UQAC
Université du Québec
à Chicoutimi

CONCEPTION D'UN OUTIL D'ANONYMISATION CONFORME AU REGLEMENT QUEBECOIS

PAR NANA KADIJA KABA

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

Aujourd'hui, les données numériques sont partout et on les réutilise de plus en plus pour la recherche et les statistiques. Mais cette utilisation massive pose un problème majeur : comment protéger la vie privée des personnes concernées ? Au Québec, la Loi 25 a mis en place des règles strictes pour encadrer l'anonymisation des renseignements personnels.

Ce mémoire propose la conception d'un outil d'anonymisation conforme aux exigences du règlement québécois. L'outil évalue les risques associés à chaque colonne de données selon trois critères (corrélation, individualisation, inférence), applique des méthodes d'anonymisation adaptées (suppression, généralisation et confidentialité différentielle), puis simule les risques résiduels de réidentification.

Développé en Python avec une interface intuitive, il vise à faciliter la conformité réglementaire accessible aux chercheurs et les petites organisations qui n'ont pas forcément les moyens d'engager des experts en sécurité.

ABSTRACT

Today, digital data is everywhere and increasingly reused for research and statistical purposes. However, this widespread use raises a critical question : how can we protect the privacy of the individuals concerned ? In Quebec, Bill 25 has established strict regulations to govern the anonymization of personal information.

This thesis proposes the design of an anonymization tool that complies with Quebec's regulatory requirements. The tool assesses the risks associated with each data column according to three criteria (correlation, individualization, inference), applies appropriate anonymization methods (suppression, masking, generalization, differential privacy), and then simulates residual re-identification risks.

Developed in Python with an intuitive interface, it aims to make regulatory compliance accessible to researchers and small organizations that may not have the resources to hire security experts.

TABLE DES MATIÈRES

RÉSUMÉ ii
ABSTRACT iii
LISTE DES TABLEAUX x
LISTE DES FIGURES xi
LISTE DES ABRÉVIATIONS xiii
DÉDICACE xiv
REMERCIEMENTS xv
AVANT-PROPOS xvi
INTRODUCTION 1
CHAPITRE I – INTRODUCTION 2
1.1 CONTEXTE 2
1.2 MOTIVATION 6
1.3 OBJECTIFS 8
1.3.1 OBJECTIF GÉNÉRAL 9
1.3.2 OBJECTIFS SPÉCIFIQUES 9
1.4 CONTEXTE DE RECHERCHE 12
1.5 CONTRIBUTIONS 12
1.6 ORGANISATION DU DOCUMENT 13
CHAPITRE II – ÉTAT DE L'ART 15
2.1 NOTIONS CLÉS 15
2.2 MODÈLES THÉORIQUES DE RÉFÉRENCE 16
2.3 TECHNIQUES D'ANONYMISATION 17
2.3.1 SUPPRESSION DES IDENTIFIANTS DIRECTS 17
2.3.2 PSEUDONYMISATION 17
2.3.3 MASQUAGE PARTIEL 18
2.3.4 GÉNÉRALISATION ET AGRÉGATION 18

2.3.5 CONFIDENTIALITÉ DIFFÉRENTIELLE 18
2.3.6 LE COMPROMIS ENTRE PROTECTION ET UTILITÉ 19
2.4 CADRE LÉGAL INTERNATIONAL ET COMPARAISONS 19
2.4.1 LE RGPD EN EUROPE 19
2.4.2 HIPAA ET NIST AUX ÉTATS-UNIS 20
2.4.3 NORMES INTERNATIONALES ISO 20
2.4.4 APPORTS SPÉCIFIQUES DU RÈGLEMENT QUÉBÉCOIS 20
2.5 RISQUES DE RÉIDENTIFICATION : CATÉGORIES ET MESURES 21
2.5.1 RISQUE DE LIAISON 21
2.5.2 RISQUE D'ATTRIBUTION 21
2.5.3 RISQUE D'INFÉRENCE 22
2.5.4 INDICATEURS QUANTITATIFS DE MESURE DES RISQUES 22
2.6 OUTILS EXISTANTS D'ANONYMISATION DES DONNÉES 24
2.6.1 ARX 24
2.6.2 AMNESIA 24
2.6.3 SOLUTIONS COMMERCIALES 25
2.7 LIMITES DES APPROCHES ACTUELLES 26
2.7.1 LES MODÈLES CLASSIQUES NE COUVRENT PAS TOUS LES RISQUES 26
2.7.2 LE COMPROMIS UTILITÉ-PROTECTION 26
2.7.3 CONFIDENTIALITÉ DIFFÉRENTIELLE : DIFFICILE À PARAMÉTRER 27
2.7.4 AUCUN OUTIL NE COUVRE LES TROIS CRITÈRES DE LA LOI 25 DU RÈGLEMENT QUÉBÉCOIS 27
2.8 POSITION DU QUÉBEC DANS LE PAYSAGE INTERNATIONAL 28
2.8.1 CE QUI REND LE RÈGLEMENT QUÉBÉCOIS UNIQUE 28
2.8.2 UNE DÉMARCHE STRUCTURÉE OBLIGATOIRE 28
2.8.3 COMPARAISON AVEC LES AUTRES CADRES 28
2.8.4 LE QUÉBEC COMME LABORATOIRE INTERNATIONAL 29
2.8.5 LE DÉFI POUR LES ORGANISATIONS QUÉBÉCOISES 29

