

UNIVERSITÉ NORBERT ZONGO

UFR-Sciences et Technologies

Département d'Informatique



Burkina Faso

*Unité - Progrès - Justice*

Année académique : 2022 - 2023

## RAPPORT DE PROJET TUTORÉ

Pour l'obtention du diplôme de Licence

Filière : Mathématique-Physique-Chimie-Informatique

Spécialité : Informatique

### Implémentation d'algorithmes permettant l'anonymisation des données

Présenté par : **Landri BAYILI**

INE : **E03515920201**

Encadré par :

**Dr Moustapha BIKIENGA**, Enseignant-chercheur à l'Université Norbert ZONGO

**Dr Dimitri OUATTARA**, Enseignant-chercheur à l'Université Joseph KI-ZERBO

Juin 2024



---

## DEDICACES

**Je** dédie ce travail à ma famille.

*Landri* BAYILI



---

## REMERCIEMENTS

Je remercie Dieu de m'avoir conduit tout au long de ce travail.

Je tiens à exprimer ma profonde gratitude à Dr Moustapha BIKIENGA, enseignant-chercheur à l'université Norbert Zongo, et à Dr Dimitri OUATTARA, enseignant-chercheur à l'université Joseph KI-ZERBO. Leur confiance en moi, ainsi que leur encadrement et leurs conseils, ont été inestimables pour la réalisation de ce projet.

J'exprime mes profonds remerciements à tout le personnel enseignant et administratif de l'Université Norbert Zongo en général, et à ceux de l'UFR/ST en particulier car, sous l'ombre de vos ailes, nous avons tout appris.

Je remercie aussi ma famille, mes amis, mes camarades de classe, et tous ceux qui ont contribué à l'élaboration de ce document et à la réussite de cette année universitaire.



---

# TABLE DES MATIÈRES

<b>LISTE DES FIGURES</b>	<b>vi</b>
<b>LISTE DES TABLEAUX</b>	<b>vii</b>
<b>LISTE DES ABRÉVIATIONS</b>	<b>viii</b>
<b>INTRODUCTION GÉNÉRALE</b>	<b>1</b>
<b>1 ÉTAT DE L'ART</b>	<b>3</b>
INTRODUCTION . . . . .	5
1.1 DÉFINITION ET CONCEPTS FONDAMENTAUX . . . . .	5
1.2 MODÈLES D'ATTAQUES DES DONNÉES TABULAIRES PUBLIÉES . . . .	7
1.3 MAXIMISER L'UTILITÉ DES DONNÉES TOUT EN PRÉSERVANT L'ANONYMAT POUR LA RECHERCHE . . . . .	9
1.4 PARADIGMES DE PROTECTION DE LA VIE PRIVÉE . . . . .	10
1.4.1 Modèle k-anonymat . . . . .	11
1.4.2 Modèle l-diversité . . . . .	12
1.4.3 Modèle t-proximité . . . . .	14
1.4.4 Modèle $\delta$ -présence . . . . .	14
1.5 ÉTUDES DES TECHNIQUES D'ANONYMISATION DES DONNÉES . . . .	16
1.5.1 Famille de généralisation des données . . . . .	16
1.5.2 Famille de randomisation . . . . .	17
1.6 ALGORITHMES DE GÉNÉRALISATION . . . . .	17
1.6.1 Principe des algorithmes de généralisation des données . . . . .	17
1.6.2 Algorithme $\mu$ -Argus . . . . .	19
1.6.3 Algorithme Datafly . . . . .	20
1.6.4 Algorithme de Samarati . . . . .	20
1.6.5 Algorithme «Bottom up generalization» . . . . .	22
1.6.6 Algorithme Incognito . . . . .	23
1.6.7 Algorithme Median Mondrian . . . . .	23

1.6.8	Algorithmes « InfoGain Mondrian » et « LSD Mondrian » . . . . .	24
1.7	ALGORITHMES DE RANDOMISATIONS . . . . .	25
1.7.1	Ajout de bruit . . . . .	25
1.7.2	Permutation . . . . .	26
1.7.3	Confidentialité différentielle . . . . .	27
1.8	QUELQUES OUTILS D'ANONYMISATION . . . . .	27
1.8.1	Outil $\mu$ -Argus . . . . .	27
1.8.2	ARX Data Anonymization Tool . . . . .	28
1.9	SYNTHÈSE . . . . .	30
	CONCLUSION . . . . .	30
<b>2</b>	<b>CONCEPTION DU SYSTÈME LOGICIEL</b>	<b>32</b>
	INTRODUCTION . . . . .	33
2.1	ANALYSE DES BESOINS . . . . .	33
2.1.1	Besoins fonctionnels . . . . .	33
2.1.2	Besoins non fonctionnels . . . . .	34
2.2	ARCHITECTURE DU SYSTÈME LOGICIEL . . . . .	34
2.2.1	Diagramme de cas d'utilisation . . . . .	35
2.2.2	Diagramme de classe . . . . .	35
	CONCLUSION . . . . .	36
<b>3</b>	<b>IMPLÉMENTATION DU SYSTÈME LOGICIEL</b>	<b>38</b>
	INTRODUCTION . . . . .	39
3.1	OUTILS UTILISES . . . . .	39
3.1.1	Technologies utilisées . . . . .	39
3.1.2	IDE de développement : Java Eclipse . . . . .	40
3.2	OBJECTIF DE L'OUTIL . . . . .	43
3.3	POURQUOI L'ALGORITHME DATAFLY ET LE MODÈLE K-ANONYMAT? . . . . .	44
3.3.1	Algorithme DataFly . . . . .	44
3.3.2	Modèle k-anonymat . . . . .	45
3.4	ORGANISATION DE L'OUTIL . . . . .	46
3.5	UTILISATION DE L'OUTIL : CAS PRATIQUE . . . . .	47
3.5.1	Dataset étudiant . . . . .	48
3.5.2	Importer une base de données . . . . .	49
3.5.3	Choix du paramètre d'anonymisation . . . . .	53
3.5.4	Construction de la hiérarchie de généralisation . . . . .	53

## TABLE DES MATIÈRES

---

3.5.5 Comparaison des résultats . . . . .	54
CONCLUSION . . . . .	55
<b>CONCLUSION GÉNÉRALE</b>	<b>56</b>
<b>Bibliographie</b>	<b>58</b>
<b>ANNEXES</b>	<b>61</b>
A.1 CODE DE DÉTECTION DES CATÉGORIES DE DONNÉES . . . . .	61

## LISTE DES FIGURES

1.1	Collecte et publication des données . . . . .	6
1.2	Taxonomie des modèles d'attaque de la vie privée . . . . .	8
1.3	Hiérarchie de généralisation de l'attribut Sexe . . . . .	18
1.4	Hiérarchie de généralisation de l'attribut Code postal . . . . .	18
1.5	Hiérarchie de généralisation de l'attribut Niveau d'étude . . . . .	19
1.6	Processus Centrale de L'algorithme Datafly . . . . .	20
1.7	Treillis de généralisation des attributs Sexe et Code postal . . . . .	21
1.8	Algorithme bottom-up generalisation . . . . .	22
1.9	Algorithme incognito . . . . .	23
1.10	Processus d'anonymisation de ARX . . . . .	29
2.1	Diagramme de cas d'utilisation . . . . .	35
2.2	Diagramme de classe . . . . .	36
3.1	IDE de développement Eclipse . . . . .	41
3.2	Interface graphique d'éclipse . . . . .	42
3.3	Maquette de fenêtre sous Éclipse . . . . .	43
3.4	Les étapes de simulations de <b>PFCL_Anonymization</b> . . . . .	47
3.5	Interface graphique de <b>PFCL_Anonymization</b> . . . . .	48
3.6	Import de base de donnée . . . . .	49
3.7	Fenêtre présentant le système de fichier de votre ordinateur . . . . .	50
3.8	Téléchargement de la base de donnée . . . . .	50
3.9	Base de données importé . . . . .	51
3.10	Obtention des types de données . . . . .	51
3.11	Sélection des attributs . . . . .	52
3.12	Interface du choix de k . . . . .	53
3.13	Résultat final de l'algorithme . . . . .	54
A.1	Code pour marquer les attributs sensibles . . . . .	61
A.2	Code pour marquer les attributs Identifiants . . . . .	61
A.3	Code pour marquer les attributs QI . . . . .	61



---

## LISTE DES TABLEAUX

1.1	Exemple de table Originale . . . . .	10
1.2	Table qui satisfait le 2-anonymat . . . . .	11
1.3	Base de données originale . . . . .	13
1.4	Table 2 qui satisfait la 2-diversité . . . . .	13
1.5	Table qui ne satisfait pas le 2-anonymat . . . . .	15
1.6	Table qui satisfait le 3-anonymat . . . . .	15
1.7	Table originale . . . . .	18
1.8	Donnée généralisée selon $\langle S1, Z1 \rangle$ . . . . .	21
1.9	Résumé sur les modèles de protection de la vie privée . . . . .	30





---

## LISTE DES ABRÉVIATIONS

- API** Interface de Programmation d'Application
- AS** Attribut Sensible
- CASC** Computational Aspects of Statistical Confidentiality
- CIL** Commission de l'Information et des Libertés
- CNIB** Carte Nationale d'Identité Burkinabè
- IDE** Environnement de Développement Intégré
- IE** Identifiant Explicite
- INE** Identifiant National Étudiant
- LSD Mondrian** Least Square Deviance Mondrian
- MSE** Mean Squared Error
- PPDM** Privacy Preserving Data Mining
- PPDP** Privacy Preserving Data Publishing
- QI** Quasi-Identifiant
- SDC** Statistical Disclosure Control
- ST** Sciences et Technologies
- UFR** Unité de Formation et de Recherche



---

## INTRODUCTION GÉNÉRALE

Les données jouent un rôle essentiel dans le développement scientifique et l'innovation, tant pour les entités publiques que privées. Elles sont une ressource fondamentale pour la production de biens et de services, fournissant des informations cruciales pour la prise de décision permettant ainsi aux organisations de maintenir leurs compétitivités. La collecte et le stockage des données restent aujourd'hui, des activités stratégiques de premier plan.

L'essor du numérique et les progrès technologiques ont facilité le partage des données au-delà des frontières organisationnelles, entraînant une augmentation considérable du volume de données disponibles. Cette tendance a été encouragée par la promotion croissante de l'« open data » par de nombreux pays, ce qui soulève des préoccupations quant à la protection de la vie privée, notamment en ce qui concerne les données personnelles sensibles.

La Commission de l'informatique et des libertés (CIL) au Burkina Faso régule la protection des données personnelles et stipule que toute donnée à caractère personnel ne peut être partagée sans le consentement des individus concernés. Cela souligne le fait que les individus sont les seuls propriétaires de leurs informations et ont des droits sur celles-ci. Cependant, étant donné que les données sont désormais intrinsèquement liées à l'innovation et au développement scientifique et technologique, il est crucial de trouver une solution qui concilie l'exploitation des données avec la protection de l'identité des individus. Cette solution doit répondre à deux exigences essentielles : permettre l'exploitation et la disponibilité des données tout en garantissant une protection robuste de l'identité des individus concernés. Pour ce faire, **l'anonymisation de données** émerge comme une réponse prometteuse à ces deux contraintes [1]. La **LOI N°001-2021/AN portant protection des personnes à l'égard du traitement des données à caractère personnel au Burkina Faso** définit l'anonymisation comme « *traitement qui consiste à utiliser un ensemble*

*de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible*». Cette définition laisse clairement voir une large différence entre anonymiser et pseudonymiser, Bien que les deux méthodes préservent la confidentialité des données, le deuxième processus est réversible[2].

Face à la reconnaissance de l'importance de la protection de la vie privée, la communauté des chercheurs, principalement constituée de statisticiens et d'informaticiens, s'est concentrée sur plusieurs aspects. Tout d'abord, elle a travaillé sur la définition des modèles de protection de la vie privée. Ensuite, elle a élaboré des algorithmes visant à dé-identifier les données sensibles(informations sensibles) tout en préservant leur utilité.

Plusieurs outils existent déjà pour mettre en place un processus complet d'anonymisation des données. Cependant, ils présentent souvent des difficultés d'utilisation et leur documentation est insuffisante pour en comprendre le fonctionnement sans une expertise approfondie dans le domaine. Pour répondre à cette problématique, notre projet de fin de cycle vise à implémenter un système logiciel robuste, efficace et évolutif pour assurer un processus d'anonymisation de qualité[3]. Notre outil contribuera à l'anonymisation des données sur les performances académiques des étudiants tout en simplifiant le processus pour les utilisateurs. Notre objectif est de fournir un programme fonctionnel au corps éducatif, permettant ainsi de partager des données anonymes de manière sécurisée et conforme aux réglementations sur la protection de la vie privée au burkina faso.

Notre travail se structurera autour de trois chapitres. Le premier chapitre explore les modèles d'attaques et de protection de la vie privée, fournissant ainsi un cadre théorique pour notre approche. Le second chapitre est consacré à la conception du système logiciel, détaillant son architecture et ses fonctionnalités. Enfin, le troisième chapitre concerne l'implémentation du système.

## Chapitre

**1**

---

**ÉTAT DE L'ART**

---

Sommaire

---

<b>INTRODUCTION . . . . .</b>	<b>5</b>
<b>1.1 DÉFINITION ET CONCEPTS FONDAMENTAUX . . . . .</b>	<b>5</b>
<b>1.2 MODÈLES D'ATTAQUES DES DONNÉES TABULAIRES PUBLIÉES . . . . .</b>	<b>7</b>
<b>1.3 MAXIMISER L'UTILITÉ DES DONNÉES TOUT EN PRÉSERVANT L'ANONYMAT POUR LA RECHERCHE</b>	<b>9</b>
<b>1.4 PARADIGMES DE PROTECTION DE LA VIE PRIVÉE .</b>	<b>10</b>
1.4.1 Modèle k-anonymat . . . . .	11
1.4.2 Modèle l-diversité . . . . .	12
1.4.3 Modèle t-proximité . . . . .	14
1.4.4 Modèle $\delta$ -présence . . . . .	14
<b>1.5 ÉTUDES DES TECHNIQUES D'ANONYMISATION DES DONNÉES . . . . .</b>	<b>16</b>
1.5.1 Famille de généralisation des données . . . . .	16
1.5.2 Famille de randomisation . . . . .	17
<b>1.6 ALGORITHMES DE GÉNÉRALISATION . . . . .</b>	<b>17</b>
1.6.1 Principe des algorithmes de généralisation des données . . . . .	17
1.6.2 Algorithme $\mu$ -Argus . . . . .	19
1.6.3 Algorithme Datafly . . . . .	20
1.6.4 Algorithme de Samarati . . . . .	20
1.6.5 Algorithme «Bottom up generalization» . . . . .	22
1.6.6 Algorithme Incognito . . . . .	23
1.6.7 Algorithme Median Mondrian . . . . .	23
1.6.8 Algorithmes « InfoGain Mondrian » et « LSD Mondrian » . . .	24
<b>1.7 ALGORITHMES DE RANDOMISATIONS . . . . .</b>	<b>25</b>
1.7.1 Ajout de bruit . . . . .	25
1.7.2 Permutation . . . . .	26
1.7.3 Confidentialité différentielle . . . . .	27
<b>1.8 QUELQUES OUTILS D'ANONYMISATION . . . . .</b>	<b>27</b>
1.8.1 Outil $\mu$ -Argus . . . . .	27

1.8.2	ARX Data Anonymization Tool . . . . .	28
<b>1.9</b>	<b>SYNTHÈSE . . . . .</b>	<b>30</b>
	<b>CONCLUSION . . . . .</b>	<b>30</b>

---

# INTRODUCTION

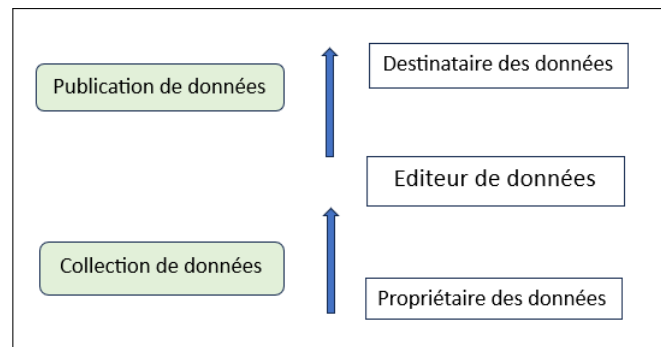
L'avènement du numérique a entraîné la collecte et la sauvegarde massive des données, dans le but de les exploiter pleinement à des fins spécifiques. Cependant, leurs disponibilités, pour les chercheurs et les professionnels du domaine de la protection des données personnelles, pose des défis en termes de respect de la vie privée, car ces données peuvent contenir des informations sensibles[4]. Dans ce contexte, l'anonymisation des données émerge comme une solution permettant de concilier l'utilité des données avec le respect de la vie privée. Ce chapitre fournira une vue d'ensemble des concepts fondamentaux entourant le sujet d'anonymisation des données tabulaires.

## 1.1 DÉFINITION ET CONCEPTS FONDAMENTAUX

L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible[5]. Ce traitement garantit la protection de la vie privée pour des fins de publication des données car une fois passées par un éditeur pour un processus d'anonymisation (transformation), l'identification d'une personne à partir des données anonymisées (transformées) devient presque impossible, quelle que soit la technique de ré-identification utilisée. Un bon processus d'anonymisation de données pour des fins de publication est constitué de deux phases (phase de collecte et phase de publication) et fait obligatoirement intervenir trois acteurs [6] :

- le propriétaire des données : toute personne dont les données ont été collectées,
- le destinataire des données : toute personne ou groupe de personnes qui doit recevoir les données anonymisées,
- un éditeur de données : c'est le système logiciel qui est chargé d'automatiser le processus d'anonymisation des données.

