date de naissance) afin que les données soient moins précises, tout en gardant la répartition globale.

Cette technique protège le jeu de données du **risque d'inférence*** c'est-à-dire qu'elle rend impossible la déduction de nouvelles informations sur une personne à partir du jeu de données.

Exemple de technique de randomisation : la permutation.

Données initiales			Données « randomisées »
Nom	Prénom	Date de naissance	L'année de naissance est permutée
Dupont	Henri	27/04/1942	05/09/2002
Grand	Marine	05/09/2002	12/03/2006
Paris	Mathilda	12/03/2006	27/04/1942

2.2. La généralisation

Il s'agit de généraliser les attributs dans un jeu de données en changeant leur échelle ou leur ordre de grandeur pour qu'ils soient communs à un ensemble de personnes ; exemple : remplacer les dates de naissance par « personnes âgées de 30 à 40 ans ».

Cette technique protège le jeu de données des **risques d'individualisation* et de corrélation*** du jeu de données avec d'autres jeux.

Exemple de technique de généralisation : l'agrégation.

Données initiales			Données « généralisées »
Nom	Prénom	Date de naissance	L'année de naissance est remplacée par un ordre de grandeur
Dupont	Henri	27/04/1942	Personnes âgées de 80 à 90 ans
Grand	Marine	05/09/2002	Personnes âgées de 20 à 30 ans
Paris	Mathilda	12/03/2006	Personnes âgées de 20 à 30 ans

Pour en savoir plus sur les techniques d'anonymisation, consultez l'avis 05/2014 sur les Techniques d'anonymisation du Groupe de travail "article 29" sur la protection des données : https://www.cnil.fr/sites/cnil/files/atoms/files/wp216_fr.pdf

2.3. Exemples de logiciels d'anonymisation

Rappel : pour l'usage d'un logiciel, l'utilisateur doit se tourner vers le Service Protection des Données désigné par son laboratoire et le chargé de la sécurité des systèmes d'information afin d'évaluer la conformité de l'outil avec la règlementation en vigueur

- **Amnesia**

Cette solution open source et gratuite est proposée par OpenAIRE. Elle permet de partager les résultats de recherche conformes au RGPD grâce à des algorithmes d'anonymisation des données.

- **Anonymization ToolBox**

Cet outil open source donne accès à des outils qui réalisent six types d'anonymisation différents.

Pour en savoir plus sur les logiciels, consultez la fiche pratique de l'université de Lorraine « Anonymiser ses données : quelques ressources ».

Afin d'éviter tout risque de violation de données*, la CNIL conseille aux responsables de traitement* de **réaliser une veille régulière sur les techniques d'anonymisation et de réidentification**. Le maintien à jour des techniques de traitement permet d'assurer, dans le temps, le caractère anonyme des données produites.

Lexique

Corrélation : « Capacité de relier entre eux, au moins, deux enregistrements se rapportant à la même personne concernée ou à un groupe de personnes concernées (soit dans la même base de données, soit dans deux bases de données différentes) » (avis 05/2014 sur les Techniques d'anonymisation).

Donnée à caractère personnel : « Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un nom, un numéro d'identification (par exemple le numéro de sécurité sociale) ou à un ou plusieurs éléments qui lui sont propres » (guide Etalab).

Données personnelles directement identifiantes : prénom et nom, coordonnées postales ou téléphoniques (y compris professionnelles), âge, sexe, date et lieu de naissance, nationalité, voix d'une personne dans le cadre d'un entretien réalisé pour une enquête, coordonnées électroniques, …

Données personnelles indirectement identifiantes : numéro de sécurité sociale, régime matrimonial, coordonnées bancaires, diplômes, emplacement géographique, adresse IP, témoins de connexion (cookies), puces RFID.

Données personnelles sensibles :
Il s'agit de « [données] qui révèlent la prétendue origine raciale ou l'origine ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale d'une personne physique ou […] des données génétiques, des données biométriques [permettant] 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 » (art.6 de la loi informatique et libertés).
Il s'agit également des données personnelles de santé : « données relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de la santé de cette personne » (art.4 du RGPD).
La collecte et le traitement de ces données est interdit sauf pour la recherche scientifique.