2.9 SYNTHÈSE DE L'ÉTAT DE L'ART 30
2.9.1 CE QUE RÉVÈLE LA REVUE DE LITTÉRATURE 30
2.9.2 LE BESOIN IDENTIFIÉ 30
2.9.3 LA SUITE DU MÉMOIRE 31
CHAPITRE III – CADRE THÉORIQUE ET RÉGLEMENTAIRE 32
3.1 CONCEPTS FONDAMENTAUX DE L'ANONYMISATION 33
3.1.1 IDENTIFIANTS DIRECTS 33
3.1.2 QUASI-IDENTIFIANTS 34
3.1.3 DONNÉES SENSIBLES 34
3.1.4 ANONYMISATION VS PSEUDONYMISATION 35
3.2 LES TROIS CRITÈRES DU RÈGLEMENT QUÉBÉCOIS 35
3.2.1 CRITÈRE 1 : INDIVIDUALISATION 36
3.2.2 CRITÈRE 2 : CORRÉLATION 36
3.2.3 CRITÈRE 3 : INFÉRENCE 37
3.3 LE PROCESSUS RÉGLEMENTAIRE D'ANONYMISATION 38
3.3.1 ÉTAPE 1 : DÉFINITION DE L'OBJECTIF DE L'ANONYMISATION . 38
3.3.2 ÉTAPE 2 : RETRAIT DES IDENTIFIANTS DIRECTS 38
3.3.3 ÉTAPE 3 : ANALYSE DES RISQUES AVANT ANONYMISATION . . 39
3.3.4 ÉTAPE 4 : APPLICATION DE TECHNIQUES D'ANONYMISATION . 39
3.3.5 ÉTAPE 5 : ANALYSE DES RISQUES APRÈS ANONYMISATION . . 39
3.3.6 ÉTAPE 6 : DOCUMENTATION ET REGISTRE 40
3.4 APPROCHE DE L'ADVERSAIRE RAISONNABLE 40
3.4.1 C'EST QUOI UN ADVERSAIRE RAISONNABLE ? 40
3.4.2 POURQUOI C'EST IMPORTANT ? 41
3.5 LIEN ENTRE LE RÈGLEMENT QUÉBÉCOIS ET LES MODÈLES THÉORIQUES 41
3.5.1 SYNTHÈSE DU CHAPITRE 42
CHAPITRE IV – CONCEPTION ET MÉTHODOLOGIE 44
4.1 LA DÉMARCHE MÉTHODOLOGIQUE 44

4.1.1 APPROCHE ITÉRATIVE 44
4.1.2 LES SEPT PHASES DU PROJET 44
4.1.3 MODÈLE DE PROTECTION PROPOSÉ 45
4.2 ANALYSE DES BESOINS 46
4.2.1 EXIGENCES RÉGLEMENTAIRES ET CHOIX D'IMPLÉMENTATION 46
4.2.2 EXIGENCES FONCTIONNELLES 47
4.2.3 EXIGENCES NON-FONCTIONNELLES 48
4.3 ARCHITECTURE DE L'OUTIL 48
4.3.1 VUE D'ENSEMBLE 48
4.3.2 POURQUOI CETTE ARCHITECTURE ? 48
4.3.3 ARCHITECTURE BACKEND 49
4.3.4 ARCHITECTURE FRONTEND 51
4.3.5 BASE DE DONNÉES 52
4.4 CHOIX TECHNOLOGIQUES 53
4.4.1 BACKEND : FASTAPI + PYTHON 53
4.4.2 FRONTEND : NEXT.JS + REACT 53
4.4.3 BASE DE DONNÉES : POSTGRESQL 53
4.4.4 BIBLIOTHÈQUES PYTHON PRINCIPALES 54
4.4.5 DÉPLOIEMENT : DOCKER 54
4.5 PROCESSUS D'ANONYMISATION 54
4.5.1 CHOIX DES TECHNIQUES D'ANONYMISATION 54
4.5.2 ÉTAPE 1 : CHARGEMENT ET VALIDATION 56
4.5.3 ÉTAPE 2 : DÉTECTION AUTOMATIQUE 56
4.5.4 ÉTAPE 3 : ÉVALUATION PRÉ-ANONYMISATION 57
4.5.5 ÉTAPE 4 : CONFIGURATION DES TRANSFORMATIONS 58
4.5.6 ÉTAPE 5 : APPLICATION DES TECHNIQUES 59
4.5.7 ÉTAPE 6 : VÉRIFICATION POST-ANONYMISATION 61
4.6 MÉTHODE D'ÉVALUATION DES RISQUES 61