Nous proposons la **Figure 1.1** qui représente de façon visuelle, notre explication textuelle.



**FIGURE 1.1 – Collecte et publication des données**

La protection de la vie privée dans un contexte de publication de données implique la non-divulgateur d'informations sensibles (informations que l'on ne souhaite pas partager), suite aux autorisations d'accès fournies aux destinataires [7]. Le principal problème du partage de données contenant des informations sensibles réside dans le fait qu'un tiers peut les utiliser pour porter atteinte à la vie privée des personnes concernées. Les propriétaires des données sont considérés comme victimes dès lors qu'ils n'ont pas été informés avant la divulgation de leurs informations.

L'anonymisation ouvre la porte à la réutilisation des données initialement interdites du fait de leur caractère personnel, permettant ainsi aux acteurs d'exploiter et de partager leurs « gisements » de données sans porter atteinte à la vie privée des personnes[5]. **On peut donc commencer à se demander, comment le processus d'anonymisation fonctionne-t-il ? Comment est-il possible que les données soient détachées de leurs caractères identifiants tout en restant utiles pour des fins de recherche ?** Avant de répondre à toutes ces interrogations, nous allons tout d'abord expliquer les raisons pour lesquelles la sécurité d'aujourd'hui entraîne l'implication des données personnelles.

## **Catégories de données dans une Base de Données tabulaire**

Une base de données tabulaire est un ensemble de données organisé en lignes et en colonnes où chaque type d'information est toujours placé au même endroit. Les données stockées dans une base de données tabulaire sont regroupées en quatre catégories[8]. Ce sont les :

1. **identifiants explicites** qui représentent les caractéristiques individuelles d'une personne par exemple, le numéro INE de l'étudiant, le nom patronymique. En pratique, cette définition n'est pas stricte, car il est possible de trouver des personnes partageant le même identifiant explicite,
2. **quasi-identifiants** qui désignent un groupe d'attributs qui lorsqu'ils sont combinés, peuvent potentiellement permettre l'identification indirecte d'au moins une personne parmi les individus décrits dans une base de données,
3. **identifiants sensibles** qui évoquent les données que les individus souhaitent généralement cacher, telles que les notes, les informations disciplinaires, ou les antécédents médicaux étudiants,
4. **identifiants non-sensibles** qui montrent des données qui n'entrent dans aucune des catégories précédentes.

**NB :** parfois, vous entendrez parler d'attribut sensible ; ne vous laissez pas perturber, car les termes attribut et identifiant gardent le même sens, sauf mention contraire.

## 1.2 MODÈLES D'ATTAQUES DES DONNÉES TABULAIRES PUBLIÉES

Dans le domaine de la protection de la vie privée, après la publication des données, un attaquant utilise généralement des stratégies basées sur sa connaissance du contexte pour accéder aux informations sensibles. Ces stratégies, aussi appelées modèles d'attaque [6], permettent à l'attaquant de déduire des informations sensibles sur une victime en établissant des liens ou en effectuant des inférences probabilistes, voir la Figure 1.2.



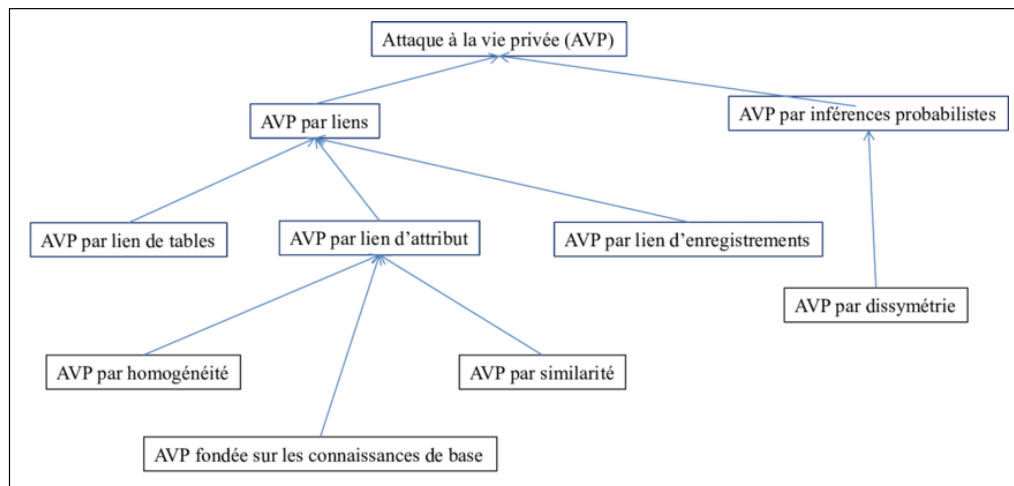


FIGURE 1.2 – Taxonomie des modèles d'attaque de la vie privée

La **Figure 1.2** fournit une taxonomie des attaques liées à la vie privée. Nous n'allons pas expliquer l'ensemble des types d'attaques dans ce document. Nous allons nous contenter des attaques les plus courantes qui sont :

- **AVP par liens** (le modèle d'attaque par liens) : dans ce modèle, l'attaquant cible une victime spécifique en connaissant son quasi-identifiant, c'est-à-dire une caractéristique distinctive permettant de l'identifier indirectement,
- **AVP par lien de table** (« table linkage ») : on parle de ce type d'attaque lorsque l'adversaire n'a pas initialement connaissance de la présence des données de sa victime dans une table, mais peut faire une déduction à partir des informations qu'il y observe,
- **AVP par liens d'attribut** (le scénario du « lien d'attribut ») : pour ce type d'attaque, le risque de ré-identification persiste car l'attaquant, connaissant le quasi-identifiant (QI) de la victime, peut identifier le groupe d'individus partageant le même QI. Cette identification permet à l'attaquant de déduire les informations sensibles de la victime en se basant sur les valeurs sensibles du groupe auquel elle appartient. Si toutes les valeurs pour un identifiant sensible sont identiques au sein d'un groupe, la déduction est directe, constituant ainsi une attaque par homogénéité,
- **AVP par lien d'enregistrements** (le scénario du « lien d'enregistrements ») : dans ce modèle, l'attaquant, en plus de connaître le quasi-identifiant de la victime, sait également que les informations la concernant sont incluses dans la table publiée,

- **AVP par inférences probabilistes** : ici, l'attaquant ne cherche pas à établir des liens directs avec des tables, des enregistrements ou des attributs sensibles. Au lieu de cela, il se base sur ses croyances probabilistes avant et après analyse de la distribution des valeurs des attributs sensibles dans la table publiée.

Dans le but de contrer ces attaques, plusieurs modèles sont proposés dans la littérature. Nous examinerons ces modèles dans les sections suivantes.

### Optimisation de l'anonymisation pour préserver l'utilité des données

Un processus d'anonymisation consiste à établir un équilibre entre deux contraintes principales :

- ☛ **ne pas dénaturer ni appauvrir la qualité des données** : les données doivent être suffisamment utiles pour des fins de recherche afin de produire de bons résultats,
- ☛ **rendre les données inutilisables pour « un attaquant supposé » qui tenterait de porter atteinte à la vie privée des personnes concernées** : l'anonymisation permet de protéger la vie privée en remplaçant les informations spécifiques permettant l'identification des individus par des informations plus générales ou agrégées.

L'anonymisation des données suit un processus complexe. Tout outil établissant un processus d'anonymisation doit obligatoirement tenir compte des contraintes ci-dessus. Ces contraintes étant plus ou moins contradictoires, il est parfois difficile d'élaborer un processus d'anonymisation sans risque de ré-identification, surtout lorsqu'on connaît finement le domaine étudié. Il existe par conséquent des modèles d'anonymisation de données élaborés par des professionnels des différents domaines [7] : domaine de la santé, domaine de recherche, etc.

## 1.3 MAXIMISER L'UTILITÉ DES DONNÉES TOUT EN PRÉSERVANT L'ANONYMAT POUR LA RECHERCHE

L'anonymisation vise à maintenir l'utilité des données tout en protégeant la vie privée. Une mauvaise anonymisation pourrait altérer la véracité des résultats obtenus à partir des données partagées, soulignant ainsi l'importance d'une transformation adéquate pour atteindre ses objectifs.

La **Table 1.1** présente un exemple d'une base de données tabulaire. Les catégories des données sont également mentionnées.

IE	QI		AS
Nom	Age	Niveau	Maladie
Jonh	19	Licence 2	maladie cardiaque
Jean	19	Licence 3	cancer
Alice	27	Licence 3	grippe
David	30	Licence 3	grippe
Bob	23	Terminale	cancer
Dupont	23	Terminale	cancer

**TABLE 1.1 – Exemple de table Originale**

Dans cette base de données tabulaire,

- chaque **Nom** permet d'indexer une personne unique, c'est-à-dire que deux personnes ne portent pas le même nom. On dit que l'attribut nom est un identifiant explicite (**IE**),
- l'attribut **Niveau** ou **Age**, à lui seul ne permet pas d'identifier une personne unique dans une base de données de grande taille. Lorsqu'on fait des liens entre attributs dans le but de remonter à une personne unique, on parle d'attribut quasi-identifiant (**QI**).
- l'attribut **Maladie** est une donnée que personne ne souhaite partager au grand public ou même à sa famille. On dit dans ce cas qu'il s'agit d'un attribut sensible (**AS**).

## 1.4 PARADIGMES DE PROTECTION DE LA VIE PRIVÉE

Les travaux des chercheurs ont permis de construire plusieurs modèles permettant d'assurer un bon équilibre entre protection de la vie privée et utilité des données. Dans ce document, nous allons nous limiter aux modèles les plus couramment utilisés tels que :

- modèle k-anonymat,
- modèle l-diversité,
- modèle t-proximité,
- modèle  $\delta$ -présence.

### 1.4.1 Modèle k-anonymat

Le k-anonymat permet une protection de la vie privée en empêchant la reconnaissance d'individus et en s'assurant qu'au moins  $k$  individus (avec  $k > 1$ ) possèdent des caractéristiques similaires dans une base de données anonymisée [9]. Il vise à protéger les données en garantissant qu'une personne disposant d'une base de données ne puisse pas établir de liens sémantiques indubitables entre quasi-identifiants c'est-à-dire, établir des liens uniquement entre attributs quasi-identifiants pour distinguer une personne spécifique dans la base de données anonymisée.

L'implémentation de ce modèle offre l'assurance que les quasi-identifiants de chaque ligne de la base de données apparaissent au moins  $k$  fois dans la table anonymisée [7]. Par conséquent, même si un attaquant a connaissance des quasi-identifiants d'un individu, il ne pourra plus l'identifier dans la table anonymisée [10]. Le k-anonymat permet de protéger la vie privée en empêchant les attaques basées sur les liens d'attributs. Il se focalise principalement sur les données quasi-identifiantes en modifiant les relations entre l'ensemble des données et les sujets des données [7]. Le **Tableau 1.2** montre l'application du k-anonymat à une table. La valeur  $k$  est appelée **degré d'anonymisation**.

Age	Éducation	Maladie
[19,23]	Secondaire	maladie cardiaque
[19,23]	Secondaire	cancer
[27,30]	Secondaire	grippe
[27,30]	Secondaire	grippe
[19,23]	Supérieur	cancer
[23,23]	Supérieur	cancer
[19,23]	Supérieur	cancer

**TABLE 1.2 – Table qui satisfait le 2-anonymat**

On dit que la **Table 1.2** satisfait le 2-anonymat car dans cette table, au moins deux personnes possèdent des quasi-identifiants identiques

### 1.4.2 Modèle l-diversité

Lorsque le modèle mis en œuvre est le k-anonymat, une connaissance des quasi-identifiants de la victime, facilite la ré-identification des données anonymisées. Le modèle **l-diversité** est une solution visant à contrer cette faille. Même si un attaquant possède des informations sur les quasi-identifiants d'une table anonymisée, la l-diversité protège les attributs sensibles de telle sorte que l'on ne puisse pas déterminer avec certitude quelles données sensibles appartiennent à un individu particulier.

La **l-diversité** vise à contrer les attaques d'homogénéité. Ces attaques permettent de déduire des valeurs pour d'autres attributs à partir de certaines valeurs d'attributs sources. Similaire au k-anonymat, une table est dite **l-diverse** si chaque groupe d'individus ayant des caractéristiques similaires pour les quasi-identifiants possède au moins **l** valeurs distinctes pour un attribut sensible donné [10].

Le principe de la l-diversité est le suivant : *une classe d'équivalence (table k-anonymisée) respecte la l-diversité s'il existe au moins l valeurs "bien représentées" pour l'attribut sensible. Une table respecte la l-diversité si chacune de ses classes d'équivalence (groupe de quasi-identifiants similaires) respecte la l-diversité.* Cette définition peut sembler complexe, car plusieurs interprétations peuvent découler de l'expression "bien représentée". En effet, une table peut disposer d'un ou de plusieurs attributs sensibles, ce qui permet de distinguer différents modèles de la **l-diversité** [7].

La l-diversité est donc un modèle d'anonymisation visant à protéger la vie privée des individus en garantissant une diversité minimale dans les données sensibles [6].

Age	Sexe	Code postal	Maladie
24	F	75001	Diabète
36	M	75003	Asthme
43	F	75005	Hypertension
29	M	75001	Diabète
38	F	75003	Asthme
41	M	75005	Hypertension
31	F	75001	Diabète
49	M	75003	Asthme
57	F	75006	Hypertension
27	M	75001	Diabète

**TABLE 1.3 – Base de données originale**

Dans le **Tableau 1.3**, Une personne connaissant le code postal et la maladie d'une personne peut l'identifier facilement. En fixant  $l$  à 2, la  $l$ -diversité s'assure qu'un groupe partageant les mêmes caractéristiques inclut au moins deux maladies différentes, réduisant ainsi le risque d'identification.

Age	Sexe	Code postal	Maladie1	Maladie2
24	F	75001	Diabète	Asthme
36	M	75003	Asthme	Hypertension
43	F	75005	Hypertension	Diabète
29	M	75001	Diabète	Hypertension
38	F	75003	Asthme	Diabète
41	M	75005	Hypertension	Asthme
31	F	75001	Diabète	Hypertension
49	M	75003	Asthme	Hypertension
57	F	75006	Hypertension	Hypertension
27	M	75001	Diabète	Hypertension

**TABLE 1.4 – Table 2 qui satisfait la 2-diversité**

Cependant, la  $l$ -diversité ne constitue pas une garantie suffisante contre les attaques par similarité, les attaques par inférences probabilistes, notamment celles par dissymétrie[7].

### 1.4.3 Modèle t-proximité

Il est possible de protéger une table en utilisant la l-diversité, mais cela ne suffit pas à empêcher la déduction d'informations sensibles. Le principe de la t-proximité est plus approprié dans ce contexte, car il limite l'obtention d'informations sensibles sur une population ciblée. La t-proximité s'assure que la distribution de l'attribut sensible dans chaque classe d'équivalence soit proche de la distribution globale de cet attribut. Pour respecter la t-proximité, la distance entre la distribution de l'attribut sensible dans une classe et la distribution globale ne doit pas dépasser un seuil prédéfini  $t$ . Une table respecte la t-proximité si toutes ses classes d'équivalence respectent ce critère.

Il existe plusieurs mesures de distance pour évaluer la proximité entre distributions dans le cadre de la t-proximité. Les deux techniques les plus couramment utilisées sont la distance euclidienne et la divergence de Kullback-Leibler (KL). La distance euclidienne est une mesure simple qui calcule la distance "en ligne droite" entre deux points dans un espace multidimensionnel. Elle est souvent utilisée pour des données numériques, telles que des histogrammes représentant la distribution de l'âge dans différents groupes. La divergence de Kullback-Leibler (KL) est une mesure de divergence entre deux distributions de probabilité. Elle quantifie la différence entre ces distributions en termes d'entropie. La divergence KL est asymétrique et nulle si et seulement si les deux distributions sont identiques.

### 1.4.4 Modèle $\delta$ -présence

Avec les modèles proposés ci-dessus, une meilleure anonymisation d'un ensemble de données est normalement satisfaite. Mais un problème persiste toujours car, il est possible pour les super attaquants d'établir des liens entre différentes tables pour inférer la présence d'un enregistrement dans une table, il est donc nécessaire de trouver un contre-modèle d'où le modèle  **$\delta$ -présence**. Dans la littérature, le **modèle  $\delta$ -présence** a pour objectif de contrer les attaques par «lien de tables»[11].

En effet, il est possible de publier plusieurs tables anonymisées par des éditeurs différents. Cependant, on ne peut pas exclure les possibilités de liens entre ces tables dès lors qu'elles partagent des attributs quasi-identifiants [7]. Voici un ensemble d'exemple qui explique le principe du modèle  $\delta$ -présence :

supposons que le **Tableau 1.5** sur les maladies a été publié au même titre que le **Tableau 1.6** sur les catégories professionnelles.

