

[Symposium « Anonymisation des données »](#)

Chaire L.R. Wilson

# L'anonymisation des données de recherche

*les bibliothèques*

Université   
de Montréal

Stéphanie Pham-Dang

Bibliothécaire | Gestion de données de recherche

Direction du soutien à la réussite, la recherche et l'enseignement

29 avril 2024



# Plan de présentation

1

Mise en contexte :  
la gestion des données  
de recherche  
en milieu universitaire

2

~~Enjeux~~  
Défis relatifs à  
l'anonymisation  
de données sensibles  
de recherche

3

~~Normes~~  
Pratiques actuelles,  
techniques et outils  
disponibles

# 1 Contexte

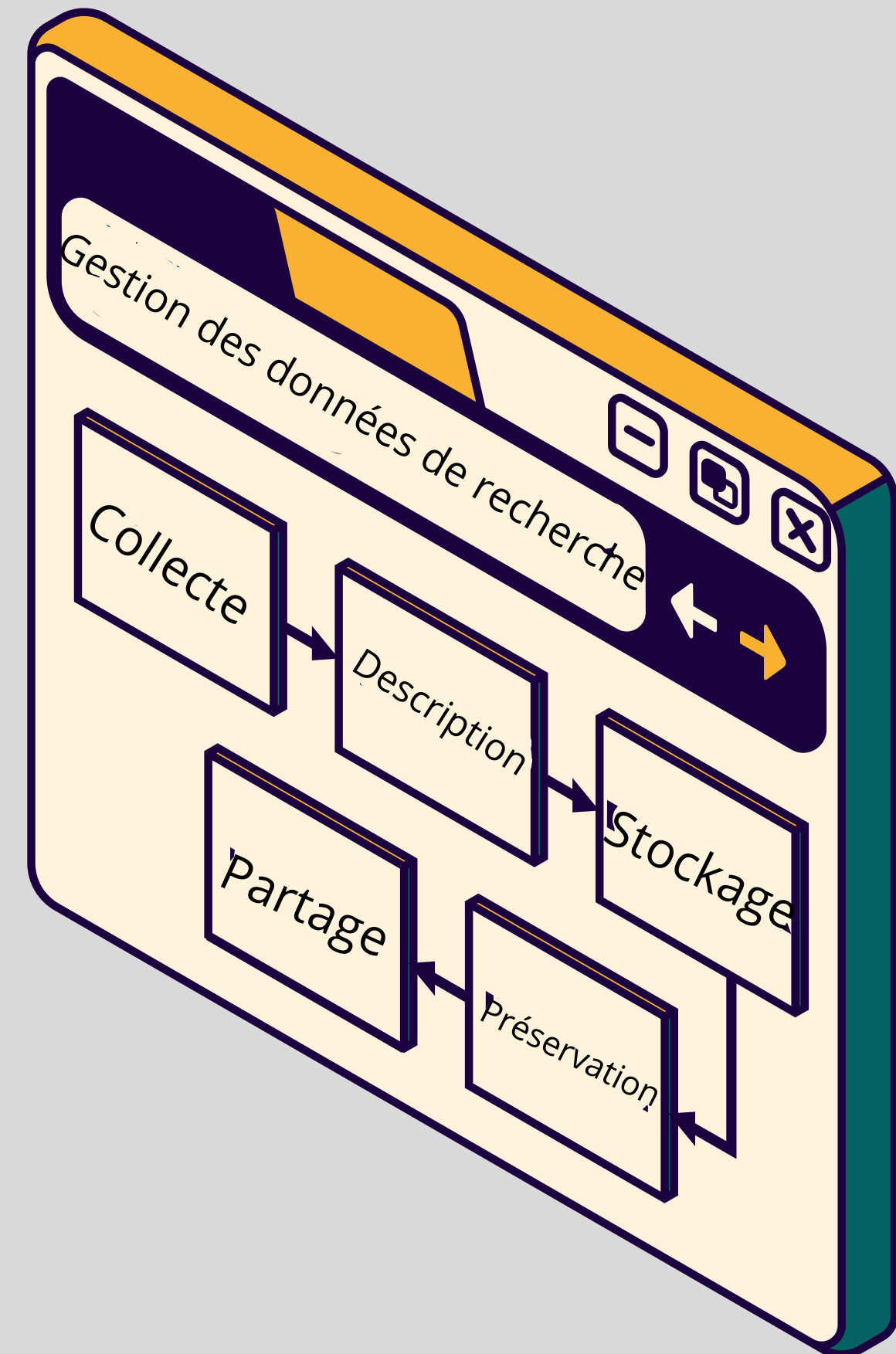
## La gestion des données de recherche (GDR) en milieu académique

“

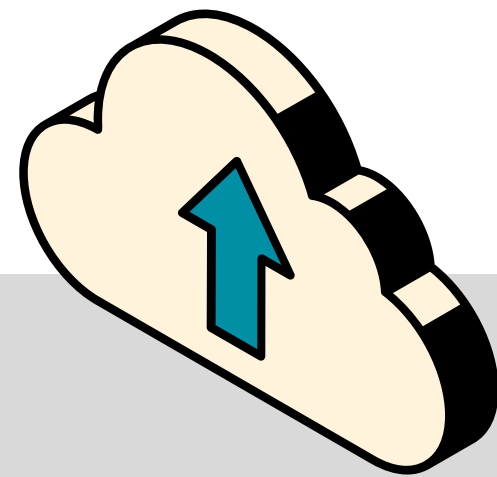
Comprend les processus utilisés tout au long du cycle de vie d'un projet de recherche pour orienter la collecte, la [description], le stockage, le partage et la préservation des données de recherche.

”

*Alliance de recherche numérique du Canada*

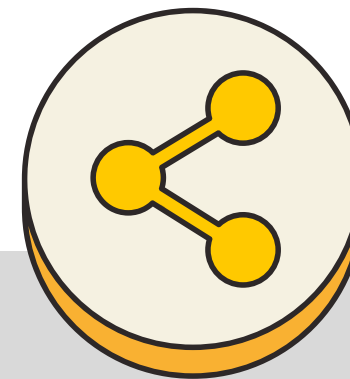


# Gestion des données sensibles de recherche

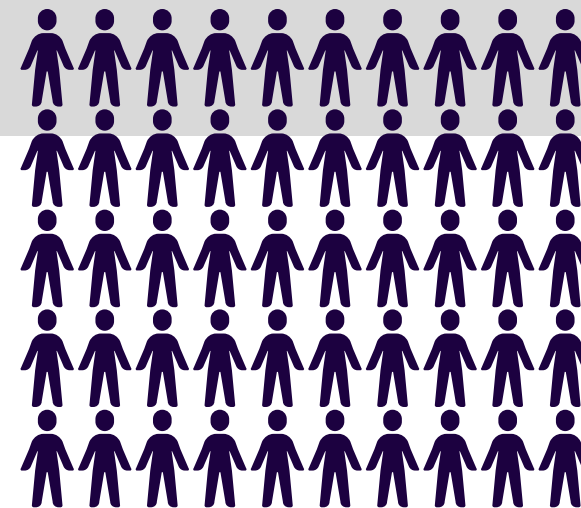


Stockage  
des données actives  
durant la recherche

“non publiques”



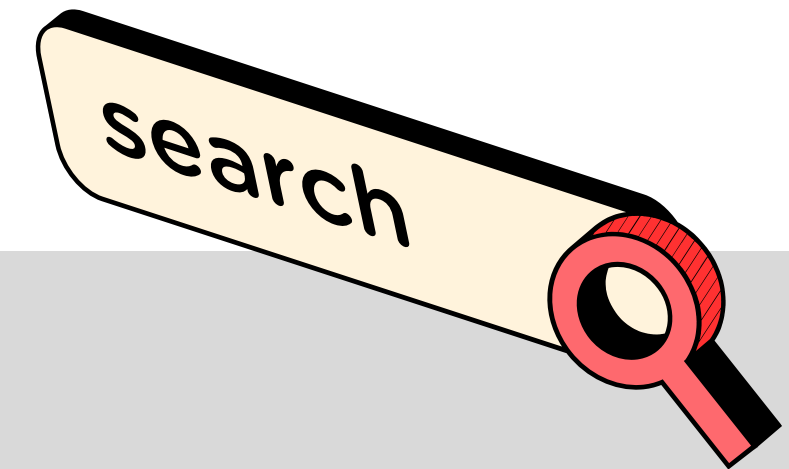
Partage  
des données finales  
une fois la recherche terminée



“publiques”?

“ouvertes”?