4.6.1 CRITÈRE 1 : INDIVIDUALISATION (40%) 61
4.6.2 CRITÈRE 2 : CORRÉLATION (35%) 62
4.6.3 CRITÈRE 3 : INFÉRENCE (25%) 62
4.6.4 SCORE GLOBAL ET CONFORMITÉ 63
4.6.5 RISQUE RÉSIDUEL MINIMUM 63
4.7 GÉNÉRATION DE RAPPORTS 63
4.8 TESTS ET VALIDATION 64
CHAPITRE V – IMPLÉMENTATION ET RÉSULTATS 65
5.1 ENVIRONNEMENT DE DÉVELOPPEMENT 65
5.1.1 CONTEXTE MATÉRIEL 65
5.1.2 LOGICIELS UTILISÉS 65
5.2 IMPLÉMENTATION DU BACKEND 66
5.2.1 STRUCTURE PRINCIPALE DU CODE 66
5.2.2 DÉTECTION AUTOMATIQUE DES COLONNES SENSIBLES 66
5.2.3 TECHNIQUES D'ANONYMISATION IMPLÉMENTÉES 66
5.2.4 ÉVALUATION DES RISQUES 67
5.3 IMPLÉMENTATION DU FRONTEND 67
5.3.1 ÉTAPE 1 : CONNEXION ET IMPORT DU FICHIER 67
5.3.2 ÉTAPE 2 : PRÉVISUALISATION ET DÉTECTION 71
5.3.3 ÉTAPE 3 : CONFIGURATION DE L'ANONYMISATION 78
5.3.4 ÉTAPE 4 : RÉSULTATS ET CONFORMITÉ 81
5.4 TESTS ET RÉSULTATS 84
5.4.1 DATASET DE TEST 84
5.4.2 RÉSULTATS DE DÉTECTION 85
5.4.3 TRANSFORMATIONS APPLIQUÉES 85
5.4.4 SCORES DE CONFORMITÉ 85
5.5 ANALYSE CRITIQUE 86
5.5.1 POINTS FORTS 86

5.5.2 LIMITES 87
5.5.3 DÉFIS TECHNIQUES 87
5.6 SYNTHÈSE DU CHAPITRE 87
CHAPITRE VI – CONCLUSION ET PERSPECTIVES 89
6.1 RAPPEL DE LA PROBLÉMATIQUE 89
6.2 SYNTHÈSE DES CONTRIBUTIONS 90
6.2.1 CONTRIBUTION THÉORIQUE 90
6.2.2 CONTRIBUTION MÉTHODOLOGIQUE 90
6.2.3 CONTRIBUTION TECHNIQUE 90
6.2.4 CONTRIBUTION EMPIRIQUE 91
6.3 RÉSULTATS PRINCIPAUX 91
6.3.1 EFFICACITÉ DE L'OUTIL 91
6.3.2 VALIDATION DES CHOIX TECHNIQUES 91
6.3.3 LIMITES IDENTIFIÉES 91
6.4 PERSPECTIVES D'AMÉLIORATION 93
6.4.1 AMÉLIORATIONS À COURT TERME 93
6.4.2 AMÉLIORATIONS À MOYEN TERME 93
6.4.3 PERSPECTIVES À LONG TERME 94
6.5 IMPLICATIONS PRATIQUES 94
6.5.1 POUR LES ORGANISATIONS QUÉBÉCOISES 94
6.5.2 POUR LA RECHERCHE EN PROTECTION DE LA VIE PRIVÉE . . . 95
6.5.3 POUR L'ÉVOLUTION DE LA RÉGLEMENTATION 95
6.6 RÉFLEXION PERSONNELLE 95
6.7 MOT DE LA FIN 96
BIBLIOGRAPHIE 97

