UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Glossaire – Données personnelles de recherche

Les définitions figurant dans ce glossaire n'ont pas de valeur juridique. Ce glossaire a pour but de faciliter la compréhension des termes utilisés dans le domaine de la protection des données.

Les explications relatives à la pseudonymisation et l'anonymisation sont génériques. Des jeux de données particuliers peuvent requérir des procédés techniques différents et plus complexes.

Sommaire

Données personnelles 2
Traitement de données personnelles 3
Données personnelles sensibles 4
Données personnelles pseudonymisées 5
Données anonymisées 6

Données personnelles

Les données personnelles désignent toute information, peu importe sa forme, son contenu ou son support, qui se rapporte à une personne identifiée ou identifiable. Dans les projets de recherche, la personne identifiée ou identifiable est souvent appelée « participant-e au projet ».

Une personne est réputée identifiée lorsque son identité ressort directement de l'information détenue. Une personne est identifiable si la corrélation d'informations tirées des circonstances ou du contexte permet son identification. Ainsi, même si aucune donnée directement identifiante (ex. nom, prénom, date de naissance) n'est récoltée, il est possible que selon les circonstances, les données récoltées et les circonstances, le contexte, les métadonnées permettent d'identifier des personnes. Il s'agirait alors tout de même de données personnelles.

Des données qui se rapportent à une personne identifiable ou identifiée, sont aussi des données personnelles même si elles n'ont pas été récoltées directement auprès de la personne concernée, p. ex. lorsque les données ont été récoltées par des collègues d'une autre institution et qu'elles sont désormais réutilisées pour un nouveau projet, ou lorsque les données sont collectées en ligne, p. ex. sur les réseaux sociaux.

Les données personnelles peuvent prendre toute forme : donnée textuelle, donnée audio, image, enregistrement vidéo, …

Exemples de données personnelles : le nom, la date de naissance, l'adresse e-mail, le sexe, les réponses des participant-e-s à un questionnaire en ligne, les données observables des participant-e-s à une étude expérimentale la photographie, la voix, les caractéristiques d'identification (par exemple, seule femme dans la cohorte de participant-e-s), le code de pseudonymisation attribué à un·e participant·e dans une étude, …

Traitement de données personnelles

Toute opération relative à des données personnelles, quels que soient les moyens et procédés utilisés : la collecte, l'enregistrement, la conservation, la pseudonymisation, l'analyse, la modification, la communication, l'externalisation, l'effacement, l'archivage ou la destruction, …

Le traitement de données personnelles par les chercheur·euse·s au sein de l'Université est soumis à la Loi sur la protection des données du canton de Fribourg. En fonction des spécificités du projet et des collaborations avec des partenaires, d'autres réglementations sur la protection des données pourraient s'appliquer, p.ex. la Loi fédérale sur la protection des données ou le Règlement général sur la protection des données de l'Union européenne.

Il incombe aux chercheur·euse·s de veiller à ce que l'ensemble du traitement des données dans le cadre de leurs projets de recherche soit conformes à la réglementation applicable en matière de protection des données. Sur demande, les DPO les soutiennent.

Données personnelles sensibles

Les données personnelles « sensibles » sont une sous-catégorie des données personnelles. Sont des données personnelles sensibles :

- les données sur les opinions ou les activités religieuses, philosophiques, politiques ou syndicales;
- les données sur la santé, la sphère intime ou l'origine raciale ou ethnique;
- les données génétiques;
- les données biométriques identifiant une personne physique de manière univoque;
- les données sur des mesures d'aide sociale;
- les données sur des poursuites ou des sanctions pénales et administratives.

Cette liste est exhaustive.

En raison de leur nature ou de leur fonction, le traitement des données personnelles sensibles présente un risque accru pour la personnalité et les droits fondamentaux des personnes concernées. Ainsi, elles exigent un niveau de protection et de sécurité plus élevé.