FAIR



Réutilisation  
des données de recherche repérées  
dans un dépôt de données

# Chronologie

**2021**

**Politique des trois organismes subventionnaires fédéraux sur la gestion des données de recherche**

- Instituts de recherche en santé du Canada (IRSC)
- Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG)
- Conseil de recherches en sciences humaines (CRSH)

**2022**

**Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains – EPTC 2**

- Chapitre 5 : Respect de la vie privée et confidentialité
- Les chercheurs
  - Les comités d'éthique de la recherche (CER)

**Février 2023**

**Stratégie institutionnelle pour la gestion des données de recherche (V1.1)**

Université de Montréal

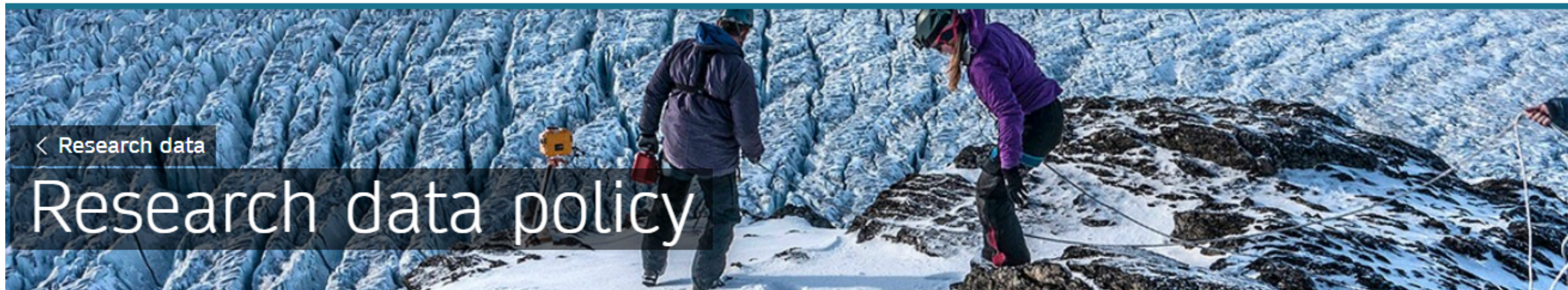
- Pour le cycle de vie complet des données de recherche
- Services-conseils
  - Infrastructures numériques

**Septembre 2023**

**Entrée en vigueur de la Loi 25 au Québec**

Mise en ligne du site web [vie-privee.umontreal.ca](http://vie-privee.umontreal.ca)

Nouvelles politiques et directives de PRP en recherche, notamment les évaluation de facteurs de risques à la vie privée (ÉFVP)



< Research data

# Research data policy

## Research data policy

Data availability statements

Data repository guidance

Sensitive data

Data policy FAQs

Research data helpdesk

## Sensitive data

Authors with sensitive data, or other data that cannot be shared openly, should apply appropriate restrictions before sharing their data. Authors should put their data in a repository where possible and only fully restrict data access if no other sharing option is available.

Putting data in a repository is the best option for transparency, long-term storage and managing access requests.

### Sensitive or restricted data includes:

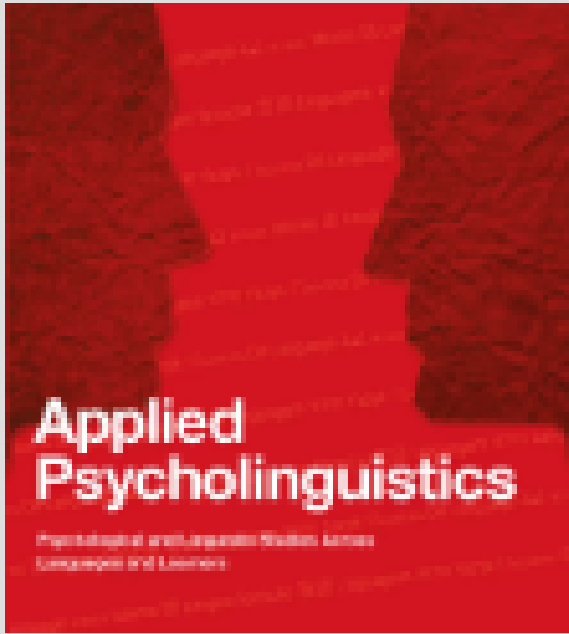
- 1 **Identifiable human data:** data involving human research participants may present a risk of reidentification if shared openly. This includes both quantitative and qualitative research data. A list of identifiers in human data is available.
- 2 **Other sensitive data:** non-human data sensitivities should also be considered, from locations of endangered



### Sensitive Data | Publish your research | Springer Nature

Authors with sensitive data, or other data that cannot be shared openly, should apply appropriat...

springernature.com



## Preparing your materials

Welcome to Cambridge Core

cambridge.org



## Openness, Transparency, and Reproducibility Policy

### Author checklist

- Confirm that your replication package tells readers where public and free access to the complete (1) study materials, (2) analysis code, and (3) data can be found.
- All replication package materials are freely and publicly available. No registration, login, or request is required to access them.
- Confirm that all links provided in the replication package statement are functional.
- Confirm that you have reviewed all of the below information and suggestions for replication package language.
- Failure to meet the replication package requirement will result in automatic administrative rejection of your manuscript.

### Data

Data refers to the full unaggregated data set(s) in .csv format. Data must be anonymized for participant privacy.

### Analysis code

Analysis code includes complete instructions and code for conducting all analyses reported in the manuscript, and must be applicable to the raw data files described above.

### Suggested language

Ideally, all research materials, data, and analysis code will be stored together in a single repository. In this case, the suggested language is:

*All research materials, data, and analysis code are available at [insert link].*

### FAQs

#### What is a trusted online repository?

Trusted online repositories include the [Open Science Framework](#), [Dataverse](#), a university repository, or other database on the [Registry of Research Data Repositories](#).

# Gestion des données sensibles de recherche

<https://stockage-recherche.umontreal.ca>

Université  de Montréal | La recherche

## Solutions de stockage pour les données de recherche

Cet outil interactif vous permet de rechercher une solution de stockage ou d'archivage de données de recherche dont l'utilisation sécuritaire a été validée par les T.I de l'UdeM. La liste des solutions ci-dessous sera bonifiée au fil de l'évolution de l'offre de services de stockage pour la recherche et des politiques de l'université.

### Questions

Réinitialiser

#### 1. Quel est le niveau de confidentialité de vos données?

- Données non confidentielles publiques à risque faible
- Données non confidentielles internes à risque modéré
- Données confidentielles à risque élevé
- Données hautement confidentielles à risque critique ou majeur

#### 2. Avez-vous des fonds disponibles pour payer pour le stockage?

### Services

Tout sélectionner

Effacer les sélections

Borealis -  
Dépôt  
Dataverse  
canadien

Partager et chercher  
des données de  
recherche  
canadienne

Calcul  
Québec/Alliance  
de recherche  
Numérique  
du Canada

Stockage, calcul  
haute performance,  
logiciels de  
recherche

DocUM -  
Recherche

Sauvegarde de  
fichiers et de  
données pour les  
équipes de recherche

Dépôt fédéré  
de données  
de recherche  
Canadienne  
(DFDR)

Partager et chercher  
des données de  
recherche  
canadienne

Hébergement  
de serveurs  
virtuels

Sauvegarde de  
données, application  
BD et opérations de  
calcul

Onedrive  
Entreprise

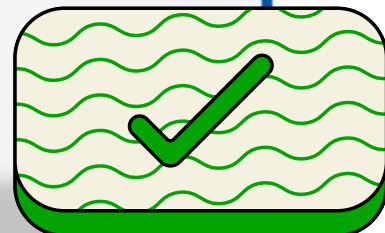
Sauvegarde de  
fichiers et de  
données avec la  
possibilité d'y  
accéder de n'importe  
où

Solution  
personnalisée

Analyse des besoins  
et solutions sur  
mesure

Teams-  
Sharepoint

Espace de  
communication pour  
l'équipe de  
recherche, partage et  
stockage de fichiers





# Gestion des données sensibles de recherche

PARTENAIRES CARACTÉRISTIQUES DÉCOUVRIR ACTUALITÉ CONTACT SE CONNECTER EN



**borealis** Le dépôt Dataverse canadien  
The Canadian Dataverse Repository

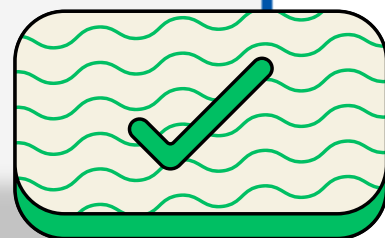
**Déposez, partagez,  
publiez et découvrez  
des données de  
recherche!**

## Questions


Réinitialiser

### 1. Quel est le niveau de confidentialité de vos données?

- Données non confidentielles publiques à risque faible
- Données non confidentielles internes à risque modéré
- Données confidentielles à risque élevé
- Données hautement confidentielles à risque critique ou majeur



borealis Search - User Guide Support English - Log In

Université  de Montréal

Université de Montréal – Dataverse

Borealis >

Contact Share

Nouvel utilisateur? Demandez la création d'un espace pour publier des données de recherche.

Université de Montréal - Dataverse étudiant

Search this dataverse... Advanced Search

Dataverses (62)  
Datasets (195)  
Files (2,385)

Dataverse Category  
Researcher (37)  
Research Group (8)  
Research Project (4)  
Organization or Institution (3)  
Laboratory (1)

Publication Year  
2024 (31)

1 to 10 of 257 Results

Institutional Trust in the World 1995-2022: Data files - world  
Apr 24, 2024 - Institutional Trust in the World - 2023  
Durand, Claire, Pena Ibarra, Luis Patricio, 2024, "Institutional Trust in the World 1995-2022: Data files - world", <https://doi.org/10.5683/SP3/EQOYKW>, Borealis, V1, UNF:6-RszKMMv2xydzwA65rHFmg== [fileUNF]  
This dataset includes data files on the individual data of all respondents to international surveys conducted and published between 1995 and 2022, including data on trust in institutions. It also includes answers to three questions about democracy, covering satisfaction, apprecia...