LISTE DES TABLEAUX

TABLEAU 2.1 : COMPARAISON DES OUTILS D'ANONYMISATION 25

TABLEAU 2.2 : COMPARAISON DÉTAILLÉE DES EXIGENCES D'ANONYMISATION 29

TABLEAU 3.1 : CORRESPONDANCE ENTRE CRITÈRES QUÉBÉCOIS ET MODÈLES THÉORIQUES 42

TABLEAU 4.1 : RELATIONS ENTRE MODÈLES DE BASE DE DONNÉES 51

TABLEAU 5.1 : SYNTHÈSE DE LA CLASSIFICATION AUTOMATIQUE DU DATASET `PATIENTS.CSV` 85

TABLEAU 5.2 : TECHNIQUES APPLIQUÉES PAR COLONNE 86

TABLEAU 5.3 : SCORES D'ÉVALUATION DES RISQUES POST-ANONYMISATION 86

LISTE DES FIGURES

FIGURE 1.1 – DÉPENSES EN CYBERSÉCURITÉ PAR SEGMENT EN 2020 ET 2021. SOURCE : (Renaud, 2024) 3

FIGURE 1.2 – RÉPARTITION DES TYPES DE BRÈCHES SUBIES PAR LES ORGANISATIONS EN 2022 4

FIGURE 5.1 – ÉCRAN DE CONNEXION DE L'OUTIL ANNOY. 68

FIGURE 5.2 – ÉCRAN DE CONNEXION AVEC LES INFORMATIONS D'AUTHENTIFICATION SAISIES. 68

FIGURE 5.3 – PAGE D'ACCUEIL AVEC LE WORKFLOW EN 4 ÉTAPES. 69

FIGURE 5.4 – ZONE DE TÉLÉVERSEMENT DU FICHIER CSV. 70

FIGURE 5.5 – FICHIER CSV SÉLECTIONNÉ PRÊT À ÊTRE TÉLÉVERSÉ. 71

FIGURE 5.6 – PRÉVISUALISATION DES DONNÉES DU FICHIER PATIENTS.CSV. 72

FIGURE 5.7 – PRÉVISUALISATION DÉTAILLÉE DES PREMIÈRES LIGNES DU DATASET.. 72

FIGURE 5.8 – ÉCRAN DE CHARGEMENT PENDANT L'ANALYSE DE DÉTECTION.. 73

FIGURE 5.9 – RÉSULTATS DE LA DÉTECTION AVEC 4 CARTES RÉSUMÉ.. 74

FIGURE 5.10 – TABLEAU DE CLASSIFICATION DÉTAILLÉE (PARTIE 1).. 75

FIGURE 5.11 – TABLEAU DE CLASSIFICATION DÉTAILLÉE (PARTIE 2).. 76

FIGURE 5.12 – TABLEAU DE CLASSIFICATION DÉTAILLÉE (PARTIE 3).. 77

FIGURE 5.13 – TABLEAU DE CLASSIFICATION DÉTAILLÉE (PARTIE 4 - FIN). . . 78

FIGURE 5.14 – CONFIGURATION DE L'ANONYMISATION AVEC TECHNIQUES RECOMMANDÉES. 79

FIGURE 5.15 – CONFIGURATION DE L'ANONYMISATION (SUITE).. 80

FIGURE 5.16 – CONFIGURATION DE L'ANONYMISATION (FIN) AVEC BOUTON ANONYMISER.. 81

FIGURE 5.17 – ÉCRAN DE CHARGEMENT PENDANT L'ANONYMISATION. . . . 82

FIGURE 5.18 – RÉSULTATS FINAUX AVEC VERDICT DE CONFORMITÉ. 83

FIGURE 5.19 – TROIS JAUGES CIRCULAIRES DES CRITÈRES D'ÉVALUATION. . 84

LISTE DES ABRÉVIATIONS

RP	Renseignements personnels
QI	Quasi-identifiant
RGPD	Règlement général sur la protection des données
NIST	National Institute of Standards and Technology
ISO	International Organization for Standardization
HIPAA	Health Insurance Portability and Accountability Act
CSV	Comma-Separated Values
RaaS	Ransomware as a Service
CaaS	Cybercrime as a Service

DÉDICACE

À mon enfant, J'ai commencé ce mémoire alors que tu étais encore en moi. Je l'ai poursuivi après ta naissance, avec toi à mes côtés chaque jour. Tu as grandi en même temps que ce projet. Tu as été ma force silencieuse, ma motivation de tous les jours et la raison pour laquelle je n'ai jamais abandonné. J'ai fait cette maîtrise avec toi, pour toi, malgré toutes les difficultés. Ce travail prouve qu'on peut être mère, faire de la recherche et aller jusqu'au bout.

À ma mère, Sans toi, ce mémoire n'existerait pas. Tu as payé toutes mes études. Tu as cru en moi quand c'était difficile. Tes sacrifices, ta confiance et tes encouragements m'ont donné la force de continuer. Ce travail, c'est aussi le tien. Il représente bien plus qu'un diplôme : c'est la preuve qu'on peut réussir malgré les obstacles, grâce à l'amour et au soutien d'une mère extraordinaire.

REMERCIEMENTS

Je remercie avant tout Dieu pour la force, le courage, la patience et la persévérance qu'il m'a accordées tout au long de ce parcours.

Je tiens à exprimer ma profonde gratitude à mon encadreur, M. Fehmi Jaafar, pour son accompagnement, ses conseils, sa rigueure, sa disponibilité et son soutien tout au long de la réalisation de ce mémoire. Ses orientations et ses remarques ont été précieuses dans l'avancement de ce travail.

Je remercie également chaleureusement M. Schallum Pierre Cofondateur et Directeur des services de gouvernance des données, et du reseau d'expertise en éthique de données, pour ses precieux conseils, ses encouragements et ses précieuses observations. Son apport a contribué à enrichir ce travail et à améliorer plusieurs aspects de ma réflexion.

Je remercie sincèrement les enseignants du programme pour la qualité de leur formation, leurs encouragements et les connaissances qu'ils m'ont transmises, qui ont largement contribué à ma progression académique et professionnelle.

Mes remerciements les plus sincères vont à ma tres chere mère, dont le soutien financier et moral a rendu possible la poursuite de mes études. Sa confiance inébranlable en moi a été une source de motivation constante.

Je remercie également mes collègues pour les échanges enrichissants, le soutien mutuel et l'entraide tout au long de ce parcours.

Enfin, je remercie ces quelques personnes qui m'ont aidé dans l'ombre, dont le soutien discret mais précieux a été important pour moi durant ce parcours.

AVANT-PROPOS

Ce projet de maîtrise m'a été proposé par mon directeur de recherche. Au début, je ne connaissais pas grand-chose à l'anonymisation des données, mais j'ai vite compris pourquoi c'est devenu un sujet important. Depuis mai 2024, le Québec a un nouveau règlement qui impose des règles strictes pour anonymiser les données. Les organisations doivent évaluer trois critères en même temps : la corrélation, l'individualisation et l'inférence. Le problème, c'est que la plupart des outils qui existent sont soit trop compliqués pour les non-spécialistes (comme ARX ou Amnesia), soit trop chers pour les petites structures. Mon travail consistait donc à concevoir un outil accessible, facile à utiliser, qui respecte les exigences du règlement québécois. C'était un défi technique intéressant : traduire des obligations légales en mesures informatiques automatiques, développer une interface intuitive, et permettre aux chercheurs et aux petites organisations de se conformer à la loi sans avoir besoin d'engager des experts. Ce projet m'a permis d'apprendre beaucoup sur la protection des données, la conformité réglementaire et le développement d'outils pratiques pour résoudre des problèmes concrets.

