

Huma-Num IR*

Atelier pratique d'anonymisation des données en SHS avec Incognito

Xiaoou Wang - Référent Humanités Numériques

Textes encadrant l'anonymisation

- 01.** Loi du 6 janvier 1978, dite « Informatique et Libertés » (LIL) modifiée
- 02.** Règlement 2016/679 du Parlement européen et du Conseil du 27 avril 2016 (RGPD)
- 03.** Article 8 de la Charte des droits fondamentaux de l'Union européenne (2000/C 364/01)7
- 04.** Loi pour une République Numérique 2016-13218

Notions de base communes I

Anonymisation

Traitement consistant à utiliser un ensemble de techniques de manière à rendre **impossible** toute identification.

Réutilisation des données : sans restriction

Pseudonymisation

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**.

Réutilisation des données : avec restriction

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

- **Données anonymisées** : diffusées en accès libre.
- **Données pseudonymisées** : diffusées en accès restreint, avec un contrôle manuel des accès par les ingénieurs du CDSP.



Source : <https://www.cnil.fr/fr/recherche-scientifique-hors-sante-enjeux-et-avantages-de-lanonymisation>

Centre de données socio-politiques : Valentin Brunel, Paul Colin, Alina Danciu, Quentin Gallis, Guillaume Garcia. Pratiques d'anonymisation et pseudonymisation des données diffusées par le CDSP. 2022. <hal-03657563>

Notions de base communes II

Identification directe

« Je m'appelle Marie, j'ai 23 ans. »

Identification indirecte

Identification contextuelle

- « Je m'appelle X, j'ai X ans. Je suis la seule professeure de japonais au lycée de Menton. »
- Je suis une vétérinaire spécialisée en rapaces dans la vallée de la Roya.

Ces trois exemples illustrent comment l'identité d'une personne peut être révélée soit par des informations nominatives explicites, soit par un recoupement de critères contextuels uniques.

La révision manuelle est indispensable.

Notions de base communes III

Panorama des Types de Données



Tableur

Données de santé



Textuelles

Entretiens



Audio



Image



Vidéo

Menu du jour : Les Données Textuelles

Les méthodes varient selon le type de données.

Privacy by Design

1. Anonymisation Précoce

Anonymiser dès que possible : le risque existe pendant toute la durée de vie du projet.

Exemples :

- Demande d'un collègue
- Exploration de textes via ChatGPT

2. Minimisation de la Collecte

Ai-je réellement besoin de les collecter ?

Exemple : 17 avenue des Acacias, Nice → Nice Nord/Cimiez

Source : <https://donnees-rgpd.fr/definitions/privacy-by-design/>

Après l'anonymisation

1. Diffusion différenciée

Anonymisation faible, moyenne, complète

FAIR, accessible != ouverte à tous, possible sur Nakala

2. Charte d'utilisation

Toute une littérature sur les **Data Use Agreements (DUA)**

- Durée limitée (engagement à détruire)
- Interdiction de ré-identification
- Interdiction de recouper les bases
- Notification d'incident & **Watermarking** personnalisé

Exemples de Variantes

Variante 1 : Ordre des documents

Corpus Alice: Doc 1, 2, 3

Corpus Bob: Doc 2, 1, 3

Variante 2 : Identifiants internes

[PERSONNE_001] vs [PERSONNE_734]

Variante 3 : Synonymes

enseignant vs professeur

Incognito - Motivation

Privacy-first qualitative data anonymization

Tool Overview

Incognito anonymizes qualitative corpora locally without sending data to the cloud.

- Multiplatform (macOS, Windows, Linux)
- Human-controlled: validate and correct before export
- Audit trail and full traceability

Access the Software and the Source Code

<https://github.com/xiaoouwang/Incognito/releases>

<https://github.com/xiaoouwang/Incognito>

Incognito - Motivation

Pourquoi cet outil ?



100 % local

Aucun appel à une API externe. Vos textes restent sur votre machine.



Multiplateforme

Même workflow sur macOS, Windows et Linux — installateurs autonomes.



Pensé pour le français

Modèles spaCy et CamemBERT adaptés aux textes qualitatifs en français.



Contrôle humain

Vous validez, corrigez et désactivez entité par entité avant l'export.



Traçabilité

Rapport d'audit et exports prêts pour l'archivage ou Label Studio.



Assistant d'anonymisation, pas une garantie d'anonymat total. Relisez toujours le résultat avant diffusion.

Incognito - Anonymisation rapide

Utile pour

1. Quelques textes

Idéal pour le traitement rapide de petits volumes de données sans configuration complexe.

2. Expérimenter des modèles

Parfait pour tester différentes approches d'anonymisation et ajuster les paramètres de détection.

Incognito - Interactivité renforcée



**Modification par
catégorie, entité**



**Modification directe sur le
texte**



Preview en temps réel

Incognito - Portabilité vers Label Studio pour des corrections avancées

<https://labelstud.io/>



Label Studio

Use Cases ▾

Learn ▾

Docs ▾

Integrations

Enterprise ▾

Q Search

 27,682

Quick Start

Label any  data.
Evaluate any AI.

Incognito - Traitement de données en masse

Pour chaque texte



1. Configuration du modèle

Possibilité de choisir un modèle distinct pour une personnalisation accrue du traitement.



2. Suivi en temps réel

Mise à jour instantanée des textes anonymisés et génération automatique des rapports d'anonymisation.

Incognito - Fonctions prévues

1. Rapport d'audit enrichi produit par un LLM local

Risque global : Élevé

Facteurs détectés :

- ✓ Localisation précise
- ✓ Profession rare
- ✓ Organisation identifiable
- ✓ Fonction unique

Facteurs absents :

- ✗ Coordonnées personnelles
- ✗ Données médicales
- ✗ Données financières

2. Plus de granularité dans des méthodes d'anonymisation



Noms propres

Martin → Jean



Professions

Professeur de piano → Travaille dans l'enseignement



Âges et dates

J'ai 23 ans → J'ai 20-30 ans.

Incognito - Conclusion

Bugs/avis ?

<https://github.com/xiaoouwang/Incognito/issues>

Références

“Anonymisation des données de recherche en SHS.” Accessed June 24, 2026.

<https://www.misha.fr/agenda/anonymisation-des-donnees-de-recherche-en-shs/>.

Brunel, Valentin. “L’anonymisation et ses enjeux : une courte sensibilisation.” Paper presented at Droit des données de recherche : Du juridique à l’éthique.

June 9, 2026. <https://hal.science/hal-05655702>.

Brunel, Valentin. “Les Pratiques d’anonymisation En SHS.” Paper presented at Journée d’étude : pleins phares sur l’anonymisation. September 24, 2024.

<https://hal.science/hal-04720171>.

Contacts



Guichet unique

Pour toutes les demandes liées au cycle de vie des données (en me mettant en copie)

✉ recherche.datazur@univ-cotedazur.fr



Questions spécifiques aux SHS et à Huma-Num

Xiaoou Wang

☎ 04 89 15 20 39

✉ Xiaoou.WANG@univ-cotedazur.fr

📍 MSHS Sud-Est, bureau 125 (1er étage)

Faites parler votre corpus ! (RAG)

Faires parler vos corpus : créer votre propre GPT personnalisé à partir de vos données.

Application pratique :

 **Corpus** : la presse kényane des années 80.

 **Questions de recherche** : Comment la société est représentée pendant cette période ? (Fragmentation, composition etc.)

...etc.