Institutional Trust in the World - 2023 (Université de Montréal)  
Apr 24, 2024 Claire Durand - Dataverse  
This is the repository for the most recent version of "Institutional trust in the world" comprising data from 17 different International Survey

Borealis -  
Dépôt  
Dataverse  
canadien

Partager et chercher  
des données de  
recherche  
canadienne



# Gestion des données sensibles de recherche



Le dépôt Dataverse canadien  
The Canadian Dataverse Repository

**Déposez, partagez,  
publiez et découvrez  
des données de  
recherche!**

## Conditions d'utilisation

Les données confidentielles et sensibles peuvent être partagées, sous certaines conditions.



• votre contenu contient une attribution et une citation appropriées dans le respect de l'intégrité académique et des normes disciplinaires.

### DONNÉES SENSIBLES ET CONFIDENTIELLES

En utilisant le service, vous confirmez que tout **vos** contenu ne contient pas d'informations qui pourraient identifier directement ou indirectement un sujet, sauf si la divulgation de ces informations d'identification ne risque pas de constituer une atteinte injustifiée à la vie privée ou une violation de la confidentialité. Vous confirmez en outre que toutes les informations personnellement identifiables dans votre contenu respectent au moins l'une des conditions suivantes :

- Les informations ont été précédemment rendues publiques.
- Les informations décrivent des personnalités publiques, lorsque ces informations se rapportent à leurs rôles publics et à d'autres sujets non sensibles.
- Un laps de temps suffisant s'est écoulé depuis la collecte de l'information pour qu'elle puisse être considérée comme historique.
- Tous les sujets identifiés ont donné un consentement éclairé explicite permettant la divulgation publique de leurs informations.
- Toutes les informations ont été collectées avec une déclaration explicite concernant leur caractère public, telles que les informations collectées à des fins réglementaires gouvernementales.
- **Pour les documents fédéraux** (c'est-à-dire le contenu créé par une agence du gouvernement fédéral canadien ou en vertu d'un contrat fédéral) **uniquement**, tous les sujets identifiés sont décédés et aucune loi fédérale ne restreint explicitement la divulgation de ces informations.

# 2 Anonymisation

Défis pour les données de recherche en milieu académique



# Gestion des données sensibles de recherche



## Encore une fracture numérique?

- A plus petite échelle?
- Nécessité d'une littératie numérique pour connaître les notions de base
- Relativité contextuelle du niveau de sensibilité de certains jeux de données de recherche

## Espaces de stockage de données actives

## Dépôts de partage de données finales

## Archivage des données semi-actives et données inactives

- Solutions actuelles: données non sensibles
- Solutions en développement pour les données sensibles
- Typologie des données de recherche : formats, textuelles, audio-visuelles, etc.

## Données administratives vs données de recherche

- Priorités institutionnelles : données administratives (étudiant.e.s, employé.s, etc), évaluations de facteurs de risque à la vie privée (EFVP) dans le contexte d'utilisation de données administratives détenues par l'université à des fins de recherche, sans le consentement des personnes concernées
- Typologie du degré de confidentialité diffère entre les deux types de données en milieu universitaire

# Cas de figure en recherche académique

La sensibilité d'une donnée de recherche varie grandement selon le contexte de la recherche. Bien que des renseignements personnels constituent des données hautement confidentielles et sensibles dans un contexte de recherche, d'autres variables contextuelles existent et peuvent impacter le niveau de sensibilité d'une donnée :

- 1 Sujet de la recherche
- 2 Lieu où se déroule la recherche
- 3 Types de données qui sont collectées
- 4 Types de participant.e.s
- 5 Analyse effectuée
- 6 Outils de collecte et d'analyse utilisés
- 7 Niveau de préjudice pour les participants et les chercheurs dans un cas de bris de confidentialité au niveau social, politique, économique, etc.

# Cas de figure en recherche académique

## Exemple 1

**La recherche est effectuée, en partie ou entièrement, dans une zone active d'un conflit qui déroule dans un autre pays**

Collecte de données qualitatives où des participants sont activement engagés dans le conflit.

Le sujet de la recherche concerne-t-il le conflit?

Existe-t-il un danger pour le chercheur ou les participants?

## Exemple 2

**La recherche porte sur le comportement des enfants d'une classe de 3e année. Leurs visages seront observés lors de l'analyse de données. Impossible de flouter leurs visages.**

Collecte de données audio-visuelles de participants d'âge mineur

Quelle donnée sera collectée?  
Sur quoi portera la recherche?

Comment sera recueilli le consentement?  
Les participants seront-ils identifiables à un certain moment lors de la collecte, l'analyse ou la diffusion?

## Identifying information

- Human participants' names and other HIPAA identifiers must be removed from all sections of the manuscript, including supplementary information.
- Written informed consent must be obtained for the publication of any other information that could lead to the identification of a participant (e.g. clinical images and videos).
- A statement confirming that informed consent to publish identifying information/images was obtained must be included in the methods section.
- Identifying images/video/details which authors do not have specific permission to use must be removed from the manuscript.
- Please note that the use of coloured bars/shapes to obscure the eyes/facial region of study participants is NOT an acceptable means of anonymisation.

### Exemple 2

**La recherche porte sur le comportement des enfants d'une classe de 3e année. Leurs visages seront observés lors de l'analyse de données. Impossible de flouter.**

Collecte de données audio-visuelles de participants d'âge mineur

Quelle donnée sera collectée?  
Sur quoi portera la recherche?

Comment sera recueilli le consentement?  
Les participants seront-ils identifiables à un certain moment lors de la collecte, l'analyse ou la diffusion?

# Gestion des données sensibles de recherche

**Approches  
quantitatives**



**mixte**

**Approches  
qualitatives**







## International Journal of General Systems

Taylor & Francis Group  
an informa business

### Methods and tools for healthcare data anonymization: a literature review

Healthcare is a rapidly evolving field. Such development creates opportunities to provide better quality, evidence-based treatment, however, increasing privacy violations. Anonymization can be appl...

Taylor & Francis / Apr 3, 2023

- Le principal défi de l'anonymisation des données réside dans l'équilibre, en minimisant les risques pour la vie privée tout en maintenant l'utilité et la qualité des données. "Le diamant"!
- Chaque type de jeux de données nécessite des approches spécifiques, et les méthodes peuvent être efficaces contre certains types d'attaques mais vulnérables à d'autres.
- Des difficultés surviennent avec certaines méthodes, comme t-proximity, qui sont complexes à appliquer dans des jeux de données réels.
- Les attaques de liaison constituent une préoccupation majeure.
- Bien que les algorithmes cryptographiques offrent des résultats prometteurs, leur mise en œuvre nécessite des ressources computationnelles importantes.
- l'anonymisation n'est pas qu'un problème technique, mais nécessite des réglementations et directives pour être effectuée de manière efficace et éthique.

Issue	Methods
Privacy risk	Applicable to all methods
Utility of data	Applicable to all methods
Vulnerability to different attacks	<i>k</i> -anonymity <i>k</i> -Map <i>(k, k)</i> -anonymity <i>(1, k)</i> -anonymity <i>(k, 1)</i> -anonymity <i>k<sup>m</sup></i> -anonymity Privacy-constrained anonymity <i>k</i> -join-anonymity Cryptographic algorithms <i>k</i> -diversity <i>(a, k)</i> -anonymity <i>t</i> -Closeness <i>ρ</i> -Sensitive <i>k</i> -anonymity <i>θ</i> -sensitive <i>k</i> -anonymity <i>h, k, ρ</i> -Coherence PS-rule based anonymity <i>ρ</i> -Uncertainty <i>ρ</i> -sensitive <i>k</i> -anonymity <i>(ρ, σ)</i> -sensitive <i>k</i> -anonymity <i>ρ +</i> -sensitive <i>k</i> -anonymity balanced <i>ρ +</i> -sensitive <i>k</i> -anonymity <i>(l, e)</i> -Diversity <i>σ</i> -Presence <i>c</i> -Confident <i>σ</i> -presence <i>(X, Y)</i> -Privacy model
Different methods for different types of data (microdata, big data, transaction data)	<i>h, k, ρ</i> -Coherence <i>ρ</i> -uncertainty PS-rule based anonymity <i>m</i> -invariance <i>k</i> -join-anonymity <i>S</i> -diversity <i>(a, k)</i> -Anonymity <i>(k, e)</i> -anonymity <i>(X, Y)</i> -Privacy model <i>(l, e)</i> diversity <i>h</i> -ceiling
Linkage attack	<i>k</i> -anonymity <i>k</i> -Map <i>l</i> -diversity <i>t</i> -Closeness <i>(a, k)</i> -Anonymity <i>ρ</i> -Sensitive <i>k</i> -anonymity <i>(k, k)</i> -anonymity
Trustfulness of data	e-differential privacy
Not only technical problem	Applicable to all methods
Computational resources requirements	Cryptographic algorithms
Difficult to implement in real-life data	<i>t</i> -Closeness

# Baby's First Years

Première étude causale visant à tester les liens entre la réduction de la pauvreté et le développement du cerveau chez les très jeunes enfants