INTRODUCTION

Dans un contexte de transformation numérique accélérée, la collecte et l'utilisation massive des données personnelles sont devenues incontournables pour la recherche, les statistiques et l'innovation organisationnelle. Toutefois, cette exploitation accrue soulève des enjeux majeurs en matière de protection de la vie privée et de conformité réglementaire. Au Québec, la modernisation de la Loi sur la protection des renseignements personnels impose désormais des exigences strictes en matière d'anonymisation des données, fondées notamment sur les critères de corrélation, d'individuation et d'inférence. C'est dans ce cadre que s'inscrit ce mémoire, dont l'objectif est de concevoir un outil d'anonymisation conforme au règlement québécois, permettant d'évaluer les risques de réidentification, d'appliquer des méthodes d'anonymisation adaptées et de soutenir les chercheurs et organisations dans leurs démarches de conformité.

CHAPITRE I

INTRODUCTION

Aujourd'hui, la protection des données personnelles est une étape importante dans le développement d'un système d'information (Hassane & Patrick, 2022). Une caractérisation des solutions à choisir, à construire ou à mettre en œuvre sera également proposée (Kalam *et al.*, 2010).

1.1 CONTEXTE

Les renseignements personnels désignent l'ensemble des informations ou données qui permettent par exemple de repérer ou d'identifier, une personne physique, soit directement (nom, adresse) ou indirectement (date de naissance, téléphone). (Government of Canada, iale). Ces renseignements étant personnels et privés par nature, ils doivent être protéger et leur accès doit être strictement contrôlé pour éviter toute diffusion non autorisée. Dans notre société dominée par la transformation numérique, où les cyberattaques se multiplient, où l'exposition de la vie privée est constante et permanente, où l'explosion des cybermenaces, notamment via des services malveillants comme le Ransomware as a Service (RaaS) ou Cybercrime as a Service (CaaS), qui ont rendu populaire les attaques informatiques en rendant facile l'accès aux outils de piratage (Renaud, 2024). Par exemple, selon Renaud (Renaud, 2024) 18 % des entreprises canadiennes ont subi une cyberattaque en 2021, et le Canada s'est retrouvé en 2022 parmi les trois pays les plus touchés en termes de coûts associés, derrière les États-Unis et le Moyen Orient. En parallèle, on constate qu'en 2021 les organisations canadiennes ont investi plus de 10 milliards de dollars en cybersécurité, mais la sécurité applicative reste encore moins priorisée (Renaud, 2024). Cela se reflète également dans la répartition des dépenses en cybersécurité, où l'on observe une faible part accordée à la sécurité applicative par rapport à d'autres segments (voir Figure 1.1).

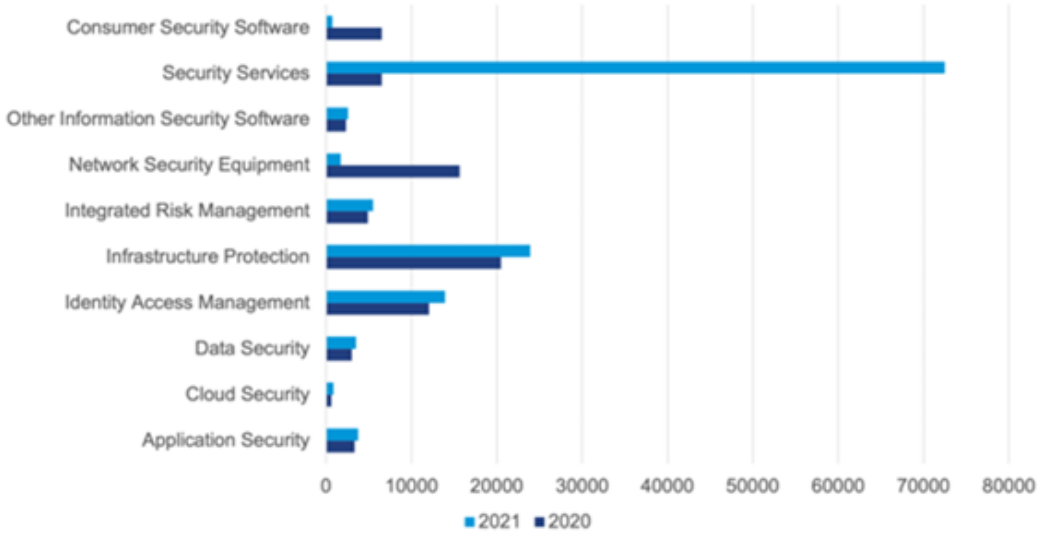

FIGURE 1.1 : Dépenses en cybersécurité par segment en 2020 et 2021.
Source : (Renaud, 2024)

Donc les renseignements personnels jouent aujourd'hui un rôle crucial et fondamental dans le fonctionnement de cette société numérique. Ces renseignements sont devenus un atout stratégique et essentiel. Ils permettent aux institutions, aux entreprises et aux chercheurs d'analyser les comportements, de mieux les comprendre, d'anticiper les besoins puis d'adapter ou d'optimiser leurs services. (Office of the Privacy Commissioner of Canada, 2024).