Age	Éducation	Maladie
[19,23]	Secondaire	maladie cardiaque
[19,23]	Secondaire	cancer
[27,30]	Secondaire	grippe
[27,30]	Secondaire	grippe
[19,23]	Supérieur	cancer
[23,23]	Supérieur	cancer
[19,23]	Supérieur	cancer

**TABLE 1.5 – Table qui ne satisfait pas le 2-anonymat**

Age	Éducation	Nom
[19,23]	Supérieur	Malik
[19,23]	Supérieur	George
[27,30]	Supérieur	Fred
[27,30]	Supérieur	Jean
[19,23]	Supérieur	Pierre
[27,30]	Supérieur	Paul
[19,23]	Supérieur	Alice

**TABLE 1.6 – Table qui satisfait le 3-anonymat**

Les Tables 1.5 et 1.6 révèlent respectivement la tranche d'âge et le niveau d'éducation des individus. Si un attaquant possède ces deux tables, le risque de ré-identification des données est très élevé car ces deux tables partagent des quasi-identifiants identiques. Prenons l'exemple d'Alice comme victime de l'attaque. Supposons qu'elle a un âge compris entre 19 et 23 ans



et qu'elle possède un niveau d'études supérieur. En rapprochant ces deux tables, l'attaquant peut déduire qu'Alice a une probabilité de  $3/4$ , soit 75% d'être atteinte d'un cancer (le chiffre 3 correspond à la taille de la classe d'équivalence du QI « [19, 23], Supérieur » dans la table 1.6 et 4 correspond à celle du même QI dans la table 1.5).

Pour prévenir la possibilité d'inférence de la présence d'un enregistrement dans une base de données publiée,  $\delta$ -présence impose que la probabilité de présence d'un enregistrement soit comprise entre un intervalle  $\delta$  ( $\delta_{min}$ ,  $\delta_{max}$ ) prédéfini.

## 1.5 ÉTUDES DES TECHNIQUES D'ANONYMISATION DES DONNÉES

Le travail des chercheurs a permis d'élaborer plusieurs techniques d'anonymisation. Ces techniques sont regroupées en deux grandes familles à savoir la :

- **généralisation des données,**
- **randomisation des données.**

### 1.5.1 Famille de généralisation des données

Généraliser un attribut implique le remplacement des données spécifiques par des données plus générales. Par exemple, une ville peut être remplacée par sa région, et une semaine par un mois. Bien que la généralisation puisse empêcher l'individualisation, elle ne garantit pas une anonymisation complète et nécessite des approches quantitatives sophistiquées pour éviter la corrélation et l'inférence[12]. Cette technique est largement explorée et mise en œuvre par divers algorithmes. Ces algorithmes visent non seulement à protéger la vie privée, mais aussi à préserver la qualité des données anonymisées. Ils sont utilisés dans des logiciels commerciaux et des prototypes de recherche, dont l'utilisation peut être complexe.

### 1.5.2 Famille de randomisation

La **randomisation** consiste à modifier les attributs dans un jeu de données de telle sorte qu'ils soient moins précis, tout en conservant la répartition globale. Cette technique permet de protéger le jeu de données du risque d'inférence [5].

La randomisation anonymise les données en les modifiant aléatoirement pour éviter qu'elles ne soient associées à des individus spécifiques. Cette méthode préserve l'anonymat tout en maintenant la qualité et l'utilité des données pour l'analyse et la recherche.

## 1.6 ALGORITHMES DE GÉNÉRALISATION

Les algorithmes de généralisation sont des moyens qui permettent d'implémenter les différents modèles d'anonymisation. Ces algorithmes sont très utiles dans l'élaboration d'un outil d'anonymisation car, ils impliquent l'utilisation de stratégies adéquates pour l'implémentation des modèles présentés ci-dessus [12].

Plusieurs algorithmes de généralisation sont proposés dans la littérature. Dans ce document, nous n'allons pas tout étudier mais nous verrons les plus exploités.

Les algorithmes de généralisation les plus connus sont : «  $\mu$ -argus », « Datafly », l'algorithme de Samarati, « Incognito », « Median Mondrian », « Infogain Mondrian » et « LSD Mondrian », « Bottom up generalization », « Incognito » [7].

### 1.6.1 Principe des algorithmes de généralisation des données

La famille de généralisation altère les caractéristiques des individus dans un ensemble de données en transformant des données spécifiques en données plus générales. *Par exemple, plutôt que de spécifier la date de naissance précise d'une personne, on utilise son année de naissance.* Cela réduit la précision des données tout en maintenant leur utilité globale pour les analyses et les applications[7]. L'objectif principal de la généralisation est de rendre plus difficile l'identification individuelle tout en préservant la pertinence des données.

Quasi-identifiant			Attribut sensible
Sexe	Code postal	Niveau d'étude	Salaire
M	13050	5 <sup>ème</sup>	1200
F	113051	3 <sup>ème</sup>	1300
M	13050	Seconde	1200
M	13050	Seconde	1300
M	13051	1 <sup>ier</sup> et 2 <sup>ème</sup> cycle	1500
F	13050	1 <sup>ier</sup> et 2 <sup>ème</sup> cycle	1500
F	13061	1 <sup>ier</sup> et 2 <sup>ème</sup> cycle	1600
F	13061	Master	2000
F	13060	Master	2100
M	13061	Doctorat	3000
M	13060	Doctorat	4000
M	13061	Doctorat	4500

TABLE 1.7 – Table originale

La **Table 1.7** définit les quasi-identifiants : Sexe, Code postal, Niveau d'étude et des attributs sensible : Salaire.

L'objectif est de construire une hiérarchie de généralisation de chaque attribut quasi-identifiant pour établir une généralisation des données.

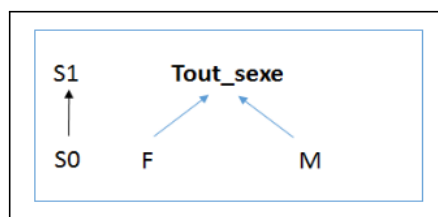


FIGURE 1.3 – Hiérarchie de généralisation de l'attribut Sexe

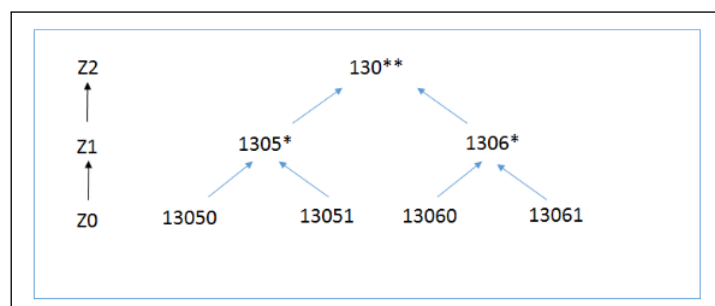


FIGURE 1.4 – Hiérarchie de généralisation de l'attribut Code postal



### 1.6.3 Algorithme Datafly

Datafly est un algorithme mis en œuvre pour assurer la protection des données médicales. Cet algorithme a été développé dans les années 1997-1998 par Latanya Arvette Sweeney dans le but de conserver le maximum d'informations utiles aux analyses de données. Datafly, contrairement à l'algorithme  $\mu$ -Argus, autorise des suppressions globales d'attributs quasi-identifiants. Au début du processus, l'algorithme fixe à  $k$  la valeur choisie par l'utilisateur, c'est-à-dire le nombre maximum de lignes qu'il a l'autorisation de supprimer dans la table [13]. Datafly procède par itération **Figure 1.6**. Il commence par calculer le nombre de valeurs distinctes de chaque attribut quasi-identifiant et sélectionne l'attribut qui a la plus grande valeur pour le généraliser. Après cette première étape, il reprend  $n$  fois le même processus jusqu'à trouver une généralisation optimale de la base de données. Ce processus itératif vise à éviter les grandes généralisations qui ne sont pas nécessaires afin de tenir compte de l'utilité des données. Pour plus de compréhension, voici un logigramme qui montre le processus de l'algorithme Datafly [7].

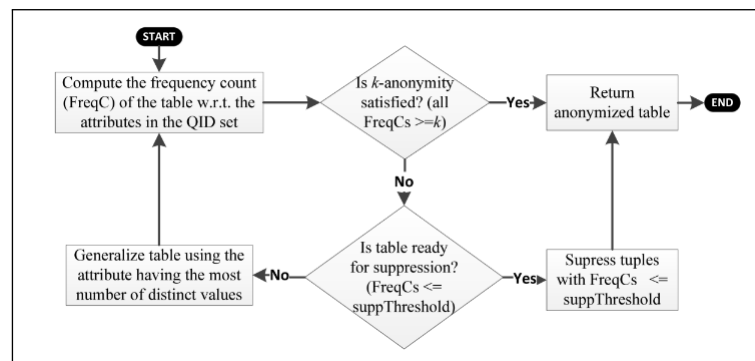


FIGURE 1.6 – Processus Centrale de L'algorithme Datafly

### 1.6.4 Algorithme de Samarati

L'algorithme de Samarati est fondé sur un treillis qui représente les combinaisons possibles des niveaux de généralisation de tous les attributs quasi-identifiants. Un treillis, en termes plus simples, est un réseau de nœuds où chaque nœud correspond à la mise en œuvre d'une hiérarchie de généralisation possible de la table originale[7]. Soit **S** pour représenter le sexe et **Z** pour le code postal.

Sexe	Code postal	Salaire
tout_sexe	1305*	1200
tout_sexe	1305*	1300
tout_sexe	1305*	1200
tout_sexe	1305*	1300
tout_sexe	1305*	1500
tout_sexe	1305*	1500
tout_sexe	1306*	1600
tout_sexe	1306*	2000
tout_sexe	1306*	2100
tout_sexe	1306*	3000
tout_sexe	1306*	4000
tout_sexe	1306*	4500

TABLE 1.8 – Donnée généralisée selon  $\langle S1, Z1 \rangle$ 

**Table 1.8** représente la généralisation des données quasi-identifiant : **sexe** et **code postal**. Ici, notre objectif est de construire un treillis de généralisation des deux attributs.

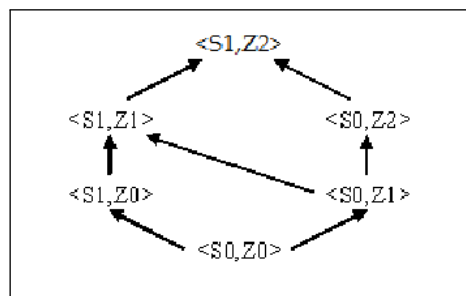


FIGURE 1.7 – Treillis de généralisation des attributs Sexe et Code postal  
de la **Table 1.8**

La **Figure 1.7** représente la construction d'un treillis de généralisation des attributs **code postal** et **sexe**. Samarati soutient que pour obtenir les meilleurs résultats en matière d'anonymisation, il est essentiel de trouver les nœuds qui décrivent une généralisation qui respecte le concept de k-anonymat, tout en limitant les suppressions afin de ne pas dépasser le seuil autorisé. De plus, afin de limiter la perte d'informations, il est préférable de choisir des nœuds qui se situent le

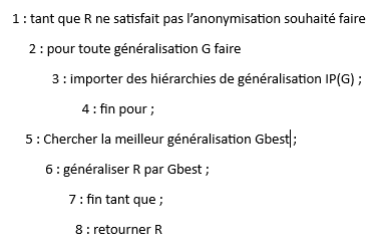
plus près possible du niveau le plus bas du treillis, correspondant à l'état initial (par exemple,  $\langle S_0, Z_0 \rangle$  dans notre cas).

Bien que monter dans le treillis soit nécessaire pour atteindre le niveau de k-anonymat requis, cela se fait au détriment de la précision des données. Par conséquent, trouver un équilibre optimal de généralisation est essentiel et peut être réalisé en choisissant des nœuds spécifiques dans le treillis.

Dans l'optique de trouver des nœuds optimaux, l'algorithme agit par itération et en considérant les nœuds de niveau  $h/2$ , où  $h$  représente la hauteur de la partie inexplorée du treillis[7].

### 1.6.5 Algorithme «Bottom up generalization»

Proposé par **Wang, Yu, et Chakraborty en 2004**, le but est d'assurer une protection des données pour un type spécifique de traitement statistique ; la classification[7]. Cette technique est une approche itérative du traitement des données pour généraliser les informations. L'avantage est qu'elle rend difficile l'établissement de liens entre sources même si les données généralisées restent utiles pour la classification. La généralisation ascendante transforme les données précises en données moins précises en gardant une sémantique cohérente.



```
graph TD; 1[1 : tant que R ne satisfait pas l'anonymisation souhaitée faire] --> 2[2 : pour toute généralisation G faire]; 2 --> 3[3 : importer des hiérarchies de généralisation IP(G) ;]; 3 --> 4[4 : fin pour ;]; 4 --> 5[5 : Chercher la meilleur généralisation Gbest]; 5 --> 6[6 : généraliser R par Gbest ;]; 6 --> 7[7 : fin tant que ;]; 7 --> 8[8 : retourner R];
```

1 : tant que R ne satisfait pas l'anonymisation souhaitée faire  
2 : pour toute généralisation G faire  
3 : importer des hiérarchies de généralisation IP(G) ;  
4 : fin pour ;  
5 : Chercher la meilleur généralisation Gbest ;  
6 : généraliser R par Gbest ;  
7 : fin tant que ;  
8 : retourner R

**FIGURE 1.8 – Algorithme bottom-up generalisation**

La **Figure 1.8** représente un aperçu de l'algorithme **bottom-up generalisation**, [7] contient une explication plus détaillée sur cette méthode.

### 1.6.6 Algorithme Incognito

Proposé par **LeFevre, DeWitt, et Ramakrishnan en 2005**, incognito repose également sur l'utilisation d'un treillis, mais son approche est itérative et progressive pour une efficacité accrue. Au cours de la première itération, Incognito élabore tous les treillis associés à un attribut de l'information sensible (QI). Chaque treillis représente une hiérarchie de généralisation. Ces treillis distincts sont ensuite raffinés en éliminant tous les nœuds qui ne permettent pas d'atteindre une généralisation  $k$ -anonyme.

Dans la deuxième itération, Incognito fusionne les treillis résultant de l'étape précédente pour former des treillis à deux attributs, et procède comme précédemment pour les nettoyer. Ce processus itératif se poursuit jusqu'à ce que le treillis englobant tous les attributs du QI soit construit et épuré.

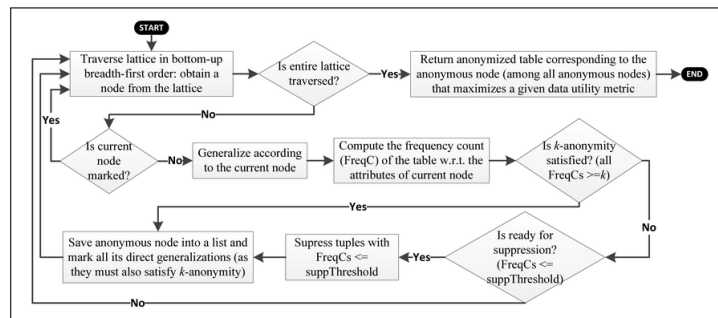


FIGURE 1.9 – Algorithme incognito

Il convient également de noter que, bien qu'étant plus rapide que Samarati, Incognito présente une complexité exponentielle, que ce soit en termes de temps d'exécution ou d'espace mémoire, par rapport à la taille des données[14].

### 1.6.7 Algorithme Median Mondrian

L'algorithme Median Mondrian vise à anonymiser les données en divisant les enregistrements de la table originale en groupes où chaque groupe contient au moins ' $k$ ' individus ayant la même valeur de quasi-identifiant (QI). Pour atteindre cet anonymat, l'algorithme opère dans un espace tridimensionnel où chaque dimension correspond à un attribut du QI.



Initialement, les enregistrements sont positionnés dans l'espace tridimensionnel correspondant en fonction de leurs valeurs d'attributs du QI. L'algorithme divise l'espace en zones en utilisant la médiane des valeurs d'une dimension spécifique. À chaque itération, il choisit une dimension et vérifie si la division d'une zone en deux sous-groupes respecte la contrainte de 'k-anonymat', c'est-à-dire que chaque sous-groupe contient au moins 'k' individus.

Les zones qui ne peuvent pas être divisées sans violer le 'k-anonymat' sont marquées. L'algorithme passe ensuite à une autre dimension et répète le processus jusqu'à ce que toutes les dimensions aient été explorées. Une fois toutes les zones marquées, il applique les généralisations nécessaires en remplaçant les différentes valeurs d'une même zone par la valeur de leur premier parent commun dans un processus appelé recodage.

Median Mondrian divise l'espace tridimensionnel des données en zones respectant le 'k-anonymat', puis recode les valeurs des zones pour assurer l'anonymisation tout en préservant l'utilité des données pour l'analyse ou la recherche.

### 1.6.8 Algorithmes « InfoGain Mondrian » et « LSD Mondrian »

Ces deux algorithmes sont des extensions de l'algorithme Median Mondrian proposés par **LeFevre, DeWitt, et Ramakrishnan en 2008** et conçus pour préserver soit la classification dans le cas de l'InfoGain Mondrian, soit la régression dans le cas du LSD Mondrian. Pour cela, ils intègrent le recodage multidimensionnel de Median Mondrian avec des heuristiques de partitionnement spécifiquement orientées vers la classification pour InfoGain Mondrian et vers la régression pour LSD Mondrian.