2018		2019		2020		2021		2022		2023	
Data publicly available		Data publicly available		Data publicly available		Data publicly available		Data publicly available		Data collection complete	
Recruitment/Baseline May 2018 – June 2019		Age 1 data collection July 2019 – June 2020		Age 2 data collection July 2020 – June 2021		Age 3 data collection July 2021 – June 2022		Age 4 data collection July 2022 – ongoing			
Hypotheses preregistered with Clinicaltrials.gov		Revised set of hypotheses submitted, before age 1 data collection		Revised set of hypotheses submitted, before age 2 data collection		Revised set of hypotheses submitted, before age 3 data collection		Revised set of hypotheses submitted, before age 4 data collection			
In process		Age 5 maintain contact July 2023 – ongoing		Age 6 data collection July 2024 – June 2025		Age 7 maintain contact July 2025 – June 2026		Age 8 data collection July 2026 – June 2027			
2023		2024		2025		2026		2027			

Étude longitudinale

Méthodes mixtes

Participants

- approx. 1000 mères
- revenus < au seuil de pauvreté
- dans 4 grandes villes américaines,
- venant tout juste d'accoucher

## Baby's First Years (BFY), New York City, New Orleans, Omaha, and Twin Cities, 2018-2022 (ICPSR 37871)

Version Date: Mar 19, 2024 [Cite this study](#) | [Share this page](#)

Principal Investigator(s):

[Katherine A. Magnuson](#), University of Wisconsin--Madison; [Kimberly Noble](#), Columbia University. Teachers College; [Greg J. Duncan](#), University of California, Irvine; [Nathan A. Fox](#), University of Maryland, College Park; [Lisa A. Gennetian](#), Duke University; [Hirokazu Yoshikawa](#), New York University; [Sarah Halpern-Meekin](#), University of Wisconsin--Madison

<https://doi.org/10.3886/ICPSR37871.v6>

Version V6 ([see more versions](#))

Explore Data | Analyze Online (SDA) | Download

At A Glance | Data & Documentation | Variables | Data-related Publications | Export Metadata

Project Description

**16,281** Downloads \* [Usage Report](#)  
\* past three years

**32** [Data-related Publications](#)

# Baby's First Years

## Personally Identifiable Information (PII)

Personally identifiable information (PII; e.g., date of birth) or potentially PII (e.g., child development measure items specific to child age in months) is protected under Health Insurance Portability and Accountability Act (HIPAA). We refer to HIPAA protected information as PII. We collect PII with the survey, so we have excluded these items in the data file that we deposit to ICPSR. In order to protect PII, these variables have either been removed or converted into a dummy variable that indicate that the mother provided a response. Some of these variables may become available in the future under more restrictive terms. However, as some of these variables can be essential for analysts, in some cases, we generated new variables that partially or completely mask the sensitive information. These variables are HIPAA compliant and useful for analysis (See Table 1). Some of these variables are described in the table below.

**Table 1. Masked Personal Information in Age-1 Public Release**

Sensitive information	Variable Name(s)	Description
Names (e.g., Mother, Child, Household Member)	respfnameal childfnameal hhmemnamea_*	Information replaced with a dummy indicator for whether there is a response to each item.
Name (Father)	dadnamefal	Father names are not provided in the data but this variable contains information such as whether the father is in the household. Please see "DadNameF" of Survey instrument for full details.
Child's birthdate	monthbirthal	A "masked" month of birth for each child. This variable consists of 14 values ranging from 17-30 to represent a month between May 2018 and June 2019. Each value corresponds to a distinct birth month but the months themselves are scrambled.
Child's age at interview	iwdatate_age_mask_a1	Child's age at the time of the Age-1 interview recoded to a binary indicator for whether the child was at least one-year-old.
Household members' birthdates	month_a_X dob_mo_a_X	Information replaced with a dummy indicator for whether there is a response to each item.
ASQ Measures (age-specific)	tot_casqlangal casqlangal casqlangcutoffal casqlangbelowal casqlangcloseal	ASQ items specific to developmental age period have been dropped, and summary measures are provided.
Interviewer ID	intervieweral	Randomly generated interviewer identification number. (Not linked over

## Administrative data and transaction tracking consent form

information shared with someone who is not permitted to see or know about that information. We will do everything we can to keep the data secure and to make sure your data is not seen by anyone outside of the research team, but we cannot promise complete confidentiality.

All additional data requested from the debit card company and any administrative records will be stored securely on a password protected server at the University of Wisconsin. Unanticipated problems, like a stolen password, may occur, although such incidences are highly unlikely. Our research team will take the utmost care to protect your privacy.

The researchers plan to publish the results obtained from additional data. To protect your privacy, they will not include any information that could directly identify you. They will protect the confidentiality of your research records by assigning a unique participant number to any additional data obtained and never associating your name and any identifying information with any of the collected data. Any data obtained, will be stored in secure data storage at the University of Wisconsin where the data will be processed for use by the research team. The files linking your name to the participant number will be kept in a password-protected database to which only key research staff will have access. When researchers report the study findings they will only report information in general aggregated terms so participants cannot be identified.

The following individuals and/or agencies will have access to the additional data we may obtain:

- Researchers at University of Wisconsin maintaining secure data storage.
- Researchers from University of California – Irvine, Columbia University's Teachers College, New York University, and University of Wisconsin will use the data for research analysis purposes. Your identity as a research subject will be protected and your name will not be associated with any additional data.
- Authorities from [university], including the Institutional Review Board, and from the Office for Human Research Protection may also access your data

This study holds a Certificate of Confidentiality (CoC) that offers additional protections for your identifiable research information, and records. The most important protection is that members of the research team cannot be forced to disclose or provide any of your private identifiable information, in any Federal, State, or local civil, criminal, administrative, legislative, or other proceeding unless you provide permission. Disclosure of your research information may only occur in limited instances.

If you tell us something that makes us believe that you or others have been or may be physically harmed, we may report that information to the appropriate agencies.

# Baby's First Years

## ICPSR PROCESSING NOTES FOR #37871

Baby's First Years (BFY), New York City, New Orleans, Omaha, and Twin Cities, 2018-2019

### DS #1: Baseline Data

- Missing Value Differences:** Due to limitations across the various statistical packages, missing values ".d" and ".r" were recoded to -888 and -999 respectively.
- Data/Documentation Discrepancies:** The variable labels in the P.I. codebook for **MWORK3TO6MONTHSA0** and **MWORK6TO12MONTHSA0** are inconsistent with those in the data; the labels in the data are correct.

Previewing 25 of 1050 total rows as exploration

[Clear all](#)

S.NO	MOTHER'S SAMPLE ID PUBLICSAMP LEID	[RAW] PRELOAD - R FIRST NAME FIRSTNAMEA0	[RAW] PRELOAD - R LAST NAME LASTNAMEA0	[RAW] APPLICATION VERSION VERSIONDATEA0	[RAW] TOTAL MINUTES OF BASEL ... TOTALMINSBASELINEA0	[RAW] CHILD INFORMATION INTR ... CIINTROA0
1	P8861679	(1) Yes, there is an answer	(1) Yes, there is an answer	05-JUN-2018	23.03	(1) Continue
2	P6209515	(1) Yes, there is an answer	(1) Yes, there is an answer	05-JUN-2018	17.9	(1) Continue
3	P8849707	(1) Yes, there is an answer	(1) Yes, there is an answer	05-JUN-2018	21.02	(1) Continue
4	P150060	(1) Yes, there is an answer	(1) Yes, there is an answer	25-SEP-2018	21.78	(1) Continue
5	P9437612	(1) Yes, there is an answer	(1) Yes, there is an answer	25-SEP-2018	27.35	(1) Continue

## BFY Codebook Baseline

Created: 7/23/2020

### Raw Variables

#### 1. publicsampleid

Mother's sample ID, which can be used to merge datasets.

#### 2. firstnamea0

Flag for whether mother provided her first name

[raw] Preload-R First Name

N Value  
934  
116 .