Par exemple, dans le domaine de la santé, les données permettent de prédire les risques de maladies, de personnaliser les traitements et de renforcer les politiques de santé publique. En finance, elles facilitent l'évaluation des risques de crédit, la détection des fraudes et l'offre de produits personnalisés. (Office of the Privacy Commissioner of Canada, 2024). Cependant, cette exploitation massive des données n'est pas sans risque : elle rend d'autant plus urgente la mise en place de mécanismes fiables pour en assurer la protection, notamment l'anonymisation. Parallèlement, les cybermenaces se multiplient et deviennent de plus en plus dangereuses. Entre 2023 et 2024, 693 brèches ont entraîné la compromission d'environ 25 millions de comptes canadiens, c'est-à-dire que des pirates ont accédé à ces comptes sans autorisation, soit un doublement par rapport à l'année précédente ; de plus, 29 % des organisations ont

signalé une fuite de données en 2022, contre seulement 18 % en 2019 (Office of the Privacy Commissioner of Canada, 2024)(Authority, 2022).

En complément, la figure suivante présente la répartition des différents types de brèches subies par les organisations en 2022 (voir Figure 1.2).

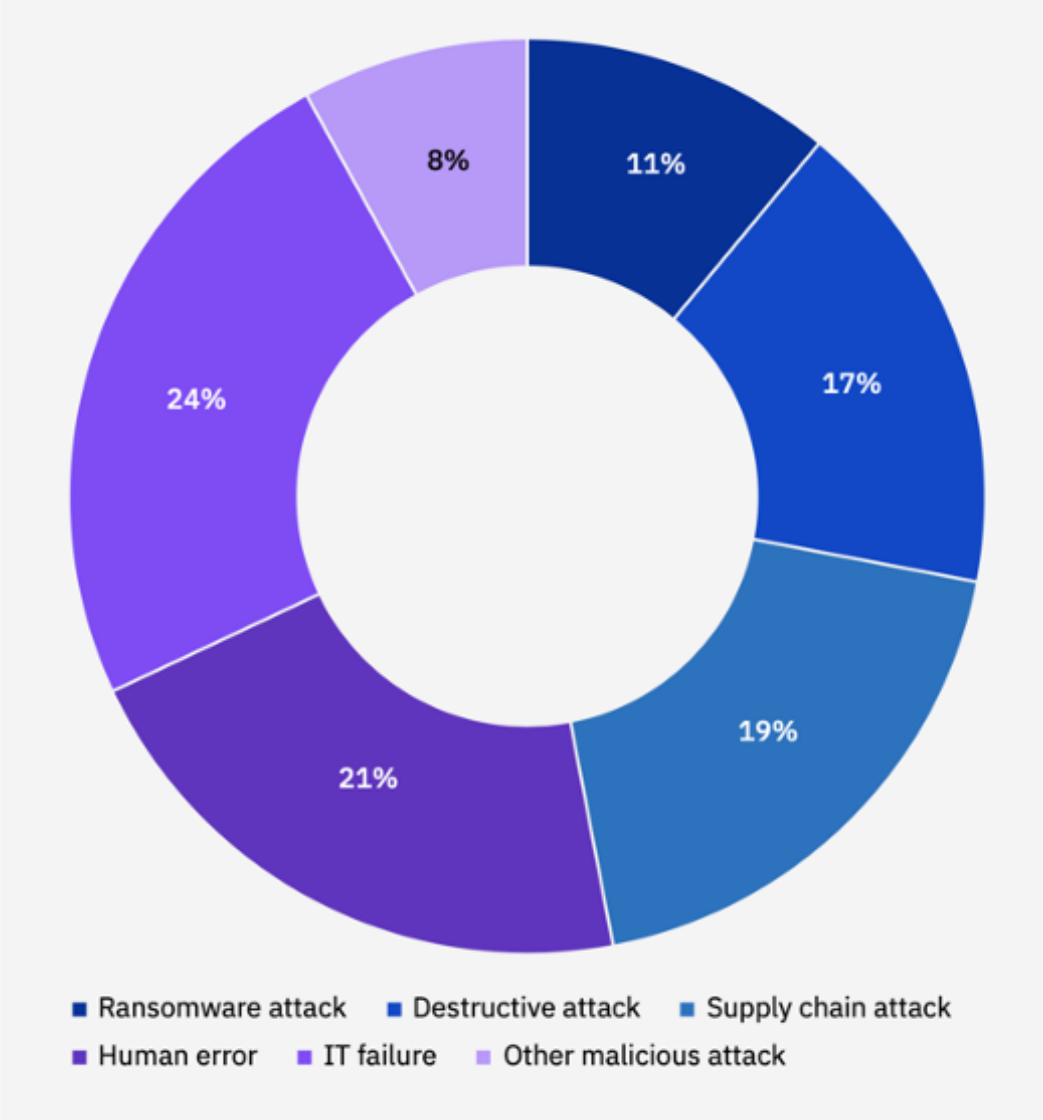

FIGURE 1.2 : Répartition des types de brèches subies par les organisations en 2022

Source : (Renaud, 2024)

Ces chiffres illustrent les défis majeurs en matière de protection de la vie privée et expliquent pourquoi il est crucial de mettre en place des mécanismes d'anonymisation rigoureux, conformes aux reglements quebecois en vigueurs. Pour donc assurer la protection de la vie privée ou des renseignements personnels, le Québec a adopté le Règlement sur l'anonymisation des renseignements personnels, entré en vigueur le 30 mai 2024. Celui-ci définit l'anonymisation comme un processus qui permet de supprimer ou transformer les données, de manière irréversible, afin de garantir que le risque de réidentification soit presque impossible ou très faible. Il repose sur trois critères essentiels :

Corrélation : impossibilité de pouvoir relier plusieurs sources différentes concernant une même personne, c'est-à-dire s'assurer qu'il n'y a aucun moyen qui pourra permettre de lier de nombreuses sources différentes a propos d'une personne ;

Individualisation : incapacité de pouvoir isoler, distinguer ou repérer une personne dans un jeu de données, c'est-à-dire veiller à ce qu'une personne ne puisse pas être indexer dans un ensemble de données ;

Inférence :impossibilité de pouvoir deviner ou déduire des renseignements personnels à partir d'autres éléments disponibles,c'est-à-dire empêcher toute possibilité qui pourrait interpréter les renseignements personnels à partir d'autres données (Gouvernement du Québec, 2024).