Intuitivement, il s'agit dans Infogain, de choisir, à chaque itération, le partitionnement qui minimise l'entropie pondérée de l'ensemble des partitions résultantes tout en préservant la contrainte d'anonymat. L'utilisation de cette métrique, selon les auteurs de cet algorithme, favorise l'obtention de partitions homogènes[6]. La formule de calcul de l'entropie pondérée est la suivante :

$$Entropy(P, C) = \sum_{\text{partitions } P'} \frac{|P'|}{|P|} \sum_{c \in D_c} -p(c|P') \log p(c|P')$$

P est la partition courante, P' est l'ensemble des partitions résultantes pour les divisions candidates,  $p(c|P')$  est le pourcentage des enregistrements labellisés avec l'étiquette de la classe c. LSD Mondrian, quant à lui, s'inspire de l'algorithme CART de construction d'un arbre de régression. Par conséquent, à chaque itération, il choisit la division qui minimise la somme pondérée de MSE (Mean Squared Error) pour l'ensemble des partitions résultantes.

$$MSE(P') = \frac{1}{|P'|} \sum_{i \in P'} (r_i - \bar{r}(P'))^2$$

|P| est constante pour toutes les divisions candidates, l'algorithme choisit la division qui minimise l'expression suivante :

$$Error^2(P, R) = \sum_{\text{partitions } P'} \sum_{i \in P'} (r_i - \bar{r}(P'))^2$$

$$PondrMSE = \sum_{\text{partitions } P'} \frac{|P'|}{|P|} (MSE(P')) = \frac{1}{|P|} \sum_{i \in P'} (r_i - \bar{r}(P'))^2$$

Où P est la partition courante, P' est l'ensemble des partitions résultantes pour les divisions candidates, i est un enregistrement,  $r_i$  est la valeur de l'attribut cible R de l'enregistrement i,  $\bar{r}(P')$  est la moyenne des valeurs de l'attribut cible des enregistrements appartenant à P' [7].

## 1.7 ALGORITHMES DE RANDOMISATIONS

### 1.7.1 Ajout de bruit

La technique d'ajout de bruit est particulièrement utile lorsque les attributs peuvent avoir un effet négatif significatif sur les individus. Elle consiste à modifier les attributs de l'ensemble de données de manière à réduire leur précision tout en préservant la distribution globale. Lorsqu'un ensemble de données est traité, un observateur peut supposer que les valeurs sont exactes, mais en réalité, elles peuvent être légèrement altérées. Par exemple, si la taille d'un individu était initialement mesurée avec précision au centimètre près, après l'anonymisation, les données pourraient indiquer une taille précise dans une marge de  $\pm 10$  cm seulement. Si cette technique

est appliquée de manière efficace, elle rendra difficile pour un tiers d'identifier une personne, de restaurer les données ou de détecter les modifications apportées[4]. Lorsque l'ajout de bruit est bien appliqué, elle garantit une [15] :

- forte confidentialité des individus,
- bonne performance,
- conservation de la distribution générale

### 1.7.2 Permutation

La technique de **permutation** consiste à modifier les valeurs des attributs d'un jeu de données tabulaire, afin que certaines des valeurs, soient artificiellement liées aux différentes personnes concernées.

**Principe :** *Consiste à mélanger les valeurs des attributs dans un tableau de telle sorte que certaines d'entre elles soient artificiellement liées à des personnes concernées différentes. La permutation altère donc les valeurs au sein de l'ensemble de données en les échangeant simplement d'un enregistrement à un autre[15].*

Elle est indispensable lorsqu'il est nécessaire de garder la distribution exacte de chaque attribut au sein de l'ensemble de données[15].

Cette méthode applique une modification des valeurs du jeu de données en faisant une permutation d'une donnée à l'autre. Elle conserve, l'étendue et la distribution des valeurs mais ne garde pas les liens réciproques entre les valeurs et les individus concernés. Dès lors que deux attributs ont une relation logique ou un lien statistique et sont indépendamment permutés, cette dernière sera détruite. En foi de quoi, il est parfois nécessaire de conserver les liens logiques lors de la permutation des valeurs des attributs possédant des relations entre eux. Sinon le pirate en quête d'une ré-identification des individus pourrait, récupérer les valeurs des attributs permutés et d'inverser la permutation. Par exemple, *si nous considérons un sous-ensemble d'attributs dans un ensemble de données médicales tel que "raisons de l'hospitalisation/symptômes/département responsable", une forte relation logique liera les valeurs dans la plupart des cas et la permutation d'une seule des valeurs sera ainsi détectée et pourra même être inversée"[4].*

### 1.7.3 Confidentialité différentielle

La **confidentialité différentielle** ou **Differential Privacy** en anglais, a pour vocation de produire un ensemble de données anonymes tout en gardant une copie des données originales. L'ensemble de données anonymes est obtenu suite à une requête faite par un tiers sur la base de données et dont le résultat sera associé à un ajout de bruit. Pour être considéré « differentially private », la présence ou l'absence d'un individu particulier dans la requête ne doit pas pouvoir changer son résultat[15]. La protection différentielle de la vie privée indique au responsable du traitement des données le nombre de bruit qu'il doit ajouter, et sous quelle forme, pour obtenir les garanties nécessaires en matière de protection de la vie privée". Cette technique ne [4] :

1. modifie pas directement les données car il s'agit d'un ajout de bruit à posteriori et relatif à une requête. Les données originales sont donc toujours présentes. À ce titre, les résultats peuvent aussi être considérés comme des données à caractère personnel,
2. permet pas de partager le jeu de données dans sa structure initiale, limitant ainsi le panel d'analyse réalisables.

## 1.8 QUELQUES OUTILS D'ANONYMISATION

### 1.8.1 Outil $\mu$ -Argus

$\mu$ -Argus est un logiciel open source développé pour anonymiser des données sensibles, principalement utilisé dans la recherche scientifique et statistique ainsi que dans d'autres contextes nécessitant la confidentialité des données. Il a été créé par l'Agence Nationale des Statistiques des Pays-Bas dans le cadre du projet européen CASC (Computational Aspects of Statistical Confidentiality). Ce logiciel présente plusieurs avantages significatifs.

1. **méthodes d'anonymisation** :  $\mu$ -argus propose diverses méthodes d'anonymisation pour protéger la confidentialité des données telles que la pseudonymisation, la suppression de données, la généralisation, la substitution aléatoire, et d'autres techniques visant à rendre les données moins identifiables tout en conservant leur utilité pour l'analyse,

2. **prise en charge de différents types de données** : le logiciel prend en charge divers types de données tels que les données tabulaires, spatiales et temporelles. Il supporte différents formats de données, ce qui en fait un outil polyvalent pour l'anonymisation dans différents domaines,
3. **configuration flexible** : l'outil offre une configuration flexible qui permet aux utilisateurs de définir des règles d'anonymisation adaptées à leurs besoins spécifiques. Cela inclut la possibilité de créer des stratégies d'anonymisation personnalisées en fonction des caractéristiques des données et des exigences de confidentialité,
4. **documentation et support communautaire** :  $\mu$ -argus est accompagné d'une documentation qui détaille son utilisation.

L'outil  $\mu$ -argus est largement utilisé dans le domaine de la recherche scientifique et statistique, ainsi que dans d'autres domaines où la confidentialité des données est une priorité.

### 1.8.2 ARX Data Anonymization Tool

**ARX** est un logiciel open source conçu pour anonymiser une base de données originale contenant des données personnelles sensibles. Il repose sur un algorithme unique appelé "flash", qui lui permet de construire un treillis de généralisation, suivant le même principe que les algorithmes Incognito ou Samarati. ARX prend en charge l'implémentation d'une grande variété de :

- modèles de confidentialité et de risque,
- méthodes de transformation des données,
- méthodes d'analyse de l'utilité des données de sortie.

Le logiciel a déjà servi dans divers contextes à savoir dans des projets de recherche, le partage de données d'essais cliniques et à des fins de formation. Voici un schéma qui montre le processus d'anonymisation du logiciel arx :

La **Figure 1.10** montre que arx met en œuvre un processus constitué de trois phases précédées de **la phase d'importation** de données originales et suivies de **l'exportation** de données anonymes[7].

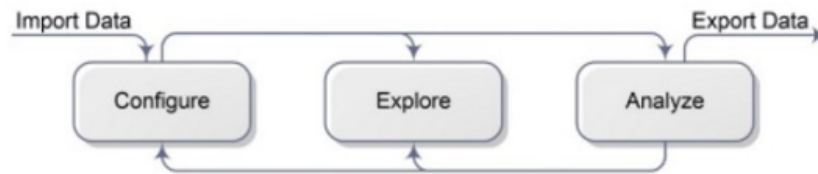


FIGURE 1.10 – Processus d’anonymisation de ARX

1. **phase configuration** : cette phase permet de spécifier plusieurs paramètres qui vont permettre par la suite d’appliquer l’anonymisation à une base de données importer dans le logiciel,
2. **phase d’exploration** : arx propose une fonctionnalité permettant d’afficher le treillis de généralisation construit par l’algorithme 'flash'. Chaque nœud de ce treillis représente une anonymisation potentielle et est coloré pour indiquer sa signification. les nœuds sont également triés en fonction de la perte d’information associée, ce qui aide l’utilisateur à choisir parmi les solutions potentielles à analyser plus en détail par la suite,
3. **phase d’analyse** : elle permet à l’utilisateur d’évaluer un nœud choisi dans la phase précédente en termes de sécurité et d’utilité. Cette analyse permet de déterminer les propriétés de la table anonymisée, telles que le nombre de classes d’équivalence et la valeur de la métrique d’utilité. De plus, elle permet de comparer les données originales aux données anonymisées en examinant les fréquences de chaque valeur de chaque attribut, la distribution des attributs et les tables de contingence. La phase d’analyse permet également d’évaluer le risque associé à chaque enregistrement, chaque attribut quasi-identifiant et à l’ensemble des données.

Le manuel d’utilisation d’arx se trouve sur son site officiel [16]. Il décrit toutes les fonctionnalités de l’outil à l’aide de textes et de vidéos.

## 1.9 SYNTHÈSE

modèle de protection de la vie privée	Liaisons d'enregistrement	Liaison d'attribut			Liaison de table
		Attaques d'homogénéité	Attaque de reconnaissance acquise	Attaque de l'inférence probabiliste	
k-anonymat	*	*			
l-diversité	*	*	*		
(l,c)-diversité	*	*	*	*	
l-diversité d'entropie	*	*	*	*	
t-proximité		*			
$\delta$ -présence					*

**TABLE 1.9 – Résumé sur les modèles de protection de la vie privée**

La **Table 1.9** est un résumé portant sur les modèles de protection de la vie privée, ce tableau laisse voir qu'aucun modèle de protection ne permet de contrer toutes attaques liées à la vie privée.

Il existe de nombreux états de l'art sur les techniques d'anonymisation, qui peuvent provenir de différentes communautés telles que SDC (Statistical Disclosure Control), PPDM (Privacy Preserving Data Mining) ou PPDP (Privacy Preserving Data Publishing) Nous avons choisi de présenter brièvement quelques techniques d'anonymisation de micro-données (donnée tabulaire), sachant qu'il existe plusieurs variantes et algorithmes pour chaque technique. Certaines techniques altèrent les données résultantes, tandis que d'autres les préservent. Par exemple, la généralisation ne modifie pas fondamentalement les données, ce qui permet de les utiliser à des fins de tests ou de statistiques.

De plus, ces techniques sont généralement applicables uniquement à certains types d'attributs, tels que les attributs continus ou catégoriels, les attributs sensibles ou ceux faisant partie du quasi-identifiant.

## CONCLUSION

Dans ce chapitre, nous avons fourni une vue d'ensemble de l'anonymisation des données tabulaires, en mettant l'accent sur leur publication. Nous avons souligné que le choix des

techniques d'anonymisation est influencé par le modèle de protection de la vie privée adopté et par le domaine d'application spécifique.

Pour notre projet, nous avons choisi de nous concentrer sur la famille de généralisation en raison de sa popularité, de sa capacité à protéger les données et de son aptitude à maintenir l'utilité des données anonymisées. La généralisation réduit la précision des données pour prévenir la ré-identification tout en conservant leur valeur pour l'analyse.

Nous avons également abordé la famille de randomisation, qui offre une forte protection de la vie privée en altérant les données de manière significative. Cependant, cette méthode peut compromettre l'utilité des données anonymisées en faussant les résultats des analyses.



---

## CONCEPTION DU SYSTÈME LOGICIEL

### Sommaire

---

<b>INTRODUCTION . . . . .</b>	<b>33</b>
<b>2.1 ANALYSE DES BESOINS . . . . .</b>	<b>33</b>
2.1.1 Besoins fonctionnels . . . . .	33
2.1.2 Besoins non fonctionnels . . . . .	34
<b>2.2 ARCHITECTURE DU SYSTÈME LOGICIEL . . . . .</b>	<b>34</b>
2.2.1 Diagramme de cas d'utilisation . . . . .	35
2.2.2 Diagramme de classe . . . . .	35
<b>CONCLUSION . . . . .</b>	<b>36</b>

---

## INTRODUCTION

Ce présent chapitre porte sur la conception architecturale de notre système logiciel baptisé **PFCL\_Anonimization**. La phase de conception est cruciale dans le développement de notre système logiciel d'anonymisation de données. Elle nous permettra de transformer les besoins des utilisateurs en une solution technique concrète.

Notre objectif principal est de construire une base solide pour assurer une bonne implémentation de notre système d'anonymisation.

Nous commencerons par analyser les besoins des utilisateurs et finirons par une description de l'architecture du système.

## 2.1 ANALYSE DES BESOINS

L'analyse des besoins consiste à identifier les attentes des utilisateurs en termes de besoins fonctionnels et non fonctionnels.

### 2.1.1 Besoins fonctionnels

Les besoins fonctionnels concernent les fonctionnalités que le système doit offrir.

Les obligations fonctionnelles que notre logiciel doit implémenter sont :

- ☛ **l'importation des données** : le système doit permettre l'importation d'une base de données tabulaires aux formats de fichier Excel (xlsx ou xls),
- ☛ **l'anonymisation des données** : l'outil doit fournir l'anonymisation de base de données tabulaire,
- ☛ **la sélection automatique des catégories de données** : l'utilisation du logiciel ne doit pas nécessiter des connaissances dans le domaine de la protection des données personnelles,
- ☛ **la possibilité de configuration** : le système doit permettre à l'utilisateur de choisir les catégories de données de la base de données,

- ☛ **protection de la qualité des données anonymisées** : le système ne doit pas faire des généralisations excessives qui détruiraient la qualité des données,
- ☛ **exportation des données anonymisées** : le système doit permettre l'exportation des données anonymisées dans les mêmes formats que ceux utilisés pour l'importation.

### 2.1.2 Besoins non fonctionnels

Les besoins non fonctionnels sont les exigences des utilisateurs en terme de design et de convivialité du logiciel.

Ces obligations pour notre outil sont :

- ☛ la **maintenabilité** : la maintenabilité d'un logiciel désigne la facilité avec laquelle un logiciel peut être modifié après son déploiement initial. Les besoins étant en constant évolution, notre système doit pouvoir évoluer pour l'intégration de nouvelles fonctionnalités,
- ☛ l'**accessibilité** : l'accessibilité à un logiciel désigne la capacité du logiciel à être utilisé efficacement et confortablement par tous les utilisateurs. Dans le but de satisfaire les utilisateurs, notre système logiciel doit être facile d'accès avec une interface utilisateur intuitive et compréhensible.

## 2.2 ARCHITECTURE DU SYSTÈME LOGICIEL

L'architecture d'une application est le squelette de celle-ci. La conception d'une architecture logicielle est la base de tout développement. Pour notre outil, une architecture bien pensée est cruciale pour s'assurer d'avoir un système évolutif.

Notre architecture de développement repose sur :

1. **la modularité** : c'est la structuration du logiciel en modules ou composants distincts qui peuvent être modifiés indépendamment les uns des autres. Elle permet une implémentation structurée et garantit une maintenabilité (adaptabilité) du système,
2. l'**interopérabilité** : c'est l'aptitude d'un système à fonctionner avec d'autres applications et services. L'utilisation d'APIs standardisées permettra une communication fluide entre

notre outil et les autres systèmes favorisant une meilleure collaboration et un échange de données efficace.

### 2.2.1 Diagramme de cas d'utilisation

Le diagramme de cas d'utilisation est crucial pour comprendre le fonctionnement de notre outil d'anonymisation. Il permet de définir clairement les exigences fonctionnelles en illustrant les interactions entre les utilisateurs (acteurs) et le système. Il identifie les rôles et responsabilités des acteurs, aidant à gérer la portée du projet et à prioriser le développement. Il sert également de documentation précieuse et de base pour la conception détaillée, garantissant que toutes les spécifications techniques sont dérivées de scénarios d'utilisation bien définis.

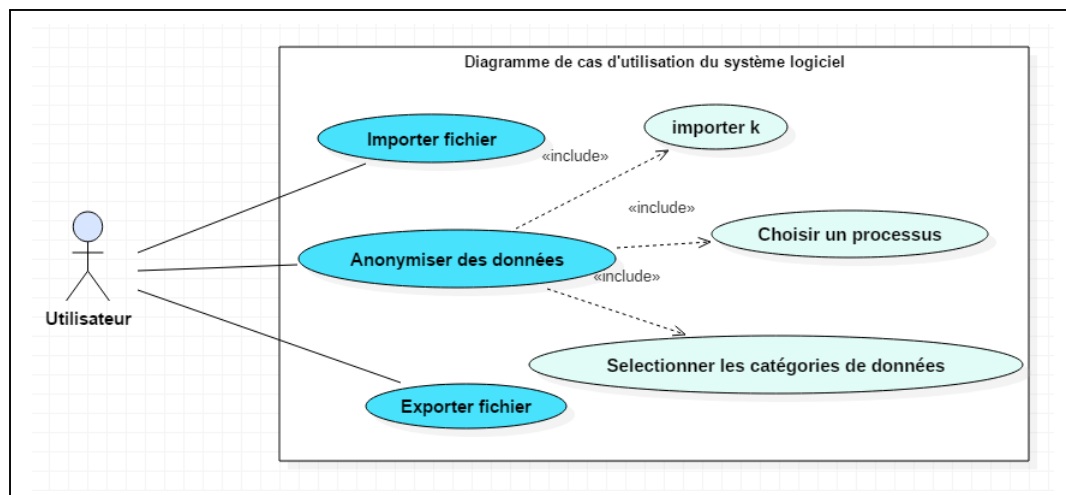


FIGURE 2.1 – Diagramme de cas d'utilisation

La **Figure 2.1** représente le diagramme de cas d'utilisation de notre système logiciel, il décrit les interactions entre le système et l'utilisateur.