Code Value  
1 Yes, there is an answer

#### 3. lastnamea0

Flag for whether mother provided her last name

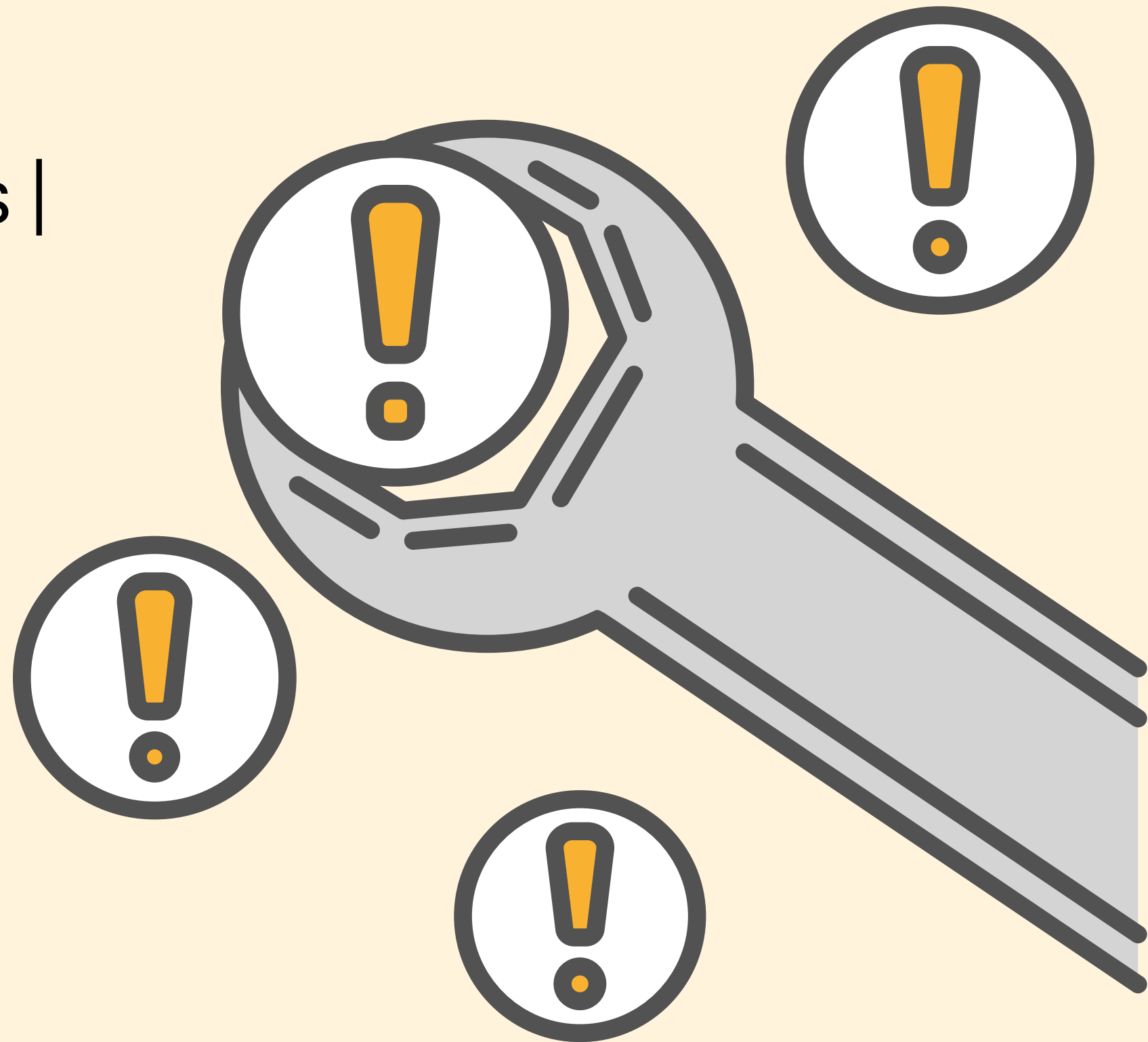
[raw] Preload-R Last Name

N Value  
914  
136 .

Code Value  
1 Yes, there is an answer

# 3 Anonymisation

Méthodes | Outils | Processus |  
Solutions complémentaires



# Anonymiser en 3 étapes

1

Déterminer +  
supprimer ou masquer  
les identifiants directs

Quanti

Supprimer certaines variables.

Quali

Ça dépend! Remplacer par pseudonymes des fois.

2

Déterminer  
les identifiants indirects  
(quasi-identifiants)

Quanti

Âge/Date de naissance

Profession

Revenu

Sexe

Géographie: région/ville/village, etc.

Origine ethnique/ethnicité

Religion

Important! Bonnes métadonnées:

- étiquettes de variables
- étiquettes de de valeurs

3

Examiner

Quanti

- les fréquences d'info potentiellement révélatrices
- les valeurs aberrantes.
- les variables textuelles en cas d'info personnelles potentiellement révélatrices ou sensibles

"Ma soeur souffre d'une maladie rare."

"J'ai été victime de violence domestique et j'ai utilisé l'association X pour obtenir de l'aide."

"J'ai travaillé pour l'organisation X pendant 35 ans."

# Données quantitatives

No.	Anonymization method
1	$k$ -anonymity <b>la méthode la plus conservatrice</b>
2	$l$ -diversity
3	$t$ -closeness
4	$\rho$ -sensitive $k$ -anonymity
5	$\rho+$ -sensitive $k$ -anonymity
6	Cryptographic algorithms (RSA, ElGamal, DES, AES)
7	balanced $\rho+$ -sensitive $k$ -anonymity
8	$(\rho, \sigma)$ -sensitive $k$ -anonymity
9	Pseudonymization, tokenization, hashing
10	$\epsilon$ -differential privacy <b>datasets adjacents</b>
11	$(\epsilon, m)$ -Anonymity
12	$k$ -join-anonymity model
13	$m$ -invariance
14	$(\sigma, k)$ -Anonymity

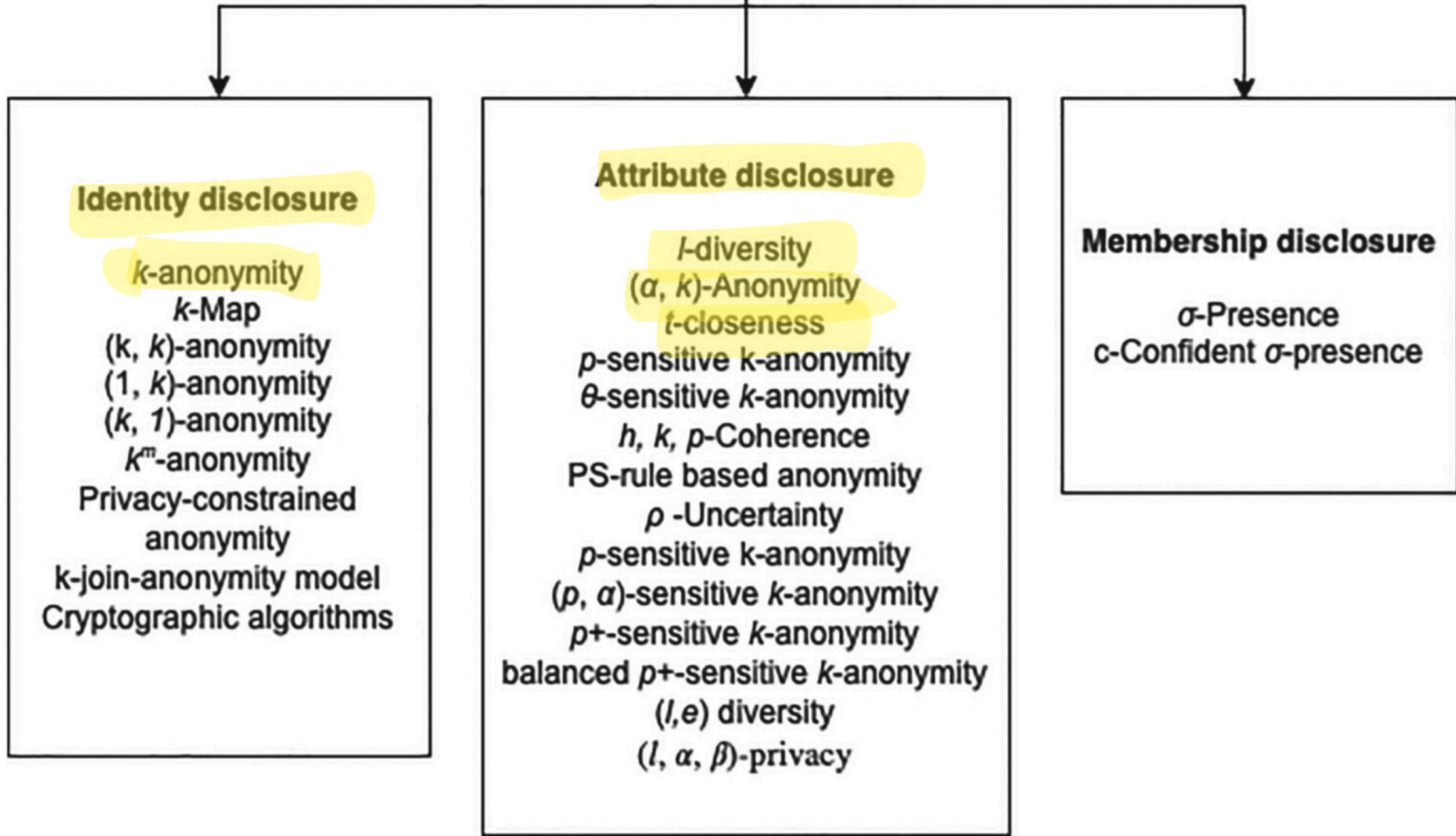
15	Closed $l$ -diversity/Closed $l$ -diversification
16	Semantic linkage $k$ -anonymity (SLKA)
17	$\theta$ -sensitive $k$ -anonymity
18	$(k, e)$ -anonymity
19	S-diversity
20	$(X, Y)$ -Privacy
21	$(k, k^m)$ -anonymity
22	$h, k, \rho$ -Coherence
23	$\rho$ -Uncertainty
24	Privacy-constrained anonymity
25	PS-rule based anonymity
26	$(l, e)$ diversity
27	$h$ -ceiling

31	$k$ -Map
32	$\sigma$ -Presence
33	$c$ -Confident $\sigma$ -presence
34	$(k, \epsilon, \delta)$ -anonymization
35	F-Classify
36	$f$ -Slip
37	$(\epsilon, \sigma)$ -differential privacy
38	$\beta$ -likeness
39	$(l, \sigma, \beta)$ -privacy
40	$\tau$ -safety



# Données quantitatives

Types of attacks and methods against them



Les limites de k-anonymat et ses variantes ont permis de développer d'autres méthodes!