De plus, ce règlement impose un processus rigoureux c'est à dire une méthode strictement encadrée : définir l'objectif de l'anonymisation, retirer les identifiants directs, réaliser des analyses de risques avant et après anonymisation, appliquer des techniques reconnues, mettre en place des mesures de sécurité, et tenir un registre à partir du 1er janvier 2025 (Gouvernement du Québec, 2024) ;(McCarthy, 2024). L'anonymisation ne se limite donc pas à supprimer des noms ou des e-mails : elle nécessite une démarche structurée conforme aux exigences du règlement québécois. Ce mémoire propose une application web composée d'une interface Next.js/React et d'un moteur FastAPI/Python, permettant de suivre un processus

guidé depuis l'importation d'un fichier CSV jusqu'au téléchargement du fichier anonymisé accompagné d'un rapport PDF. Le détail de ce processus est présenté dans la section suivante.

1.2 MOTIVATION

Dans le contexte de la transformation numérique et de la multiplication des brèches de données, les organisations sont de plus en plus tenues à anonymiser les renseignements personnels qu'elles possèdent, puis à démontrer que le risque de réidentification reste très faible. Selon le cabinet BLG (Borden Ladner Gervais LLP, 2024), le règlement exige un standard élevé : une analyse valide doit montrer que les données anonymisées ne permettent plus d'identifier une personne, même indirectement. Le risque ne doit pas être nul, mais il doit rester très faible. Pour atteindre un tel niveau de protection, selon Le Règlement sur l'anonymisation des renseignements personnels (A-2.1,r.0.1) (Gouvernement du Québec, 2024) il faut respecter trois critères, critères d'individualisation, de corrélation et d'inférence. Ces critères doivent être respectés pendant tout le long du processus d'anonymisation (Gouvernement du Québec, 2024)(Gouvernement du Québec, s.d). Ces critères reprennent les principes internationalement reconnus en matière d'anonymisation et posent la barre très haut : si ces trois conditions sont satisfaites, on peut considérer que les données ne permettent effectivement plus d'identifier les individus et ne portent plus atteinte à leur vie privée (Fasken, 2024). Le critère de corrélation est particulièrement difficile à garantir, car il exige d'anticiper les renseignements externes « raisonnablement disponibles » pouvant être croisés aux données publiées (Amir, 2024). De plus, le risque résiduel doit être démontré comme « très faible », ce qui nécessite une analyse approfondie, souvent impossible pour les petites organisations qui ne disposent pas des moyens ni des experts (Borden Ladner Gervais LLP, 2024). Or, toutes les organisations, que ça soit les petites entreprises, organisations, les institutions publiques ou les équipes de recherche avec peu de moyens sont concernées par les exigences légales en matière d'anonymisation des données. Pourtant, elles n'ont pas toujours des compétences

techniques ni des outils nécessaires qui pourrait leur permettre de respecter ces règles. Il existe bien des logiciels spécialisés, mais ils ne sont pas toujours adaptés à ces structures. Par exemple, l'outil libre ARX permet de faire différentes opérations comme la généralisation, la suppression ou l'ajout de bruit, et qui contrôle également le niveau de risque de réidentification. Toutefois, il reste difficile à utiliser et s'adresse surtout à des utilisateurs expérimentés, mais aussi ses fonctionnalités très intéressantes sont payantes (Hackread, sd). En revanche, certaines solutions commerciales comme Privitar ou K2View sont plus faciles à utiliser, mais elles coûtent chers et sont conçues principalement pour les grandes entreprises (Hackread, sd). Ce manque d'outils accessibles crée un véritable handicap : d'un côté, les lois exigent des rigueurs sur la protection des renseignements personnels ; de l'autre, les petites structures ou organisations n'ont pas les moyens necessaires de s'y conformer facilement. Le même problème a été observé en France après l'arrivée du RGPD (Règlement Général sur la Protection des Données) : un an après son entrée en vigueur, près de 80 % des petites entreprises françaises n'étaient toujours pas en conformité, car les procedures sont jugées complexes, techniques et stressantes (BLOCKPROOF, sd). Face à cette réalité, il devient essentiel de rendre les outils d'anonymisation plus simples et plus accessibles, pour que toute organisation puisse protéger les données qu'elle partage, même sans être experte en sécurité informatique. Question de recherche : Comment concevoir un outil web d'anonymisation des données qui évalue automatiquement les trois critères du règlement québécois (individuation, corrélation, inférence) et applique des techniques d'anonymisation adaptées tout en restant simple à utiliser par des petites et moyennes organisations ? Ce mémoire propose de relever ce défi en créant un outil web moderne composé d'une interface Next.js/React et d'un moteur FastAPI/Python. Concrètement, l'outil développé permet de suivre un processus guidé complet. D'abord, il permet de charger un fichier CSV et de prévisualiser son contenu afin de vérifier la structure des données avant traitement. Ensuite, il effectue une analyse pré-anonymisation avec détection automatique des données sensibles, catégorisation des colonnes selon quatre types (identifiants directs, quasi-identifiants, données sensibles, données non-sensibles), évaluation du risque