### 2.2.2 Diagramme de classe

Un diagramme de classe est un élément clé de la modélisation d'un système logiciel dans le cadre de la méthode de développement orienté objet. Utilisé couramment dans le cadre du langage de modélisation UML (Unified Modeling Language), un diagramme de classe décrit la structure statique d'un système en montrant ses classes, leurs attributs, leurs méthodes et les

relations qui les unissent. Dans le but de bien organiser notre implémentation, un diagramme de classe bien détaillé sera un atout majeur.

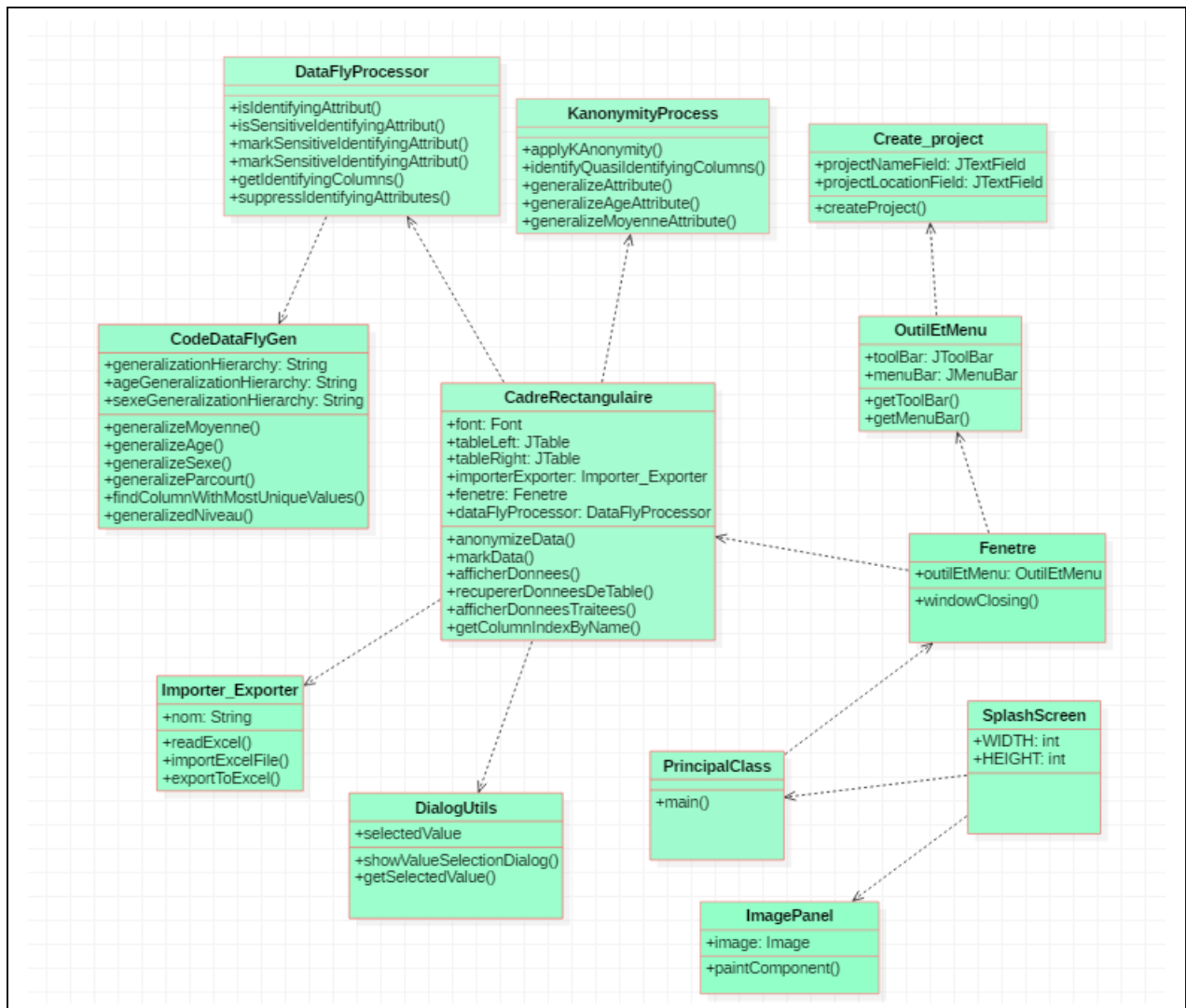


FIGURE 2.2 – Diagramme de classe

La Figure 2.2 illustre la modélisation du diagramme de classe de l’outil que nous devons concevoir. Dans le chapitre sur l’implémentation, il sera question d’implémenter les tables présentées et de les faire interagir.

## CONCLUSION

Ce chapitre a présenté les fondements de la conception de notre outil d’anonymisation, **PFCL\_Anonymization**. En commençant par une analyse détaillée des besoins, nous avons

identifié les exigences fonctionnelles et non fonctionnelles essentielles pour garantir l'efficacité et la convivialité de l'outil. Cette analyse a permis de définir clairement les attentes des utilisateurs et les objectifs à atteindre.

Ensuite, nous avons détaillé l'architecture du système logiciel, illustrée par un diagramme de classe et un diagramme de cas d'utilisation. Le diagramme de classe a permis de structurer les différentes composantes du système et leurs interactions, offrant une vue d'ensemble claire et cohérente de l'architecture logicielle. Le diagramme de cas d'utilisation, quant à lui, a mis en lumière les interactions entre les utilisateurs et le système, en soulignant les fonctionnalités principales et les scénarios d'utilisation.

Cette conception détaillée établit une base solide pour le développement de notre outil, en garantissant que toutes les exigences sont prises en compte et que l'architecture est optimisée pour répondre aux besoins identifiés. Les diagrammes offrent une visualisation claire et structurée du système, facilitant ainsi la transition vers la phase d'implémentation.

---

## IMPLÉMENTATION DU SYSTÈME LOGICIEL

### Sommaire

---

<b>INTRODUCTION . . . . .</b>	<b>39</b>
<b>3.1 OUTILS UTILISES . . . . .</b>	<b>39</b>
3.1.1 Technologies utilisées . . . . .	39
3.1.2 IDE de développement : Java Eclipse . . . . .	40
<b>3.2 OBJECTIF DE L'OUTIL . . . . .</b>	<b>43</b>
<b>3.3 POURQUOI L'ALGORITHME DATAFLY ET LE MODÈLE     K-ANONYMAT ? . . . . .</b>	<b>44</b>
3.3.1 Algorithme DataFly . . . . .	44
3.3.2 Modèle k-anonymat . . . . .	45
<b>3.4 ORGANISATION DE L'OUTIL . . . . .</b>	<b>46</b>
<b>3.5 UTILISATION DE L'OUTIL : CAS PRATIQUE . . . . .</b>	<b>47</b>
3.5.1 Dataset étudiant . . . . .	48
3.5.2 Importer une base de données . . . . .	49
3.5.3 Choix du paramètre d'anonymisation . . . . .	53
3.5.4 Construction de la hiérarchie de généralisation . . . . .	53
3.5.5 Comparaison des résultats . . . . .	54
<b>CONCLUSION . . . . .</b>	<b>55</b>

---

## INTRODUCTION

L'implémentation de notre logiciel consiste à transformer les tables et les liens entre les tables du diagramme de classe en un programme exécutable. Ce programme doit offrir les fonctionnalités décrites par le diagramme de cas d'utilisation.

Dans ce chapitre, nous décrirons l'implémentation de notre outil d'anonymisation de donnée : **PFCL\_Anonymization**.

Notre objectif principal est de développer un outil robuste, convivial et facile d'accès capable d'anonymiser une base de données contenant des informations sur des étudiants, en garantissant à la fois l'anonymat des étudiants concernés et l'utilité des données anonymisées.

Nous débuterons par une présentation des outils de développement utilisés. Ensuite, nous définirons les objectifs spécifiques de notre outil d'anonymisation. Nous poursuivrons par une justification des méthodes utilisées. Enfin, nous décrirons en détail l'outil implémenté.

## 3.1 OUTILS UTILISES

### 3.1.1 Technologies utilisées

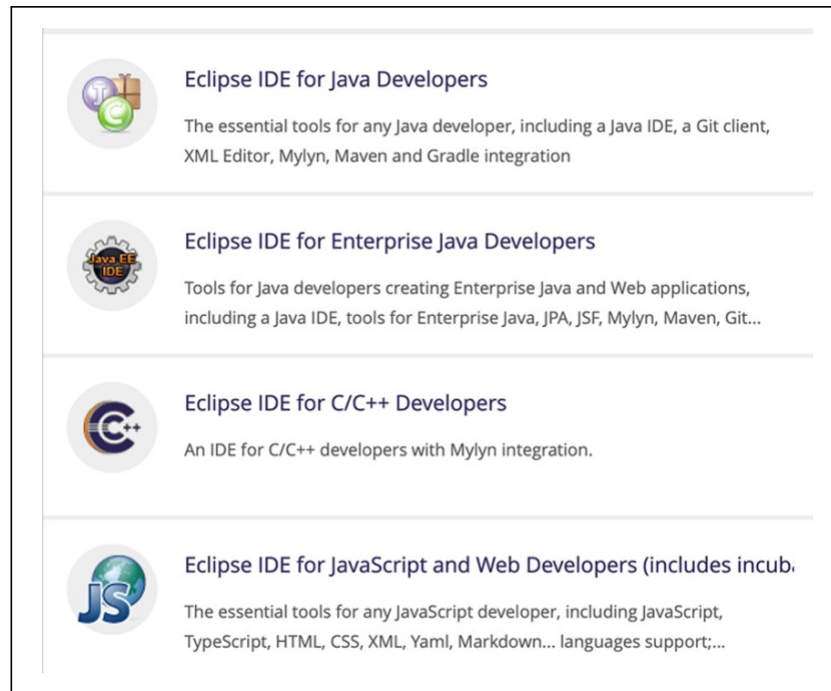
- ☛ **le matériel utilisé** : pour le développement de notre système logiciel, nous utilisons un ordinateur portable utilisant Windows 10 professionnel 64 bits comme système d'exploitation, une mémoire RAM de 4,00Go et de processeur Intel® Core™ i5-2520M CPU @ 2,50GHz, 2501MHz, 2 cœurs et 4 processeur(s) logique(s),
- ☛ **le langage de programmation** : comme langage de programmation, nous avons fait le choix du langage Java car il est multi-plateforme (c'est-à-dire, le code java une fois édité est utilisable sur divers système d'exploitation tel que Windows, UNIX, GNU/Linux, Mac OS, avec peu ou aucune modification), facile à maintenir, orienté objet et propose un ensemble de librairie qui permet de ne pas tout coder à la main[17],



☛ **les librairies utilisées** : pour développer notre système logiciel, nous utilisons deux librairies offertes par Java. Premièrement nous utilisons la **librairie Swing** qui nous permet le développement d'interface utilisateur indépendant du système d'exploitation ce qui est le contraire pour AWT qui demande au système d'exploitation de fournir les composants graphiques. Nous n'avons pas utilisé AWT par ce que nous voulons que notre interface utilisateur soit construit indépendamment du système d'exploitation (conserver la même interface utilisateur peu importe le système utilisé). Swing permet de demander à java de dessiner lui-même les composants graphiques nécessaires à l'interface utilisateur tel que demander par le développeur[18]. Nous utilisons la bibliothèque Apache POI, essentielle pour manipuler les documents Microsoft Office (Excel, Word, PowerPoint). POI permet de lire, écrire et modifier des fichiers Office dans divers formats, ce qui est crucial pour notre projet. Elle s'intègre parfaitement avec notre stack Java et est compatible avec les principales IDEs, facilitant son adoption. POI prend en charge les formats modernes (.xlsx, .docx) et anciens (.xls, .doc), assurant une rétrocompatibilité.[19].

### 3.1.2 IDE de développement : Java Eclipse

Un environnement de développement intégré (IDE) est une application logicielle qui aide les programmeurs à développer efficacement le code logiciel. Eclipse Java est un IDE open source et personnalisable. Il est ouvert pour le langage de programmation Java, distribué et maintenu par la Fondation Eclipse [20]. Nous avons opté pour cette plate-forme car elle offre une multitude de plugins en fonction du langage de programmation. La communauté **Eclipse Foundation** est très active sur le Web, ce qui facilite la recherche de solutions et le partage d'expériences.



**FIGURE 3.1 – IDE de développement Eclipse**

La **Figure 3.1** présente les différents IDE de développement proposés par la fondation Eclipse.

Éclipse possède un interface convivial, facile à maîtriser et utilise l'API Swing qui est facile d'accès pour le développement d'application à interface graphique multi-plateforme. Eclipse permet de suivre la rédaction de votre code source en vous proposant de corriger directement les erreurs syntaxiques.

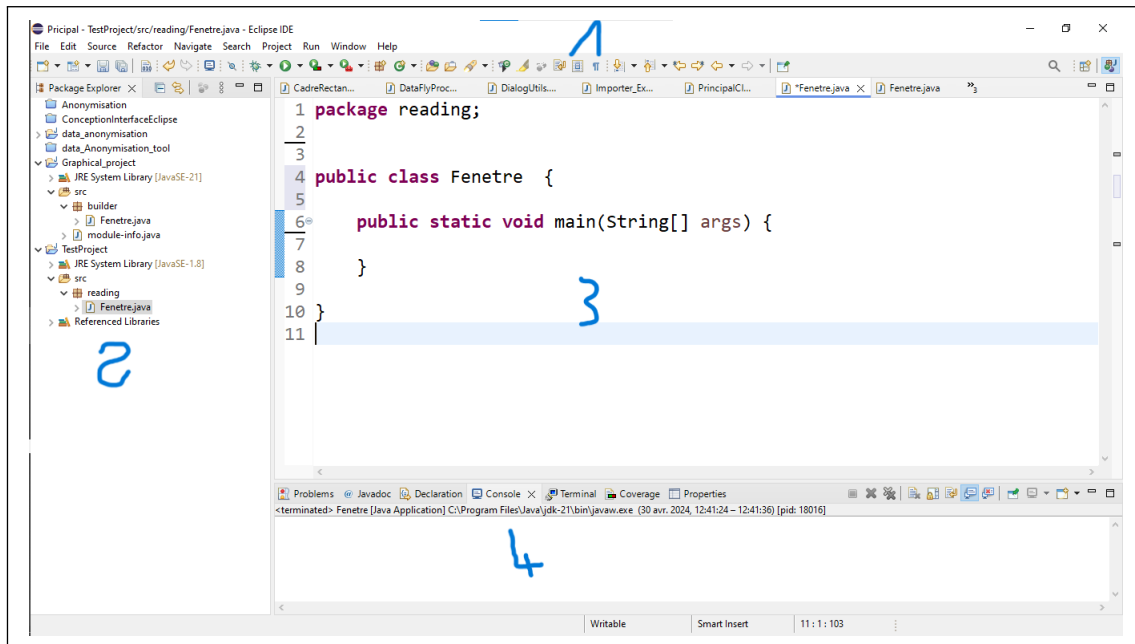


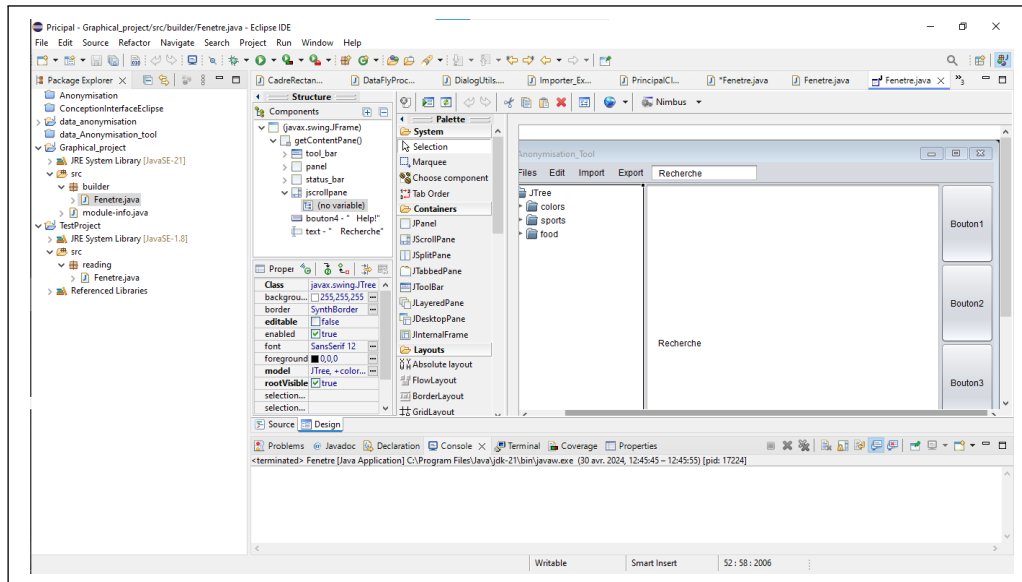
FIGURE 3.2 – Interface graphique d'éclipse

La **Figure 3.2** montre les différentes parties de l'interface graphique de l'environnement de développement Éclipse qui sont marquées par des numéros. Le numéro :

1. représente les composants **JToolBar** et **JMenuBar** en français on parle de **Barre d'outils** et de **Menu**. Elles contribuent de manière significative à l'expérience utilisateur, à l'efficacité de l'utilisation et à la convivialité globale du logiciel,
2. présente la structure de votre code : lorsque vous développez avec éclipse IDE, vous avez la possibilité de voir directement l'organisation de votre code c'est à dire la disposition des packages (répertoire pour organiser les classes du programme), des classes, des librairies utilisées,
3. montre votre espace de développement : c'est l'espace réservé pour l'édition du code de votre application,
4. désigne la **console** : la console représente une interface textuelle permettant aux développeurs d'interagir avec le système ou l'application en cours de développement. Il s'agit de l'endroit pour afficher les erreurs liées à votre application après compilation.