**Couplage d'enregistrement**  
L'adversaire parvient à restaurer de l'identification des enregistrements figurant dans l'ensemble de données.

**Couplage d'attribut**  
L'adversaire parvient à coupler un participant à un sous groupe dont les membres ont certaines propriétés

# Quelques conseils ...

Données  
quantitatives



Agréger ou réduire la précision



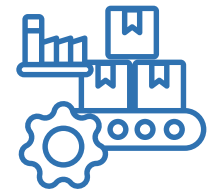
Recoder les variables clés catégorielles en moins de catégories

(k-anonymat)



Supprimer des valeurs spécifiques des variables clés pour certaines unités

(k-anonymat)



Généraliser le sens des variables textuelles

remplacer les réponses libres potentiellement révélatrices par un texte plus général



Restreindre les plages supérieures ou inférieures d'une variable continue pour masquer les valeurs aberrantes. Par exemple, l'âge - recoder en 70 ans et plus

Comment décider ? Examiner la distribution de cette variable.



Anonymiser les données géoréférencées

remplacer les coordonnées ponctuelles par des variables non révélatrices

# Données quantitatives

Tool	Methods
ARX	<ul style="list-style-type: none"> <li><math>k</math>-anonymity</li> <li><math>l</math>-diversity</li> <li><math>t</math>-closeness</li> <li><math>\sigma</math>-disclosure privacy</li> <li><math>\beta</math>-likeness</li> <li><math>\Sigma</math>-presence</li> <li><math>k</math>-map</li> <li><math>(\epsilon, \sigma)</math>- differential privacy</li> </ul>
scdMicro	<ul style="list-style-type: none"> <li><math>k</math>-anonymity</li> <li><math>l</math>-diversity</li> <li>Randomization</li> <li>Adding noise</li> <li>Rank swapping recording</li> </ul>
$\mu$ -ANT	<ul style="list-style-type: none"> <li><math>k</math>-anonymity</li> <li><math>t</math>-closeness</li> </ul>
ShinyAnonymizer	Encryption and hashing methods
The Anonymizer	Salting and hashing methods
Amnesia	<ul style="list-style-type: none"> <li><math>k</math>-anonymity</li> <li><math>k^m</math> anonymity</li> </ul>
Anonimatron	Pseudonymization
$\mu$ -ARGUS	$k$ -anonymity

Calculer le bruit nécessaire pour camoufler une contribution individuelle (Confidentialité différentielle)

**Table 2**

Basic properties of the tools identified

Tool	Institution	Country	Language(s)	Release	Last update	License
μ-Argus	Centraal Bureau voor de Statistiek	Netherlands	C++, Java	1998	2021	EUPL
sdcmicro	Statistics Austria	Austria	R	2007	2021	GPL 2
Open Anonymizer	University of Vienna	Austria	Java	2008	2009	Unknown
CAT	Cornell University	USA	C++	2009	2014	Unknown
Tiamat	Purdue University	USA	Java	2009	Unknown	Unknown
UTD	The University of Dallas	USA	Java	2010	2012	GPL 2
Anon	University of Klagenfurth	Austria	Java	2012	Unknown	Unknown
ARX	BIH@Charité	Germany	Java	2012	2022	Apache 2
SECRET	University of Peloponnes	Greece	C++, Qt	2013	Unknown	Unknown
Probabilistic Anonymization	University of Cyprus, Cyprus and Newcastle University, UK	Greece/UK	R	2018	2018	Unknown
μ-Ant	Center for Cybersecurity Research of Catalonia	Spain	Java	2019	2019	MIT
Amnesia	University of Thessaly	Greece	Java, JavaScript	2019	2022	BSD 3-Clause
PrioPrivacy	Research Studio Data Science	Austria	Java	2019	2021	Unknown

Données  
quantitatives

Anna C Haber, Ulrich Sax, Fabian Prasser, the NFDI4Health Consortium, Open tools for quantitative anonymization of tabular phenotype data: A literature review, Briefings in Bioinformatics, Volume 23, Issue 6, November 2022, bbac440, <https://doi.org/10.1093/bib/bbac440>

Données  
qualitatives

# QDR

## The Qualitative Data Repository

### De-Identification

Often – although not always – research data need to be collected, managed, used, shared, and potentially re-used in ways that protect human participants' identities, i.e., in ways that do not allow the data t...

 QDR / Sep 7, 2021

Myers CA, Long SE, Polasek FO. Protecting participant privacy while maintaining content and context: Challenges in qualitative data De-identification and sharing. Proc Assoc Inf Sci Technol. 2020;57:e415. <https://doi.org/10.1002/pr2.415>

**Données  
qualitatives**

# Advance Care Planning in Hospice Organizations: A Qualitative Pilot Study


Version 2.0



Harrison, Krista. 2021. "Advance Care Planning in Hospice Organizations: A Qualitative Pilot Study". Qualitative Data Repository. <https://doi.org/10.5064/F6YMWPUX>. Palliative Care Research Cooperative QDR. V2

[Cite Data Project](#) ▾[Learn about Data Citation Standards.](#)[Download Data Project](#) ▾[Link Dataset](#)[Contact Owner](#)[Share](#)

This dataset includes transcripts of 51 semi-structured interviews from a four-site qualitative study; 33 documents could not be de-identified and available only upon request from PI. All participants gave verbal consent before participating in a semi-structured interview whose domains included (1) contextual information about the participant and hospice organization; (2) processes and practices of eliciting and documenting preferences for care among hospice enrollees; and (3) professional opinions on eliciting/ documenting preferences in the context of the hospice philosophy, including changes in practices over time.

**Make Data Count (MDC) Metrics**   
since 2019-10-01

**6,971 Views** 

**1,058 Downloads** 

## Data Collection Overview:

All data were collected by the depositor (a PhD qualitative researcher) during a two-day site visit to each of four non-profit, community-based hospices affiliated with the Palliative Care Research Cooperative (PCRC), between April and September 2016. Semi-structured in-depth interviews were conducted with key informants. All interviews were digitally recorded. **Audio recordings** of interviews were transcribed by a **professional transcription service and reviewed for accuracy**. All data were **converted to electronic format**, then uploaded to a **qualitative data analysis software program: Atlas.ti, version 8**. Any identifying information was redacted by **deleting audio-recorded portions of the tapes, deleting words from the transcript, or blacking out the words in hand-written notes**. Member checking was used to validate and establish credibility of the findings by returning transcripts to the participant for review and clarification, and presenting preliminary findings to a diverse audience of hospice and palliative care researcher-clinicians, to solicit views and interpretations of the credibility of the findings. Other documents were also collected (brochures, internal training materials), but cannot be redacted and therefore are not being shared publicly.

# Données qualitatives

**INTERVIEWER:** That's right. So begin by telling me about yourself and your role in the organization.

**P16:** Sure, my name is [redacted]. I'm the president and CEO of [redacted]. I've been here since 2008. It's my second time around as CEO. I was here from [redacted]. That's my role.

**INTERVIEWER:** How long have you been in hospice generally?

**P16:** I have actually kind of grown up in the industry. I started when I was [redacted], so since back in [redacted]. So it's been about 20 plus years now.

**INTERVIEWER:** Wow, what roles have you been in?

**P16:** CFO, I started off as a CFO for a hospice in [redacted-city name] and then I've been a CEO since 2002.

**INTERVIEWER:** Got it.

**P16:** So kinda 15 years now, so 10 years as a CFO, 15 as a CEO.

**INTERVIEWER:** Great. Tell me about the patient population that [redacted] serves.

**P16:** Sure, so geographic wise we serve the region of Western [redacted – state name]. So a beautiful mountainous area. Some population centers like [redacted – city]. So [redacted] county is 200,000 people. [redacted – city] is 100,000 people but then you get to some pretty rural counties like [redacted – three listed]. So I'm at counties of 20/15/30,000 people. So very rural and very picturesque but a lotta space in between, so it's very difficult to get to those patients. Large chronically ill population. So a lot of kinda bad life habits, eating habits et cetera manifest themselves in multiple chronic diseases. So we're a pretty unhealthy population when you look throughout kind of the [redacted] Mountains area, whichever term you wanna use to describe the area. We have a large-- I think our greatest diagnosis is dementia from a patient population standpoint. Specifically, in [redacted] County which is where kind of our corporate headquarters is based here. We have a large retiree population so our population since 2002 has been indicative of where the country is going. So 20 percent of our population has been greater than 65 years or older. So, when you look at nursing homes-- so we have 1,000 nursing home beds in a county of 100,000 people. That's pretty unprecedented and so that's indicative of the large retiree population that we have here. In fact, I've always believed that one of the interesting things about [redacted] and why I just love this organization is that it's always felt like we've had a microcosm of the broader population. So if we can make some innovative projects work here, it would have applicability to the rest of our country. While the diversity of our population-- we do have a large Hispanic population, a lotta migrant workers. So that is indicative of our country. We don't have a large African American population. So we don't have diversity that way. But, if



# Données qualitatives

[Our text anonymisation helper tool](#) can help you find disclosive information to remove or pseudonymise in qualitative data files. The tool does not anonymise or make changes to data, but uses MS Word macros to find and highlight numbers and words starting with capital letters in text. Numbers and capitalised words are often disclosive, e.g. as names, companies, birth dates, addresses, educational institutions and countries.