Au sens de la loi suisse, la catégorie des participant-e-s à la recherche n’est pas un critère pour déterminer si les données sont sensibles. Mais suivant les circonstances, un devoir de diligence accru peut être nécessaire pour des raisons d’éthique de la recherche lors du traitement de données qui se rapportent à des enfants, des personnes en situation de handicap, en situation de précarité, des personnes incarcérées, des personnes dans l’incapacité de donner leur consentement, des étudiant-e-s lorsque le ou la responsable de la recherche est également leur enseignant-e, ou d’autres catégories de personnes considérées comme vulnérables.

Données personnelles pseudonymisées

La pseudonymisation est un procédé qui consiste à **substituer les données permettant l'identification d'une personne (généralement le nom) par un code** neutre afin qu'elles ne puissent plus être reliées à une personne spécifique sans informations supplémentaires ou sans efforts disproportionnés**.**

Les données qui sont pseudonymisées ou codées par le chercheur ou la chercheuse après la récolte sont toujours considérées comme des données personnelles dont le traitement est soumis à la réglementation applicable en matière de protection des données.

Pourquoi pseudonymiser des données personnelles de recherche ?

C'est une mesure de sécurité des données pour réduire le risque de mise en corrélation des données avec l'identité de la personne concernée. On assure ainsi un meilleur niveau de protection des participant-e-s à la recherche. La pseudonymisation des données se fait généralement dès la récolte des données et avant leur analyse.

Pour que la sécurité soit effective :

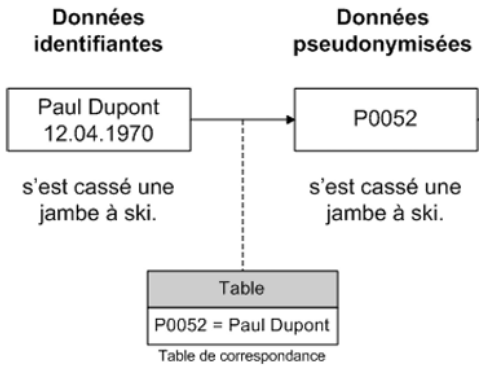
- Stocker séparément les données pseudonymisées et la clé de codage (dans des dossiers séparés avec des droits d'accès distincts)
- Restreindre l'accès à la clé de codage à un minimum de personnes, p. ex. seulement le ou la responsable du projet de recherche. Les données pseudonymisées peuvent rester accessibles à l'ensemble de l'équipe de recherche.
- Les données brutes, contenant les données permettant l'identification, devraient être supprimées dès que la pseudonymisation a été effectuée, seul les données pseudonymisées et la clé de codage étant conservés séparément.

La pseudonymisation peut aussi faciliter le partage des données avec une institution tierce pour la réutilisation dans d'autres projets de recherche.

Processus

- Création d'un code qui remplace typiquement le nom, la date de naissance, etc. de la personne concernée dans les bases de données.
- Simultanément, une clé de codage (appelée aussi table de correspondance) est créée afin de lier le nom à ce code. Ainsi, seules les personnes ayant accès à cette clé de codage peuvent aisément lier les données avec le nom, et de ce fait renverser la pseudonymisation

Le processus est réversible tant que la clé de codage n'est pas éliminée. Mais l'opération ne doit pas être réversible sans la clé de codage.



Source de l'image : Guide relatif aux TOM du Préposé fédéral à la protection des données et à la transparence, janvier 2024, disponible ici, p. 24.

Données anonymisées

L'anonymisation est un procédé technique qui consiste à **modifier de façon irréversible** les données personnelles de sorte à ce qu'elles **ne puissent plus être liées à une personne spécifique**, y compris par l'auteur de la collecte et du traitement, sans efforts disproportionnés. En d'autres termes, des données sont anonymisées si la réidentification ne paraît possible qu'au prix d'efforts disproportionnés.

Pourquoi anonymiser des données

- Les données anonymisées ne sont plus considérées comme des données personnelles. Les exigences de la réglementation sur la protection des données ne trouve plus application. Mais, le procédé d'anonymisation des données est considéré comme un traitement de données personnelles et est encore soumis à la législation sur la protection des données.
- Les données anonymisées pourront être conservées sans limitation de durée.
- Cela facilite la réutilisation des données et leur partage pour d'autre projets.
- Cela facilite la publication des données selon les principes FAIR.