Sur l'interface, chaque composant est bien positionner pour assurer la convivialité et faciliter la recherche des outils de paramétrage ou de débogage lors du codage[17]. Un peu Similaire à **NetBeans**, l'IDE Éclipse offre des composants graphiques directement utilisables ce qui

facilite la tâche des développeurs de logiciel desktop (logiciel de bureau). La **Figure 3.3** est une illustration mettant en œuvre le développement d'une **JFrame** (fenêtre ou interface graphique d'une application).



**FIGURE 3.3 – Maquette de fenêtre sous Éclipse**

## 3.2 OBJECTIF DE L'OUTIL

Notre objectif principal est de développer un outil capable de transformer une base de données contenant des informations précises sur les étudiants en une version anonymisée, tout en préservant l'utilité des données. Pour atteindre cet objectif, nous avons mis en œuvre plusieurs exigences et principes clés à savoir :

1. **protection des données personnelles** : assurer que toutes les données sensibles des étudiants soient efficacement anonymisées pour garantir leur confidentialité,
2. **qualité de l'anonymisation** : assurer une bonne anonymisation des données en utilisant le principe du k-anonymat. Cela permet de réduire les risques de ré-identification tout en préservant la richesse des informations anonymisées,
3. **utilité des données** : maintenir la pertinence et l'intégrité des données anonymisées pour qu'elles restent utiles à des fins d'analyse et de recherche. L'outil doit minimiser la perte

d'informations tout en garantissant un haut niveau de confidentialité, permettant ainsi aux données anonymisées de rester utiles pour des analyses et des études,

4. **Automatisation** : mettre en place des fonctionnalités d'identification automatique des types d'attributs et proposer des hiérarchies de généralisation appropriée. Le logiciel doit détecter automatiquement les types d'attributs dans la base de données importée et propose une hiérarchie de généralisation appropriée, simplifiant ainsi le processus pour l'utilisateur,
5. **Accessibilité** : développer une interface utilisateur intuitive qui permet aux utilisateurs de tous niveaux, y compris ceux sans expertise technique, de facilement anonymiser des bases de données,
6. **Robustesse et Efficacité** : utiliser l'algorithme DataFly et le modèle k-anonymat pour offrir un équilibre optimal entre anonymisation et conservation des informations,
7. **Flexibilité de Format** : permettre l'importation et l'exportation des bases de données aux formats couramment utilisés, tels que Excel (xlsx et xls),
8. **efficacité pour les professionnels** : permettre de choisir les catégories d'attributs selon les besoins spécifiques. Il s'agit de mettre en place une fonctionnalité qui permet à l'utilisateur de renseigner lui même les types d'attributs.

### 3.3 POURQUOI L'ALGORITHME DATAFLY ET LE MODÈLE K-ANONYMAT ?

#### 3.3.1 Algorithme DataFly

L'algorithme datafly est un ensemble d'instruction permettant d'élaborer un système d'anonymisation de données. Nous l'avons choisi pour garantir une anonymisation efficace des données. Il est largement utilisé dans le domaine médical pour assurer l'anonymat en ajustant les niveaux de généralisation et de suppression. Sa réputation dans la protection des données sensibles témoigne de sa robustesse et de son efficacité en matière de protection de la vie privée. Les raisons sont les suivantes :

- datafly permet une rapidité et une efficacité dans le traitement des données, ce qui est crucial pour anonymiser de grandes bases de données étudiantes en un temps raisonnable,
- Il permet également l'utilisation du modèle de k-anonymat, qui est un standard éprouvé pour la protection des données,
- en utilisant des techniques de généralisation, datafly équilibre efficacement la confidentialité des données et leur utilité, ce qui est essentiel pour maintenir la pertinence des données anonymisées pour les utilisateurs finaux,
- datafly peut être adapté pour répondre à des besoins spécifiques et pour inclure des améliorations futures, assurant ainsi la pérennité et l'évolution de notre outil.

L'algorithme datafly est un choix optimal pour notre outil d'anonymisation car il allie efficacité, simplicité d'utilisation, flexibilité et robustesse dans la protection des données tout en maintenant leur utilité.

### 3.3.2 Modèle k-anonymat

Le modèle k-anonymat est un modèle de protection des données personnelles. Ce modèle permet d'établir un équilibre entre l'anonymat et l'utilité des données anonymisées, protégeant ainsi la vie privée des individus tout en maintenant la véracité des données. D'après notre analyse de l'état de l'art, le modèle k-anonymat est le plus recommandé pour atteindre cet équilibre, car il s'assure que chaque enregistrement est indistinguishable d'au moins k-1 autres enregistrements, réduisant ainsi le risque de ré-identification. En plus, le k-anonymat :

- peut être appliqué à diverses bases de données et peut être adapté en ajustant la valeur de k selon les besoins spécifiques de confidentialité et d'utilité des données,
- est un modèle robuste qui a été largement étudié et utilisé dans le domaine de la protection des données. Sa robustesse repose sur sa capacité à protéger contre diverses attaques de ré-identification,
- peut être combiné avec d'autres techniques de protection des données, comme l-l-diversité et t-closeness, pour améliorer encore plus la confidentialité des données anonymisées,

- est largement adopté dans divers domaines, ce qui en fait un choix éprouvé pour les applications nécessitant une protection des données,
- permet de maintenir une balance entre la protection de la confidentialité des données et la préservation de leur utilité. En généralisant ou en supprimant les données de manière contrôlée, il permet aux bases de données de rester utilisables pour des analyses statistiques et des recherches.

### 3.4 ORGANISATION DE L'OUTIL

La **Figure 3.4** est la représentation des différents étapes de simulation mises en œuvre dans **PFCL\_Anonymization**.

Notre outil simplifie l'anonymisation en reconnaissant automatiquement les catégories de données (quasi-identifiants, attributs sensibles, attributs identifiants) dans une base de données. Conçu pour améliorer les processus d'anonymisation, il est accessible tant aux novices qu'aux professionnels. En cas de besoins spécifiques, notre outil gère efficacement la détection des catégories de données, facilitant ainsi un processus d'anonymisation rapide et précis dès son lancement. Il offre la possibilité :

- d'importer une base de données à anonymiser,
- de détecter les types d'attributs contenus dans la base de données,
- de choisir les types d'attributs manuellement,
- de créer des hiérarchies de généralisation de façon automatique pour généraliser (remplacer les valeurs spécifiques par des valeurs générales) les valeurs de chaque attribut quasi-identifiant,
- de supprimer automatiquement les attributs identifiants (identifiants directs : email),
- d'enregistrer la base de données anonymisée.

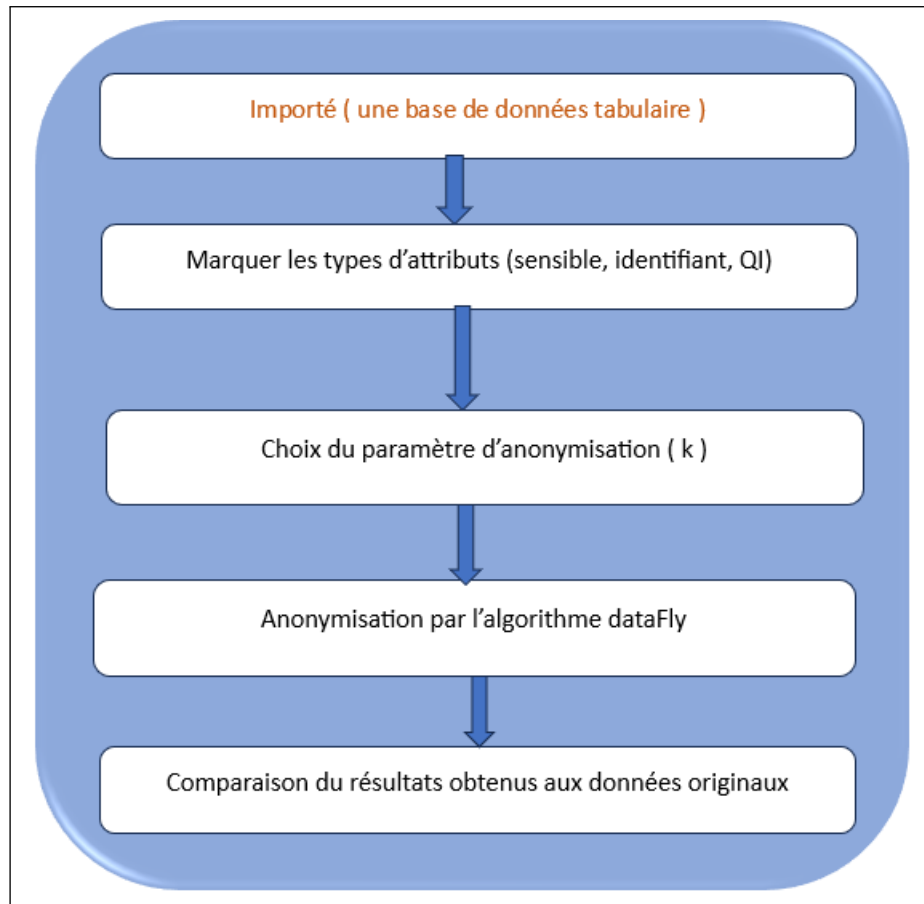
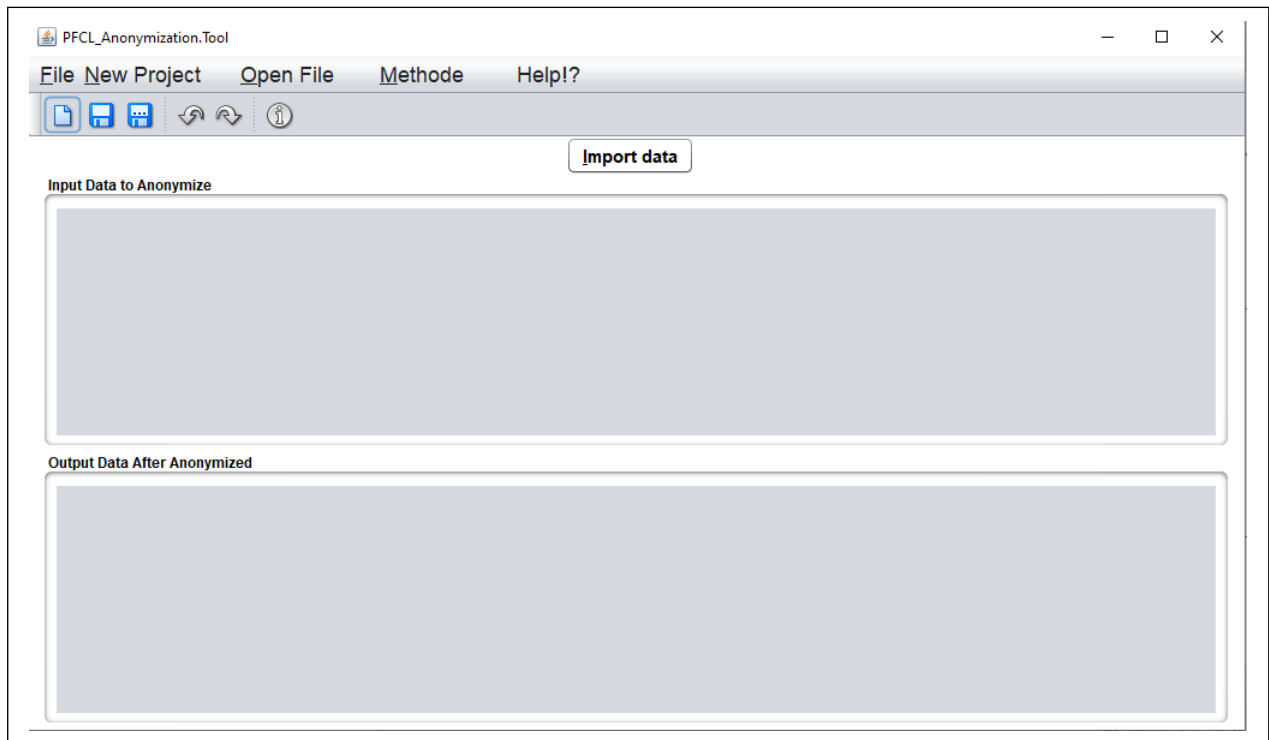


FIGURE 3.4 – Les étapes de simulations de PFCL\_Anonymization

### 3.5 UTILISATION DE L'OUTIL : CAS PRATIQUE

Dans cette section, nous utiliserons un modèle de données (dataset) pour décrire le fonctionnement de **PFCL\_Anonymization**. Nous avons développé une interface utilisateur où l'on peut visualiser simultanément la base de données source et la version anonymisée. Cette fonctionnalité permettra d'observer en temps réel les transformations subies par les données. Ainsi, l'utilisateur pourra facilement décider s'il souhaite sauvegarder les résultats ou relancer le processus d'anonymisation voir **Figure 3.5**. Cette approche promet de rendre l'anonymisation des données non seulement plus transparente, mais aussi plus contrôlable et efficace. Découvrez comment notre interface innovante simplifie et sécurise le processus d'anonymisation des données sensibles, offrant une solution avancée qui guide efficacement les utilisateurs vers une anonymisation optimale des données, adaptée tant aux professionnels qu'aux non-initiés.





**FIGURE 3.5 – Interface graphique de PFCL\_Anonymization**

## 3.5.1 Dataset étudiant

Un dataset est un modèle de données utilisé pour simuler le fonctionnement d'un logiciel. Notre dataset doit être constitué uniquement d'information estudiantine.

L'obtention une base de données réelle pour tester un logiciel est souvent difficile. Par conséquent, nous avons recherché un ensemble de données en ligne, en utilisant Google, qui offre des informations détaillées sur les étudiants.

La base de données que nous avons choisie contient 20 enregistrements et 11 attributs (tels que Nom, Prénom, etc). Le modèle de dataset que nous utilisons est constitué d'attribut identifiant, quasi-identifiant et sensible.

Le dataset que nous utilisons contient les attributs suivants :

### ■ identifiants :

- **nom, prenom** : ce sont des informations patronymiques servant à identifier une personne parmi un ensemble. Une exception est l'existence des homonymes parfaits, ce qui n'est pas le cas de notre base de données,

- **matricule** : chaîne de caractère attribué à une et une personne à la fois,
- **email** : adresse électronique indexant une unique personne à la fois,
- **contact** : numéro permettant d'établir une communication avec une unique personne,

### ■ quasi-identifiants :

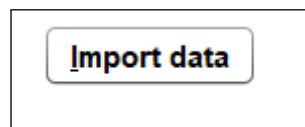
- **parcours** : représente la filière d'un étudiant,
- **niveau** : désigne le niveau d'étude actuel de l'étudiant,
- **age** : information obtenu de la date de naissance,
- **lieu de naissance** : c'est lieu de naissance d'un individu,
- **sexe** : information portant sur le caractère biologique d'un individu,
- **moyenne** : représente la moyenne d'un étudiant,

### ■ sensibles :

- **résultat** : statut d'un étudiant prenant deux valeurs possible : Validé ou Rejeté,
- **mention** : représente la valeur du résultat.

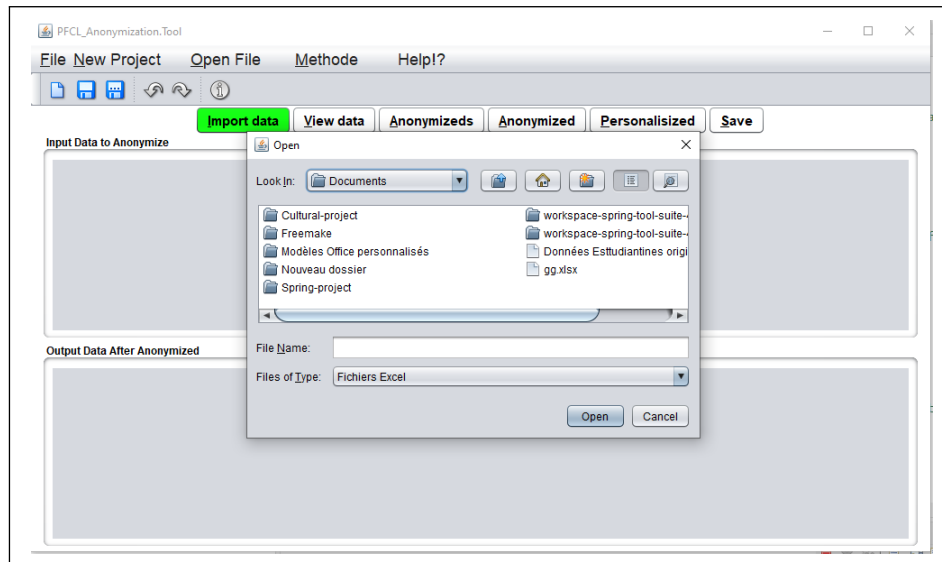
### 3.5.2 Importer une base de données

Avant de découvrir le processus d'anonymisation proposé par **PFCL\_Anonymization**, il est nécessaire d'importer une base de données. Cette étape est possible via l'interface utilisateur du logiciel en utilisant l'option dédiée à l'importation **Figure 3.6**.