## Anonymising qualitative data

A comprehensive resource funded by the ESRC to support researchers, teachers and policymakers who depend on high-quality social and economic data.

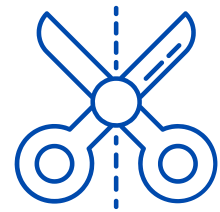
UK Data Service

```
TestData.txt - Bloc-notes
Fichier Edition Format Affichage Aide
I:      So the wartime Secretaries of State, Lord Lloyd, Moyne, Cranborne, Aby and then Oliver Stanley, their
R:      Yes, it had to be. And of course the period when we had Lord Lloyd was a bit of a revelation. He just
I:      Why did he have this tremendous impact?
R:      I don't know. He always had this kind of forceful and ruthless personality.
I:      He was also close to Churchill?
R:      Yes. Ah, now of course one thing one's got to take into account during that wartime period is the American
I:      Did that lead to many disagreements?
R:      You mean inside the Office?
I:      Between the United States and the British Government?
R:      I don't think it did but I wasn't actually much involved in this. Most of the negotiations had taken
dn't admit that; they had to find some scapegoat and what better scapegoat than wicked British imperialism?
I:      What were the circumstances in which Parkinson came back and took over from Sir George Gater for two
R:      Yes, they did a sort of 'Box and Coxing', didn't they? I can't remember why. I think Gater just retired
```

Nom	Type
AnonymisationHelperMacros.txt	Document texte
AnonymisationHelperTab.xml	Document XML



# Quelques conseils...



Planifier ou appliquer l'édition au moment de la transcription, sauf : études longitudinales - (liens)



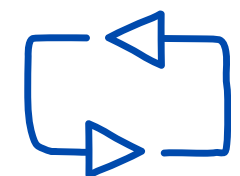
Cohérence au sein de l'équipe de recherche et tout au long du projet



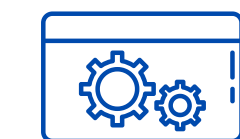
Identifier les remplacements, par exemple avec [crochets]



Tenir un journal d'anonymisation de tous les remplacements, agrégations ou suppressions effectués - le conserver séparé des fichiers de données anonymisées



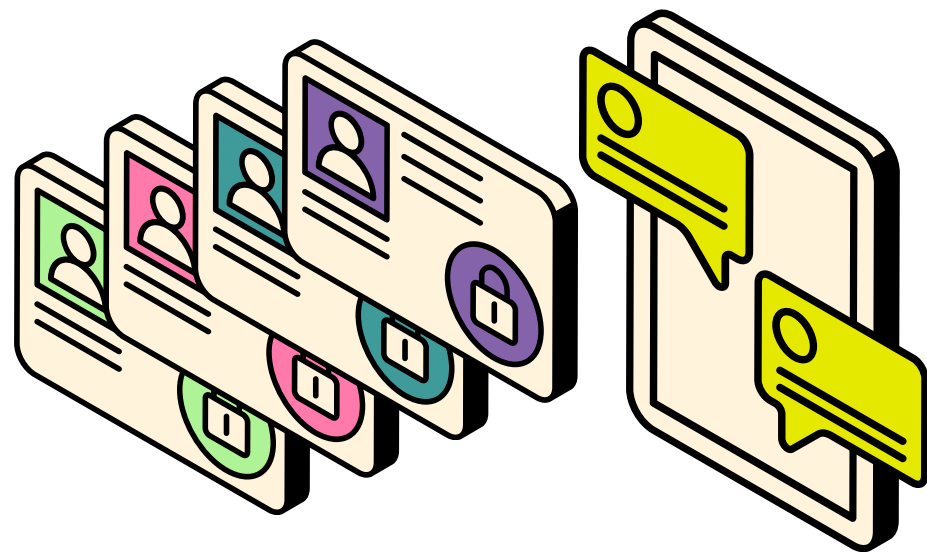
Éviter de laisser des espaces vides ; utiliser des pseudonymes ou des remplacements



Éviter une sur-anonymisation - supprimer/agréger des informations dans le texte peut déformer les données, les rendre inutilisables, peu fiables ou trompeuses

- Contrôler l'accès est une meilleure option que la sur-anonymisation

# Catalogue de solutions



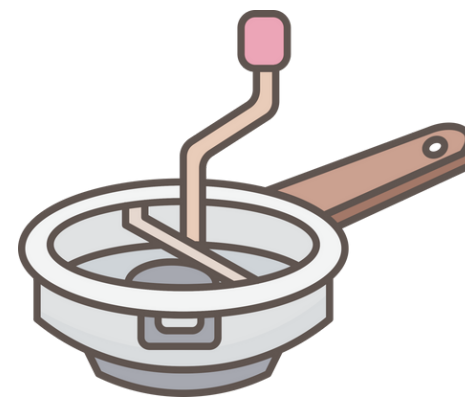
## Service conseil sur mesure

[Consortium Santé Numérique](#)

Université de Montréal

Exemple:

Packages utilisant le langage R  
dont [SDCMicro](#)



## Solution "moulinette"

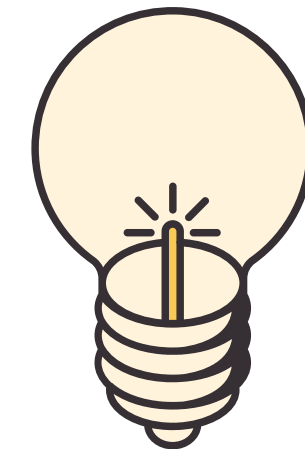
≠ panacée!

exemple

[amnesia.openaire.eu](https://amnesia.openaire.eu)

DÉMO inclus

OpenAIRE + HORIZON 2020



## Solutions alternatives

Contrôle d'accès aux données

Partage de la méthodologie de  
recherche + [principes FAIR](#)



🏠 Anonymization Wizard

🔄 Reset

📄 Source <

📄 Anonymized ▾

Manage

👤 Hierarchy <

🔒 Algorithms

📊 Solution Graph

📄 Results

@ Back to Amnesia website

## Amnesia Dashboard

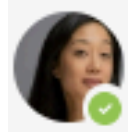
version:1.3.3 beta

Do not upload sensitive data, the online edition is for demonstration purposes only

Upload sensitive data 📄

Upload

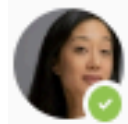
➔ **Drop files** to upload  
(or click)



Stéphanie Pham-Dang 04-26 10:57 Modifié

Kamran Afzali m'a expliqué: c'est parce que ce sont des 🦴 cyberattaques réelles et certaines hypothétiques. Toutes les méthodes 🛠️ ont été développées au fur et à mesure que les hackers sont parvenus à déjouer les méthodes, K anonymity (le baseline) d'abord. C'est donc un processus itératif 🔄 d'un véritable jeu du chat 🐱 et de la souris 🐭. Les méthodes sont développées en réaction aux cyberattaques, donc en aval 📺. Il y a en parallèle des hackathons "éthiques" où les organisations, en amont 📺 inversement, encouragent les hackers de déjouer leurs nouvelles méthodes, avec récompenses 💰.

[Afficher moins](#)



Stéphanie Pham-Dang 04-26 11:04 Modifié

Fascinant! Cette revue de littérature concerne le domaine de la santé où les pratiques sont déjà très avancées. Alors que moi ce n'est pas QUE mon public, j'ai des chercheurs qui sont maintenant obligés de partager leurs données dans certaines revues (ex. psycholinguistiques) pour fins de reproductibilité. Le comité d'évaluation leur revient avec des commentaires comme quoi leurs méthodes d'anonymisation doivent minimalement se basé sur K-anonymity, avec un package R. Et là, on me demande comment se former: des étudiants au doctorat, des petites équipes qui ne font pas dans la santé.

[Afficher moins](#)



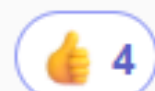
Modifié

donc pour le quanti, il y a des solutions sur mesure pour ceux qui en bénéficient (domaine santé à UdeM), mais pour le reste, on s'attend à des outils qui automatisent des méthodes de base (comme K-anonymity, l-diversity et t-proximity) pour simplifier et démocratiser cette tâche, pour se conformer aux exigences légales.



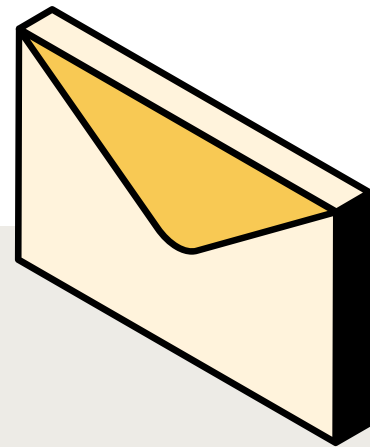
Yves Terrat 04-26 11:29

C'est compliqué de mettre en place des outils sur mesure pour des types de données qui sont extrêmement variables en fonction des domaines de recherche, des source, etc etc. Ça me paraît difficile de ne pas accompagner, au moins dans un premier temps, les chercheuses et chercheurs qui ne bénéficient pas de connaissances ou de ressources humaines dans leur équipe.