Individualisation : « Correspond à la possibilité d'isoler une partie ou la totalité des enregistrements identifiant un individu dans l'ensemble de données » (avis 05/2014 sur les Techniques d'anonymisation).

Inférence : « Possibilité de déduire, avec un degré de probabilité élevé, la valeur d'un attribut à partir des valeurs d'un ensemble d'autres attributs » (avis 05/2014 sur les Techniques d'anonymisation).

Principe de minimisation des données : « Le principe de minimisation prévoit que les données à caractère personnel doivent être adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées. Exemple : Collecter et conserver le statut marital d'un salarié n'apparaît pas nécessaire à l'activité RH ». (CNIL).

Responsable du traitement : la CNIL définit le responsable du traitement comme « la personne morale (entreprise, commune, etc.) ou physique qui détermine les finalités et les moyens d'un traitement, c'est à dire l'objectif et la façon de le réaliser. En pratique et en général, il s'agit de la personne morale incarnée par son représentant légal ».

Traitement des données personnelles : « Un traitement de données personnelles est une opération, ou ensemble d'opérations, portant sur des données personnelles, quel que soit le procédé utilisé (collecte, enregistrement, organisation, conservation, adaptation, modification, extraction, consultation, utilisation, communication par transmission ou diffusion ou toute autre forme de mise à disposition, rapprochement). Un traitement de données personnelles n'est pas nécessairement informatisé : les fichiers papier sont également concernés et doivent être protégés dans les mêmes conditions.
Un traitement de données doit avoir un objectif, une finalité déterminée préalablement au recueil des données et à leur exploitation.
Exemples de traitements : tenue du registre des sous-traitants, gestion des paies, gestion des ressources humaines, etc. ». (CNIL).

Violation de données : est « une violation de la sécurité entraînant, de manière accidentelle ou illicite, la destruction, la perte, l'altération, la divulgation non

autorisée de données à caractère personnel transmises, conservées ou traitées d'une autre manière, ou l'accès non autorisé à de telles données ». (art. 4.12 du RGPD).

Autres ressources DoRANum

Retrouvez d'autres ressources sur les sites :

DoRANum : Données de la recherche, apprentissage numérique : des ressources pour accompagner la communauté scientifique dans la gestion et le partage de leurs données.

Chaîne Canal U DoRANum : DoRANum est un service de formation à distance sur la thématique de la gestion et du partage des données de la recherche. Retrouvez ici les vidéos issues de ces formations.

Suivez-nous sur :

X DoRANum

Crédits

- Bouchet-Moneret, Florence, Bracco, Laetitia, Jouneau, Thomas. **Anonymiser ses données : quelques ressources**. Atelier de la donnée ADOC Lorraine. [en ligne]. [cité le 3 septembre 2024]. Disponible sur : https://zenodo.org/records/5717856
- CNIL. **L'anonymisation de données personnelles**. [en ligne]. 2020 [cité le 27 août 2024]. Disponible sur : https://www.cnil.fr/fr/technologies/lanonymisation-de-donnees-personnelles
- CNIL. **Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation**. [en ligne]. 2022 [cité le 28 août 2024]. Disponible sur : https://www.cnil.fr/fr/recherche-scientifique-hors-sante-enjeux-et-avantages-de-lanonymisation-et-de-la-pseudonymisation
- CNIL, CADA, Etalab. **Guide pratique de la publication en ligne et de la réutilisation des données publiques** (« Open data »). Présentation du cadre juridique de l'ouverture des données. [en ligne]. [cité le 28 août 2024]. Disponible sur : https://www.cnil.fr/sites/cnil/files/atoms/files/guide_open_data.pdf
- DoRANum. **Comprendre la science ouverte**. [en ligne]. 2024 [cité le 27 août 2024]. Disponible sur : https://callisto-formation.fr/pluginfile.php/13538/mod_scorm/content/9/scormcontent/index.html#/
- Etalab. Les guides d'Etalab. **Pseudonymiser des documents grâce à l'IA** [en ligne]. 2024 [cité le 27 août 2024]. Disponible sur : https://guides.etalab.gouv.fr/pseudonymisation/pourquoi-comment/
- Robin, Agnès. **Droit des données de la recherche : Science ouverte, innovation, données publiques**. Bruxelles : Larcier ; 2022.