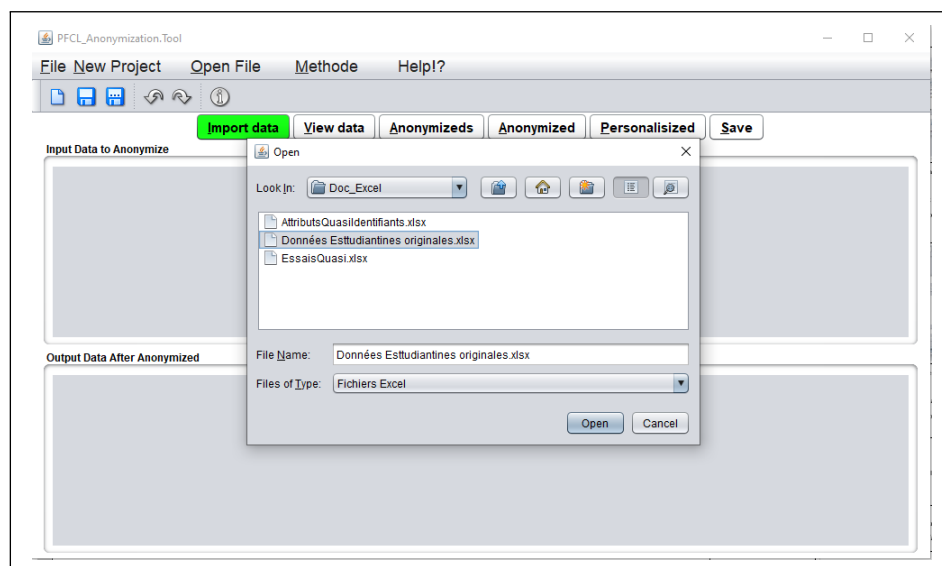
**FIGURE 3.6 – Import de base de donnée**

La base de données à importer doit être au format Excel (.xlsx ou .xls). Pour effectuer l'importation, il existe un cheminement précis à suivre, détaillé à l'aide des illustrations suivantes **Figure 3.7**.



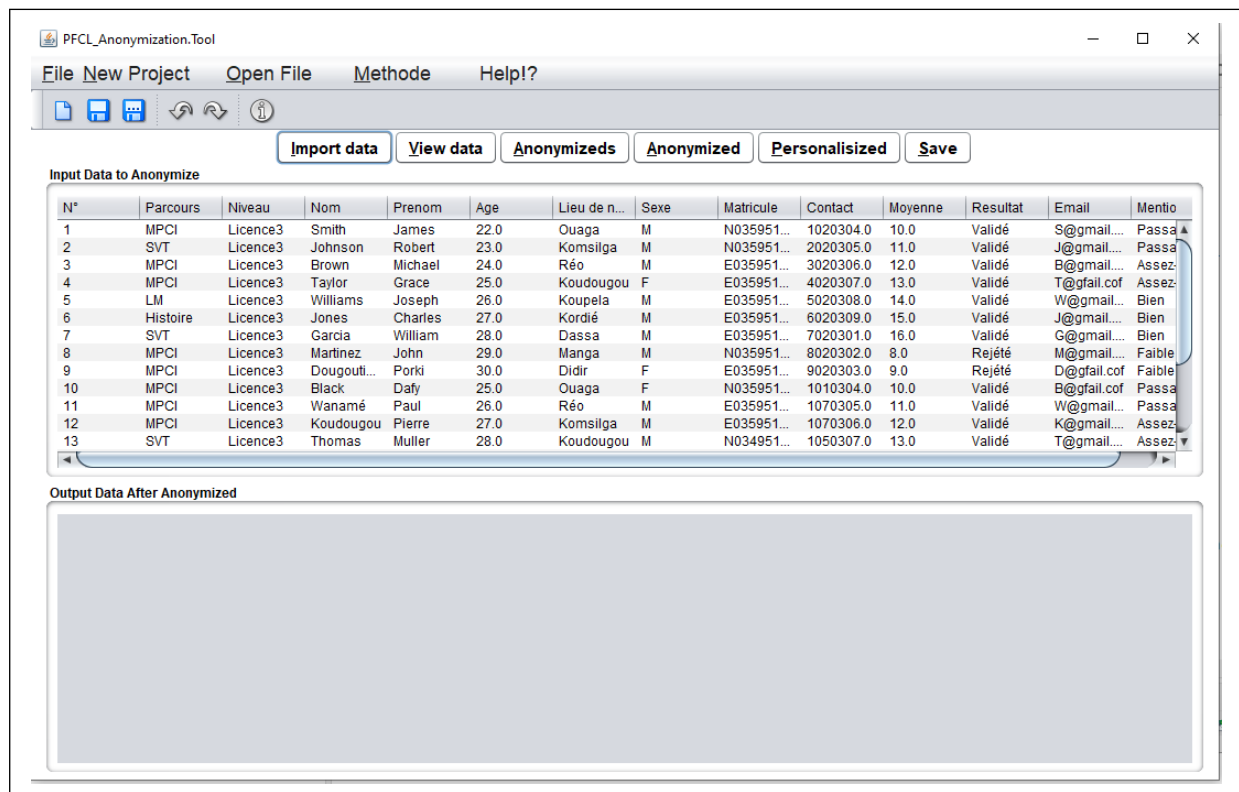
**FIGURE 3.7 – Fenêtre présentant le système de fichier de votre ordinateur**

L'interface de la **Figure 3.7** vous permet de rechercher et d'importer votre base de données directement depuis l'application. Utilisez les différents composants de cette interface pour naviguer et localiser l'emplacement de votre fichier contenant les données.



**FIGURE 3.8 – Téléchargement de la base de donnée**

Depuis l'interface utilisateur de l'application, vous avez la possibilité d'importer votre base de données. La **Figure 3.8** montre cette illustration. Lorsque vous avez localisé l'emplacement de votre base de données, sélectionnez le fichier, puis appuyez sur l'option "**Open**". Cela permettra à l'application de charger votre base de données dans l'interface principale du logiciel.



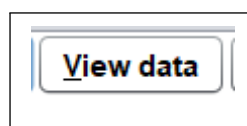
The screenshot shows the PFCL\_Anonymization.Tool application window. It has a menu bar with 'File', 'New Project', 'Open File', 'Methode', and 'Help!?' and a toolbar with icons for file operations. Below the toolbar are buttons for 'Import data', 'View data', 'Anonymizeds', 'Anonymized', 'Personalized', and 'Save'. The main area is titled 'Input Data to Anonymize' and contains a table with 13 rows of data. The table has columns for N°, Parcours, Niveau, Nom, Prenom, Age, Lieu de n..., Sexe, Matricule, Contact, Moyenne, Resultat, Email, and Mentio. The data is as follows:

N°	Parcours	Niveau	Nom	Prenom	Age	Lieu de n...	Sexe	Matricule	Contact	Moyenne	Resultat	Email	Mentio
1	MPCI	Licence3	Smith	James	22.0	Ouaga	M	N035951...	1020304.0	10.0	Validé	S@gmail...	Passa
2	SVT	Licence3	Johnson	Robert	23.0	Komsilga	M	N035951...	2020305.0	11.0	Validé	J@gmail...	Passa
3	MPCI	Licence3	Brown	Michael	24.0	Réo	M	E035951...	3020306.0	12.0	Validé	B@gmail...	Assez
4	MPCI	Licence3	Taylor	Grace	25.0	Koudougou	F	E035951...	4020307.0	13.0	Validé	T@fail.cof	Assez
5	LM	Licence3	Williams	Joseph	26.0	Koupela	M	E035951...	5020308.0	14.0	Validé	W@gmail...	Bien
6	Histoire	Licence3	Jones	Charles	27.0	Kordé	M	E035951...	6020309.0	15.0	Validé	J@gmail...	Bien
7	SVT	Licence3	Garcia	William	28.0	Dassa	M	E035951...	7020301.0	16.0	Validé	G@gmail...	Bien
8	MPCI	Licence3	Martinez	John	29.0	Manga	M	N035951...	8020302.0	8.0	Rejeté	M@gmail...	Faible
9	MPCI	Licence3	Dougouti...	Porki	30.0	Didir	F	E035951...	9020303.0	9.0	Rejeté	D@fail.cof	Faible
10	MPCI	Licence3	Black	Dafy	25.0	Ouaga	F	N035951...	1010304.0	10.0	Validé	B@fail.cof	Passa
11	MPCI	Licence3	Wanamé	Paul	26.0	Réo	M	E035951...	1070305.0	11.0	Validé	W@gmail...	Passa
12	MPCI	Licence3	Koudougou	Pierre	27.0	Komsilga	M	E035951...	1070306.0	12.0	Validé	K@gmail...	Assez
13	SVT	Licence3	Thomas	Muller	28.0	Koudougou	M	N034951...	1050307.0	13.0	Validé	T@gmail...	Assez

Below the table is a section titled 'Output Data After Anonymized' which is currently empty.

**FIGURE 3.9 – Base de données importé**

La **Figure 3.9** présente une base de données importée dans **PFCL\_Anonymization**. En raison du volume important de données, nous avons ajouté des barres de défilement pour permettre la visualisation complète de toutes les informations. La base de données originale reste visible sur l'interface, facilitant ainsi les comparaisons directes entre les données anonymisées et les données sources. La partie graphique "**Output data after anonymize**" contiendra les données de sortie après le processus d'anonymisation. Une fois la base de données affichée, cliquez sur le bouton présenté sur la **Figure 3.10**



**FIGURE 3.10 – Obtention des types de données**

Le bouton de la **Figure 3.10** permet de visualiser la base de données importée, en marquant les attributs selon leur type spécifique. Voici l'affichage obtenu **Figure 3.11** :

**Input Data to Anonymize**

N°	Parcours	Niveau	Nom	Prenom	Age	Lieu de n...	Sexe	Matricule	Contact	Moyenne	Resultat	Email	Mentio
1	MPCI	Licence3	Smith	James	22.0	Ouaga	M	N035951...	1020304.0	10.0	Validé	S@gmail...	Passa
2	SVT	Licence3	Johnson	Robert	23.0	Komsilga	M	N035951...	2020305.0	11.0	Validé	J@gmail...	Passa
3	MPCI	Licence3	Brown	Michael	24.0	Réo	M	E035951...	3020306.0	12.0	Validé	B@gmail...	Assez
4	MPCI	Licence3	Taylor	Grace	25.0	Koudougou	F	E035951...	4020307.0	13.0	Validé	T@gmail...	Assez
5	LM	Licence3	Williams	Joseph	26.0	Koupela	M	E035951...	5020308.0	14.0	Validé	W@gmail...	Bien
6	Histoire	Licence3	Jones	Charles	27.0	Kordié	M	E035951...	6020309.0	15.0	Validé	J@gmail...	Bien
7	SVT	Licence3	Garcia	William	28.0	Dassa	M	E035951...	7020310.0	16.0	Validé	G@gmail...	Bien
8	MPCI	Licence3	Martinez	John	29.0	Manga	M	N035951...	8020302.0	8.0	Rejeté	M@gmail...	Faible
9	MPCI	Licence3	Dougouti...	Porki	30.0	Didir	F	E035951...	9020303.0	9.0	Rejeté	D@gmail...	Faible
10	MPCI	Licence3	Black	Dafy	25.0	Ouaga	F	N035951...	1010304.0	10.0	Validé	B@gmail...	Passa
11	MPCI	Licence3	Wanamé	Paul	26.0	Réo	M	E035951...	1070305.0	11.0	Validé	W@gmail...	Passa
12	MPCI	Licence3	Koudougou	Pierre	27.0	Komsilga	M	E035951...	1070306.0	12.0	Validé	K@gmail...	Assez
13	SVT	Licence3	Thomas	Muller	28.0	Koudougou	M	N035951...	1050307.0	13.0	Validé	T@gmail...	Assez

**Output Data After Anonymized**

N°	Parcours	Niveau	Nom	Prenom	Age	Lieu de	Sexe	Matricule	Contact	Moyenne	Resultat	Email	Mentio
1	MPCI	Licence3	Smith	James	22.0	Ouaga	M	N035951...	1020304.0	10.0	Validé	S@gmail...	Passa
2	SVT	Licence3	Johnson	Robert	23.0	Komsilga	M	N035951...	2020305.0	11.0	Validé	J@gmail...	Passa
3	MPCI	Licence3	Brown	Michael	24.0	Réo	M	E035951...	3020306.0	12.0	Validé	B@gmail...	Assez
4	MPCI	Licence3	Taylor	Grace	25.0	Koudougou	F	E035951...	4020307.0	13.0	Validé	T@gmail...	Assez
5	LM	Licence3	Williams	Joseph	26.0	Koupela	M	E035951...	5020308.0	14.0	Validé	W@gmail...	Bien
6	Histoire	Licence3	Jones	Charles	27.0	Kordié	M	E035951...	6020309.0	15.0	Validé	J@gmail...	Bien
7	SVT	Licence3	Garcia	William	28.0	Dassa	M	E035951...	7020310.0	16.0	Validé	G@gmail...	Bien
8	MPCI	Licence3	Martinez	John	29.0	Manga	M	N035951...	8020302.0	8.0	Rejeté	M@gmail...	Faible
9	MPCI	Licence3	Dougouti...	Porki	30.0	Didir	F	E035951...	9020303.0	9.0	Rejeté	D@gmail...	Faible
10	MPCI	Licence3	Black	Dafy	25.0	Ouaga	F	N035951...	1010304.0	10.0	Validé	B@gmail...	Passa
11	MPCI	Licence3	Wanamé	Paul	26.0	Réo	M	E035951...	1070305.0	11.0	Validé	W@gmail...	Passa
12	MPCI	Licence3	Koudougou	Pierre	27.0	Komsilga	M	E035951...	1070306.0	12.0	Validé	K@gmail...	Assez
13	SVT	Licence3	Thomas	Muller	28.0	Koudougou	M	N035951...	1050307.0	13.0	Validé	T@gmail...	Assez

FIGURE 3.11 – Sélection des attributs

La **Figure 3.11** offre une vue détaillée du processus d'anonymisation grâce à l'algorithme datafly. L'outil **PFCL\_Anonymization** est autonome, capable de reconnaître les types d'attributs d'une base de données importée.

Il convient de noter que les outils d'anonymisation existants attribuent cette tâche à l'utilisateur. Cependant, dans le cadre de notre objectif visant à permettre aux utilisateurs n'ayant aucune expertise de fournir des données anonymes ou aux professionnels du domaine d'établir rapidement un processus d'anonymisation, il était impératif pour nous d'intégrer cette fonctionnalité dans notre outil. Expliquons maintenant le visuel de la figure3.11.

- ☛ **les attributs affectés d'un point rouge** représentent les identifiants explicites de la base de donnée importé,
- ☛ **les attributs marqués par le jaune** sont les attributs quasi-identifiants,
- ☛ **Les attributs marqués au vert** sont les attributs sensibles.

### 3.5.3 Choix du paramètre d'anonymisation

Dans l'outil implémenté, avant l'anonymisation des données proprement dite, il faut obligatoirement choisir le **paramètre d'anonymisation k**.

**k** est l'ensemble des attributs quasi-identifiants qui ont la même valeur.

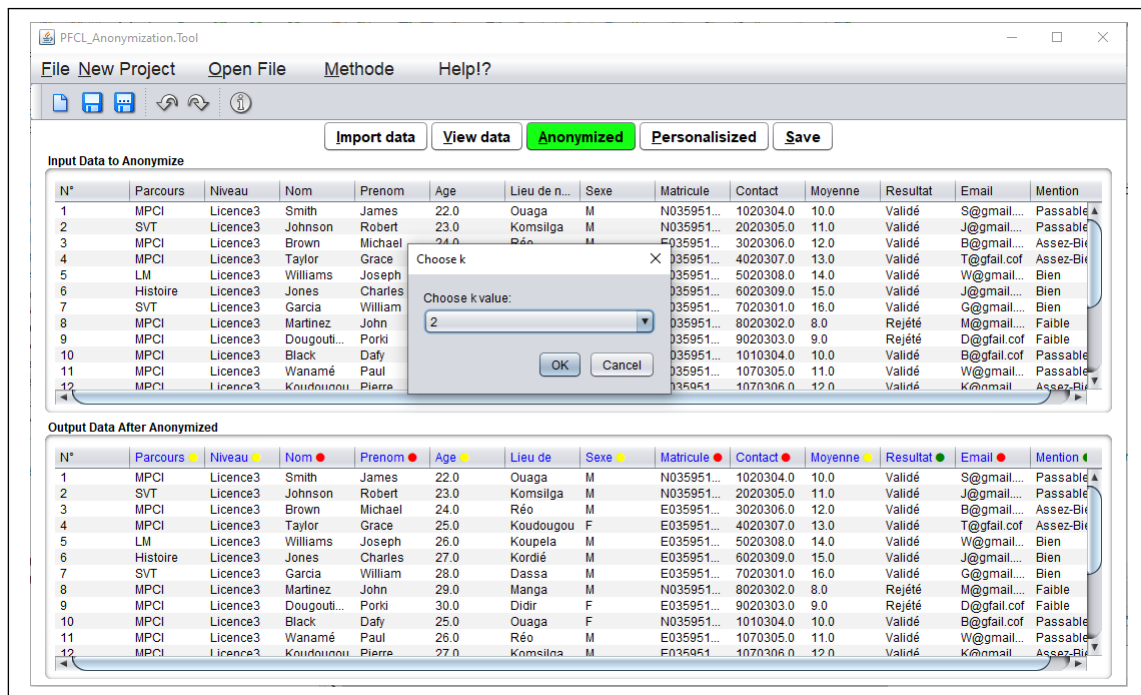


FIGURE 3.12 – Interface du choix de k

### 3.5.4 Construction de la hiérarchie de généralisation

Après la sélection du paramètre d'anonymisation, l'outil construit de manière autonome la hiérarchie de généralisation pour chaque attribut, en fonction de son type spécifique.