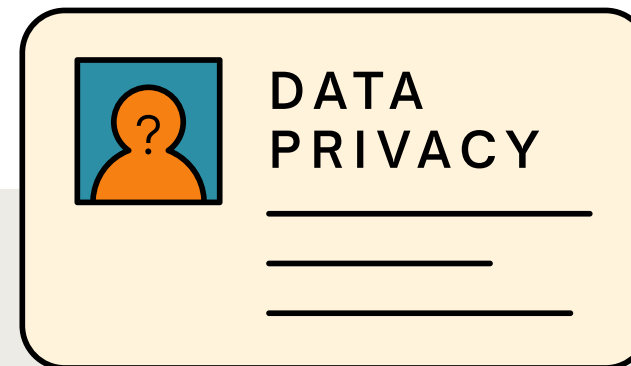


# L'anonymisation oui, mais aussi...

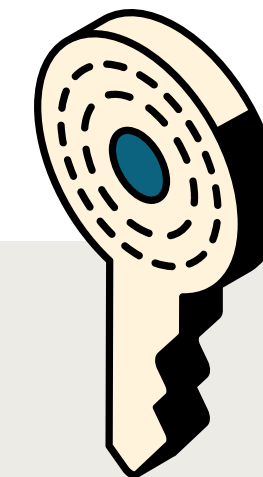
Trois volets pour protéger les personnes participant à un projet de recherche. Ces stratégies permettent le partage de la plupart des données et minimisent le risque d'identification.



**Consentement x2  
projet de recherche  
et réutilisation**



**Anonymisation  
Dépersonnalisation (ou  
pseudonymisation en UE)**



**Sécurité d'écosystème numérique  
+ Contrôle d'accès  
+ fiche de métadonnées**



## Sensitive Data | Publish your research | Springer Nature

Authors with sensitive data, or other data that cannot be shared openly, should apply appropriat...

 [springernature.com](https://springernature.com)

If any of the above are applicable, authors should consider the following methods to facilitate safe sharing of sensitive data. Participant consent to share the data (in addition to use or collection of data) should also be obtained and documented prior to data collection.

- ✓ Anonymising data to create a shareable version. It may be possible to remove or replace identifying information in the data before sharing openly. See [guidance on anonymisation from \*Trials\* and the UK Data Service](#).
- ✓ Use of a controlled access repository to manage who can access data and under what terms. Certain repositories offer this functionality, for example enabling a data owner to know who has access to the data and/or to apply additional restrictions such as an agreement not to reidentify participants. See [guidance on suitable controlled-access repositories](#), or consult services such as [ClinicalStudyDataRequest.com](https://ClinicalStudyDataRequest.com).
- ✓ Use of Trusted Research Environments or data safe havens. Certain research institutions manage environments within which data can be queried and accessed by trusted parties only, without removing data from the system. These are mainly associated with clinical health settings. Contact your research institute to check if this is an option for your data.
- ✓ Use of metadata records in repositories for data that cannot be publicly shared. This provides persistent, long-term context to another researcher on what data are available, even if the data can only be made available on request. See this [example from \*npj Precision Oncology\*](#).

Data availability subject to **controlled access**

The **data availability statement** should include the following information:

- reasons for controlled access (e.g., privacy, ethical/legal issues)
- conditions of access must be described precisely including contact details for access requests,
- timeframe for response to requests,
- restrictions imposed on data use via data use agreements.

A copy or link to the **data use agreement** should be provided if requested by editors.

### **Restrictions on controlled access datasets**

- including restrictions on downstream data reuse or authorship requirements
- must be clearly described in manuscript and to editors at the time of submission.

Editors may decline further consideration of the manuscript after evaluation if restrictions are found to be unduly prohibitive.

### **Studies involving vulnerable groups**

- For manuscripts reporting studies involving vulnerable groups where there is the potential for coercion or where consent may not have been fully informed, extra care will be taken by the editor.
- The manuscript may be referred to an internal editorial oversight group for further scrutiny.
- Consent must be obtained for all forms of personally identifiable data including biomedical, clinical, and biometric data.
- Documentary evidence of consent must be supplied if requested.

# Cybersécurité

Pour assurer un **contrôle d'accès** adéquat, l'environnement numérique doit aussi être SÉCURISÉ. Une **liste de contrôle** s'avère nécessaire!

réseaux

ordinateurs

inonuagique

transferts

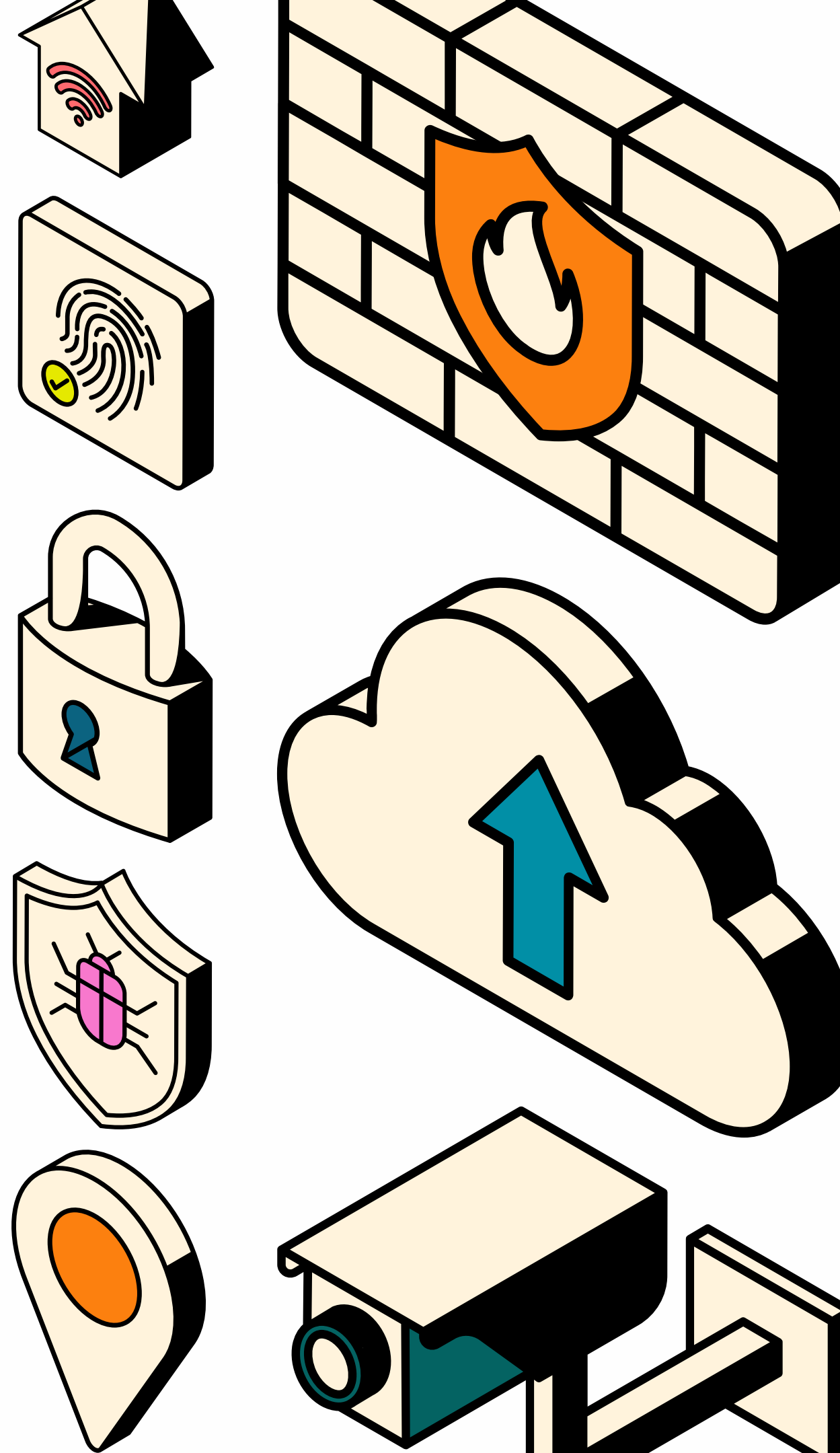
etc.

Center for Internet Security (CIS) Critical Security Controls

National Institute of Standards and Technology (NIST) Framework for Improving Critical Infrastructure Cybersecurity

Centre canadien pour la cybersécurité - informations pour le milieu universitaire

Ministère de la Cybersécurité et du Numérique (Québec)





# Autres sources

- [Directives UdeM sur l'utilisation de l'infonuagique](#)
- [Grille de stockage UdeM](#)
- [Exemples UdeM de renseignements et de documents selon leur niveau de confidentialité](#)
  
- [Conditions d'utilisation de Borealis.ca.](#)
  
- [Data Anonymization Workshop Series \(2023\)](#)
- [Douglas College Libguide - Research Data Safeguarding](#)
- [UK Data Service. Anonymisation. 2021.](#)
- [Finnish Social Science Data Archive. Identifier type table.](#)
  
- [Boîte à outils pour les données sensibles — destiné aux chercheurs :](#)
  - [Partie 1: Glossaire terminologique sur l'utilisation des données sensibles](#)
  - [Partie 2: Matrice de risque lié aux données de recherche avec des êtres humains](#)
  - [Partie 3 : Langage en matière de gestion de données de recherche pour le consentement éclairé](#)



# Merci!

*les bibliothèques*

---

Université   
de Montréal