Risque de réidentification

Pour déterminer si une réidentification est possible sans efforts disproportionnés, il faut tenir compte de **l'ensemble des moyens raisonnablement susceptibles d'être utilisés et l'ensemble des circonstances**. Il faut apprécier concrètement, pour le jeu de données concerné :

- la nature identifiante ou non des données
- le risque de réidentification
- les moyens raisonnablement susceptibles d'être mobilisés pour permettre l'identification d'une personne. Il faut tenir compte de critères objectifs, incluant notamment le coût et le temps nécessaires à l'identification, eu égard aux technologies disponibles au moment du traitement ainsi que de leur évolution prévisible.

Quelques circonstances à prendre en compte quant au **risque de réidentification** :

- Une personne peut être identifiable à partir de nombreux éléments qui la distinguent d'une autre, et pas seulement par son nom. Même si une seule donnée ne permet pas d'identifier une personne, la combinaison de différentes données collectées peut permettre de l'identifier.

  Exemple : par exemple, dans des données relatives à une étude fribourgeoise sur le suivi post-opératoire oncologique : [ID 8128136, Femme, 42 ans, habite (petit village), traitée pour un cancer du sein, date d'opération 12.07.2021] ne remplit pas les conditions d'anonymité, car cette personne pourrait être identifiée sans efforts disproportionnés.
- Vérifier si la combinaison de données peut conduire à l'identification de personnes lorsque des recherches très précises sont effectuées (technique de ciblage). Ainsi, il est préférable de remplacer l'âge ou l'année de naissance par une fourchette d'âge.
- Vérifier si la combinaison de données peut conduire à l'identification de personnes lorsqu'une collecte de données est effectuée auprès d'un cercle fermé de personnes, par exemple les collaborateurs et collaboratrices d'une certaine entreprise, les élèves d'une certaine école, les habitant·e·s d'une certaine commune, etc.

- Un risque important de réidentification existe également lorsque différents jeux de données concernent les mêmes personne sont publiés. Il faut alors vérifier si leur croisement peut conduire à l'identification des personnes. C'est aussi le cas si le jeux de données publié peut facilement être croisé avec des données publiques par exemple.

Processus et techniques d'anonymisation

Chaque jeu de données demande une réflexion individuelle, et certains jeux ne peuvent tout simplement pas être rendus anonymes sans perdre leur utilité.

Procédé standard

- Supprimer les données directement identifiantes (nom, prénom, …).
- Supprimer les valeurs rares qui permettraient une ré-identification (ex. lieu de résidence à l'étranger si c'est le cas pour un seul ou seulement quelques participant-e-s)
- Distinguer les informations importantes qui doivent être conservées pour que les données gardent leur utilité pour une réutilisation dans la recherche de celles qui sont supprimables. Supprimer ces dernières.
- Définir la précision de la donnée acceptable pour chaque information qui sera conservée et remplacer certaines valeurs d'attributs par des valeurs plus génériques ou des marges (généralisation) :

  P. ex. remplacer une certaine date de naissance (ex. 8.03.1980) par l'année de naissance (ex. 1980), ou encore par une marge d'âge (ex. [40-50] ans).

  P. ex. remplacer une certaine adresse exacte par sa ville, région ou canton (on peut très bien hiérarchiser).

  P. ex. remplacer une certaine nationalité par sa région géographique, ou continent.

!! Le seul remplacement des données directement identifiantes par un code n'équivaut pas à une anonymisation des données. Ce procédé correspond à pseudonymiser les données. C'est une étape seulement du procédé d'anonymisation.

D'autres techniques doivent également être prises en compte selon le jeu de données : randomisation (ajout de bruit, permutation de données, confidentialité différentielle), création de données synthétiques, anonymisation des personnes dans des enregistrements vidéo ou sur des photographies, anonymisation des voix sur des enregistrements audio ou vidéo, etc. Ces techniques nécessitent généralement le soutien d'un *data scientist*.

Le chercheur ou la chercheuse en charge du projet est en charge de mener l'appréciation concrète du caractère anonyme du jeu de données, une fois le procédé mis en œuvre.