Concernant les attributs identifiants, une suppression globale des valeurs de chaque attribut est appliquée, car ces données sont particulièrement critiques en termes de ré-identification des individus.

Quant aux attributs quasi-identifiants, diverses techniques sont appliquées, telles que la généralisation sous forme d'intervalle pour les attributs numériques et le remplacement des valeurs distinctes par des valeurs plus générales.

### 3.5.5 Comparaison des résultats

Après le choix de la valeur de  $k$  (degré d'anonymisation), et ayant validé ce choix, on obtient :

N°	Parcours	Niveau	Nom	Prenom	Age	Lieu de n...	Sexe	Matricule	Contact	Moyenne	Resultat	Email	Mentio
1	MPCI	Licence3	Smith	James	22.0	Ouaga	M	N035951...	1020304.0	10.0	Validé	S@gmail...	Passa
2	SVT	Licence3	Johnson	Robert	23.0	Komsilga	M	N035951...	2020305.0	11.0	Validé	J@gmail...	Passa
3	MPCI	Licence3	Brown	Michael	24.0	Réo	M	E035951...	3020306.0	12.0	Validé	B@gmail...	Assez
4	MPCI	Licence3	Taylor	Grace	25.0	Koudougou	F	E035951...	4020307.0	13.0	Validé	T@gtail.cof	Assez
5	LM	Licence3	Williams	Joseph	26.0	Koupela	M	E035951...	5020308.0	14.0	Validé	W@gmail...	Bien
6	Histoire	Licence3	Jones	Charles	27.0	Kordié	M	E035951...	6020309.0	15.0	Validé	J@gmail...	Bien
7	SVT	Licence3	Garcia	William	28.0	Dassa	M	E035951...	7020301.0	16.0	Validé	G@gmail...	Bien
8	MPCI	Licence3	Martinez	John	29.0	Manga	M	N035951...	8020302.0	8.0	Rejeté	M@gmail...	Faible
9	MPCI	Licence3	Dougouti...	Porki	30.0	Didir	F	E035951...	9020303.0	9.0	Rejeté	D@gtail.cof	Faible
10	MPCI	Licence3	Black	Dafy	25.0	Ouaga	F	N035951...	1010304.0	10.0	Validé	B@gtail.cof	Passa
11	MPCI	Licence3	Wanamé	Paul	26.0	Réo	M	E035951...	1070305.0	11.0	Validé	W@gmail...	Passa
12	MPCI	Licence3	Koudougou	Pierre	27.0	Komsilga	M	E035951...	1070306.0	12.0	Validé	K@gmail...	Assez

N°	Parcours	Niveau	Nom	Prenom	Age	Lieu de naissance	Sexe	Matricule	Contact	Moyenne	Resultat	Email
1	ST	Licence*	*	*	21-25	Burkina Faso	M, F	*	*	6-10	Validé	*
2	ST	Licence*	*	*	21-25	Burkina Faso	M, F	*	*	11-15	Validé	*
3	ST	Licence*	*	*	21-25	Burkina Faso	M, F	*	*	11-15	Validé	*
4	ST	Licence*	*	*	21-25	Burkina Faso	M, F	*	*	11-15	Validé	*
5	LSH	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	11-15	Validé	*
6	LSH	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	11-15	Validé	*
7	ST	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	16-20	Validé	*
8	ST	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	6-10	Rejeté	*
9	ST	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	6-10	Rejeté	*
10	ST	Licence*	*	*	21-25	Burkina Faso	M, F	*	*	6-10	Validé	*
11	ST	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	11-15	Validé	*
12	ST	Licence*	*	*	26-30	Burkina Faso	M, F	*	*	11-15	Validé	*

FIGURE 3.13 – Résultat final de l'algorithme

La Figure 3.13 offre une vue d'ensemble de la base de données après anonymisation, présentant une version plus générale par rapport à la base de données originale.

Les attributs identifiants ont été entièrement supprimés de la base de données. Ainsi, même si un attaquant potentiel parvient à accéder à cette base de données, il ne sera plus en mesure d'identifier directement un individu de la base de données source. Prenons l'exemple de la toute première personne de l'ensemble des données sources, aucune information dans la base de données anonymisée ne permettrait d'identifier Smith James. Désormais, la seule option pour un attaquant de remonter à une personne spécifique serait d'analyser les données quasi-identifiants.

Les données quasi-identifiants comprennent à la fois des données catégorielles (type chaîne de caractère) et des données continues (type numérique). Pour ces types de données, nous

avons appliqués une généralisation des valeurs, préservant ainsi la véracité des données tout en réduisant leur précision.

Reprenant l'exemple des données identifiantes, si nous examinons l'attribut sexe, nous ne pouvons pas être rassuré du sexe d'Alice mais les valeurs de l'attribut restent vraies, car elles appartiennent à une personne de la base de données.

Les données sensibles, bien qu'elles ne permettent pas d'identifier une personne par elles-mêmes, sont d'un intérêt primordial pour les analystes. C'est pourquoi nous avons choisi de les laisser inchangées. Une des raisons est que le modèle mis en œuvre ne permet pas de les traiter.

## CONCLUSION

Dans ce chapitre, nous avons réalisé l'implémentation de **PFCL\_Anonymization**, un outil dédié à l'anonymisation des données tabulaires, en utilisant l'algorithme datafly et mettant en œuvre le modèle k-anonymat. Notre outil fournit un ensemble de données anonymes capable de contrer les attaques par liens d'enregistrement et de prévenir l'inférence (possibilité de déduire de façon quasi certaine des informations sur une personne), la corrélation (possibilité de retrouver les informations propre à une personne) et l'individualisation (possibilité d'isoler un individu).

Il est important de souligner que notre outil se concentre sur l'application d'un seul modèle d'anonymisation. Bien qu'il soit efficace pour couvrir les attaques par liens d'enregistrement, il pourrait présenter des failles face à d'autres types d'attaques telles que les liens de table ou l'inférence probabiliste. Ces aspects n'ont pas été pris en compte dans cette version de l'outil.

Pour cette première version de **PFCL\_Anonymization**, nous avons adopté une architecture simple et efficace tout en prévoyant des possibilités d'amélioration. Dans les prochaines versions, nous envisageons d'intégrer d'autres modèles d'anonymisation pour renforcer la robustesse de l'outil et mieux répondre aux diverses menaces potentielles.





---

## CONCLUSION GÉNÉRALE

Les données sont devenues un atout concurrentiel majeur pour les organisations, mais elles contiennent également des informations personnelles sensibles telles que les salaires et les états de santé. Dans l'ère numérique actuelle, la protection de la vie privée est devenue un droit fondamental, ce qui pose le défi de concilier confidentialité et utilité des données.

Pour relever ce défi, de nombreuses techniques et algorithmes d'anonymisation des données sont disponibles. Ils modifient les données originales pour réduire les risques de ré-identification tout en préservant leur utilité. Cependant, il n'existe pas d'algorithme universel qui convienne à tous les contextes. Le choix de l'algorithme optimal dépend des spécificités des données et des exigences de confidentialité.

Dans le cadre de notre projet tuteuré, nous avons réalisé un état de l'art sur l'anonymisation des données tabulaires, identifiant deux grandes familles de techniques. Cette exploration nous a permis de concevoir **PFCL\_Anonymization**, un outil visant à anonymiser les données des étudiants.

Notre outil implémente le modèle de k-anonymat et utilise l'algorithme Datafly pour maintenir un équilibre entre protection et utilité des données. Il offre aux non-professionnels une solution conviviale pour anonymiser les données sans expertise préalable, tout en offrant aux professionnels une automatisation des tâches méticuleuses telles que la détection des types d'attributs et le choix des techniques d'anonymisation.

## PERSPECTIVES DE RECHERCHE

En ce qui concerne les perspectives de recherche, nous envisageons plusieurs pistes :

- ☛ implémentation du modèle l-diversité pour permettre une diversité dans les attributs sensibles.
- ☛ intégration de l'intelligence artificielle pour détecter automatiquement les catégories de données.
- ☛ intégration de techniques pour évaluer le degré de protection et d'utilité des données anonymisées.
- ☛ construction de hiérarchies de généralisation plus robustes.



---

## Bibliographie

- [1] *RGPD : les règles de l'anonymisation des données*, fr. adresse : <https://www.letolito.com/guides/rgpd-les-regles-de-lanonymisation-des-donnees> (visité le 23/02/2024).
- [2] *Pseudonymisation des données : principes et techniques*, fr-FR, fév. 2023. adresse : <https://www.vaadata.com/blog/fr/pseudonymisation-des-donnees-principes-techniques-et-bonnes-pratiques/> (visité le 23/02/2024).
- [3] *JUSTICE AND CONSUMERS ARTICLE 29 - Item Overview*. adresse : <https://ec.europa.eu/newsroom/article29/items> (visité le 17/03/2024).
- [4] *Anonymisation des données, une définition - ZDNet*. adresse : <https://www.zdnet.fr/lexique-it/anonymisation-des-donnees-une-definition-39925683.htm> (visité le 22/02/2024).
- [5] *L'anonymisation des données, un traitement clé pour l'open data*, fr. adresse : <https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data> (visité le 22/02/2024).
- [6] *Figure 1. Collecte et publication des données (Y. Xu et al. 2014)*, en. adresse : [https://www.researchgate.net/figure/Collecte-et-publication-des-donnees-Y-Xu-et-al-2014\\_fig1\\_324938181](https://www.researchgate.net/figure/Collecte-et-publication-des-donnees-Y-Xu-et-al-2014_fig1_324938181) (visité le 15/04/2024).
- [7] F. B. FREDJ, « Méthode et outil d'anonymisation des données sensibles, » fr, thèse de doct., Conservatoire national des arts et métiers - CNAM; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, juill. 2017. adresse : <https://theses.hal.science/tel-01783967> (visité le 02/03/2024).
- [8] F. B. FREDJ, « Méthode et outil d'anonymisation des données sensibles, » PhD Thesis, Conservatoire national des arts et métiers-CNAM; Université de Sfax (Tunisie ...), 2017.

- [9] I. K. GAYKI et A. S. KAPSE, « Privacy Preservation of Published Data Using Anonymization Technique, » en, t. 5, 2014.
- [10] D. P. O. PARTAGÉ, *Cas Pratique : K-anonymat - méthode d'anonymisation*, fr-FR, mars 2023. adresse : <https://www.dpo-partage.fr/cas-pratique-k-anonymat-methode-danonymisation/> (visité le 05/03/2024).
- [11] *Analyse des risques de la restauration de l'identification | Documentation sur la protection des données sensibles*, fr. adresse : <https://cloud.google.com/sensitive-data-protection/docs/concepts-risk-analysis?hl=fr> (visité le 15/06/2024).
- [12] A. ELOUARDIGHI, M. MAGHFOUR, H. HAMMIA et F.-Z. AAZI, « Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'apprentissage automatique, » fr,
- [13] *Datafly algorithm*, fr, Page Version ID : 1189130380, déc. 2023. adresse : [https://en.wikipedia.org/w/index.php?title=Datafly\\_algorithm&oldid=1189130380](https://en.wikipedia.org/w/index.php?title=Datafly_algorithm&oldid=1189130380) (visité le 04/03/2024).
- [14] S. MÉMOIRE, *L'anonymisation de micro-données à des fins de publication*, fr-FR, nov. 2023. adresse : <https://www.rapport-gratuit.com/lanonymisation-de-micro-donnees-a-des-fins-de-publication/> (visité le 06/03/2024).
- [15] *Quelles techniques d'anonymisation pour protéger vos données personnelles ?* Adresse : <https://www.apssis.com/actualite-ssi/618/quelles-techniques-d-anonymisation-pour-proteger-vos-donnees-personnelles.htm> (visité le 23/02/2024).
- [16] *ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing*, en-US. adresse : <https://arx.deidentifier.org/> (visité le 10/06/2024).
- [17] *Avantages et inconvénients de Java*, fr-FR, mars 2022. adresse : <https://innowise.com/fr/blog/benefits-and-drawbacks-of-java/> (visité le 30/04/2024).
- [18] *Développons en Java - Le développement d'interfaces graphiques avec SWING*. adresse : <https://www.jmdoudoux.fr/java/dej/chap-swing.htm> (visité le 23/05/2024).
- [19] *Apache POI - the Java API for Microsoft Documents*. adresse : <https://poi.apache.org/> (visité le 23/05/2024).

- [20] *The Community for Open Collaboration and Innovation | The Eclipse Foundation*. adresse : <https://www.eclipse.org/> (visité le 23/05/2024).

## ANNEXES

### A.1 CODE DE DÉTECTION DES CATÉGORIES DE DONNÉES

```
53
54 // méthode pour marquer les attributs quasi-identifiants au niveau des en-tête
55 public void markSensitiveIdentifyingAttribut(List<String> headers) {
56     // Parcourir les en-têtes pour trouver les attributs quasi-identifiants
57     List<Integer> quasiIdentifyingColumns = new ArrayList<>();
58     for (int i = 0; i < headers.size(); i++) {
59         String header = headers.get(i);
60
61         // Verifier si l'en-tête correspond à un quasi-identifiant
62         if (isSensitiveIdentifyingAttribut(header)) {
63             // Si oui, ajouter un repère visuel
64             headers.set(i, "<html><font color='blue'>" + header + "</font> <fo
65             quasiIdentifyingColumns.add(i);
66         }
67     }
68 }
69
70 }
```

FIGURE A.1 – Code pour marquer les attributs sensibles

```
74
75 méthode pour marquer les attributs identifiants au niveau des en-tête
76 public void markIdentifyingAttributes(List<String> headers) {
77
78     // Parcourir les en-tête pour trouver les attributs identifiants
79     for(int i= 0; i< headers.size(); i++) {
80         String header= headers.get(i);
81
82         // Verifier si l'attribut correspond à un identifiant
83         if(isIdentifyingAttribut(header)) {
84             // si oui, ajpouter un repère visuel
85             headers.set(i, "<html><font color='blue'>" + header + "</f
86
87         }
88     }
89 }
```

FIGURE A.2 – Code pour marquer les attributs Identifiants

```
193
194 // méthode pour marquer les attributs quasi-identifiants au niveau des en-tête
195 public void markQuasiIdentifyingAttribut(List<String> headers) {
196     // Parcourir les en-têtes pour trouver les attributs quasi-identifiants
197     List<Integer> quasiIdentifyingColumns = new ArrayList<>();
198     for (int i = 0; i < headers.size(); i++) {
199         String header = headers.get(i);
200
201         // Verifier si l'en-tête correspond à un quasi-identifiant
202         if (isQuasiIdentifyingAttribut(header)) {
203             // Si oui, ajouter un repère visuel
204             headers.set(i, "<html><font color='blue'>" + header + "</font> <fo
205             quasiIdentifyingColumns.add(i);
206         }
207     }
208 }
209
210 }
```

FIGURE A.3 – Code pour marquer les attributs QI

---

---

**Landri BAYILI**

---

---

**Résumé :**

À l'ère du numérique actuelle, la protection des données personnelles est devenue un droit fondamental, en particulier pour les données étudiantes qui contiennent souvent des informations sensibles. Ce travail a permis le développement d'un outil robuste et convivial.

"PFCL\_Anonymization" est conçu pour anonymiser efficacement les bases de données étudiantes. Cet outil utilise l'algorithme datafly et implémente le modèle de k-anonymat pour équilibrer confidentialité et utilité des données. la méthodologie utilisée comprend l'importation d'une base de données au format Excel, l'application de techniques d'anonymisation et l'exportation des données anonymisées, le tout via une interface intuitive. l'outil automatise l'identification des types d'attributs et propose des hiérarchies de généralisation, le rendant accessible même aux non-spécialistes. Les résultats montrent que "PFCL\_Anonymization" anonymise avec succès les données étudiantes tout en maintenant leur pertinence pour les analyses. Ce travail répond au besoin immédiat de protection des données étudiantes et propose une solution évolutive pour des améliorations futures.

**Mots clés :** anonymiser, Algorithme datafly, modèle k-anonymat, Hiérarchie de généralisation, informations sensibles, données.

**Abstract :**

In today's digital era, the protection of personal data has become a fundamental right, particularly in the context of student data, which often includes sensitive information. Our project aims to develop a robust and user-friendly tool. "PFCL\_Anonymization" designed to anonymize student databases effectively. This tool employs the DataFly algorithm and implements the k-anonymity model to balance data confidentiality and utility. The challenges of data anonymization include preserving data utility while ensuring confidentiality. Our methodology involves importing a database in Excel format, applying anonymization techniques, and exporting the anonymized data, all through an intuitive interface. Our tool automates the identification of attribute types and proposes generalization hierarchies, making it accessible even to non-specialists. The results show that "PFCL\_Anonymization" successfully anonymizes student data while maintaining its relevance for analysis. This work meets the immediate need for student data protection and offers an evolutionary solution for future improvements.

**Key-words :** anonymize, DataFly algorithm, k-anonymity model, generalization hierarchy, sensitive information, data.