selon les trois critères réglementaires avec leurs pondérations respectives (individualisation, corrélation, inférence), et génération de justifications pour chaque classification. Puis, il permet de configurer les techniques d'anonymisation adaptées à chaque type de données : suppression complète des identifiants directs, généralisation des quasi-identifiants (ex. année au lieu de date complète, FSA au lieu du code postal complet, tranches de revenus), et confidentialité différentielle (mécanisme de Laplace) appliquée exclusivement aux données sensibles numériques. Après cela, l'anonymisation est appliquée, puis une analyse post-anonymisation ré-évalue les trois critères de risque, indique si chaque critère présente un risque élevé ou faible, et détermine la conformité globale avec le règlement québécois ((conforme si score global < 20%, seuil opérationnel retenu pour traduire l'exigence d'un risque très faible, en s'inspirant des pratiques de la CNIL française (Commission Nationale de l'Informatique et des Libertés, 2020))). Enfin, l'outil permet de télécharger le fichier CSV anonymisé ainsi qu'un rapport PDF documentant l'ensemble des transformations appliquées, les scores avant/après, les recommandations et la conformité atteinte.

La mise en œuvre de cet outil pose plusieurs défis : traduire les trois critères en mesures automatisées (individualisation par le k-anonymat, corrélation par la détection de variables liées, inférence par la prédictibilité des attributs sensibles), choisir et calibrer les méthodes d'anonymisation pour équilibrer protection et utilité, puis valider l'outil en simulant des tentatives de réidentification selon un modèle d'adversaire raisonnable (Amir, 2024). Les prochains chapitres détaillent chacun de ces enjeux et la démarche suivie.

1.3 OBJECTIFS

Pour minimiser ces problèmes de protection des renseignements personnels tout en respectant les exigences de la Loi 25 sur l'anonymisation, ce projet a pour but de mettre en œuvre une solution pratique et adaptée. Les objectifs ci-dessous expliquent ce que je veux

atteindre dans ce mémoire, en commençant par l'objectif général et en détaillant ensuite les objectifs spécifiques.

1.3.1 OBJECTIF GÉNÉRAL

Développer un outil interactif, accessible et conforme à la Loi 25 qui évalue automatiquement les trois critères réglementaires avec leurs pondérations respectives (individualisation 40%, corrélation 35%, inférence 25%), ces poids ayant été retenus pour refléter l'importance relative des trois critères, l'individualisation correspondant au risque le plus direct de réidentification (Sweeney, 2002a), la corrélation devenant critique à l'ère du Big Data (National Institute of Standards and Technology, 2015), et l'inférence protégeant contre les attaques par connaissances de fond (Machanavajjhala *et al.*, 2007a), applique des transformations d'anonymisation adaptées comme la suppression, la généralisation et la confidentialité différentielle, et vérifie l'effet avant/après pour déterminer la conformité, le dataset étant considéré comme conforme si le score global est inférieur à 20%, ce seuil opérationnel ayant été retenu pour traduire l'exigence d'un risque très faible, en s'inspirant des pratiques de la CNIL française (Commission Nationale de l'Informatique et des Libertés, 2020).

1.3.2 OBJECTIFS SPÉCIFIQUES

Cadrer et traduire le règlement en métriques opérationnelles : une revue des principales méthodes et des principaux outils d'anonymisation existants, notamment ARX, et d'autres solutions comparables, a d'abord été menée afin d'évaluer leur niveau d'adéquation avec les exigences de la Loi 25. Cette analyse a ensuite servi de base à la définition d'un ensemble de mesures simples, objectives et calculables permettant d'opérationnaliser les trois critères retenus dans le prototype. Le critère d'individualisation, pondéré à 40%, est évalué à partir du k-anonymat minimal appliqué aux quasi-identifiants ainsi que de la proportion d'enregistrements uniques. Le critère de corrélation, pondéré à 35%, repose sur le coefficient

de Pearson, l'information mutuelle et la détection de colonnes susceptibles d'être reliées à des sources externes. Enfin, le critère d'inférence, pondéré à 25%, est mesuré à l'aide de la l-diversité, de l'entropie conditionnelle et de la prédictibilité des attributs sensibles.

Concevoir l'architecture de l'outil : une architecture full-stack moderne a été conçue, combinant un backend API REST avec FastAPI en Python, un frontend interactif avec Next.js/React, ainsi qu'une base de données PostgreSQL destinée à la traçabilité. L'outil repose également sur une architecture modulaire conteneurisée à l'aide de Docker, afin d'assurer sa portabilité, sa reproductibilité et sa facilité de déploiement. Le backend intègre neuf services spécialisés, dédiés notamment à la détection des données sensibles, à l'application des techniques d'anonymisation, à l'évaluation des risques, à la génération de rapports PDF et à la vérification post-anonymisation. Enfin, des mécanismes de traçabilité des opérations, fondés sur une base de données conservant l'historique complet, ont été intégrés, de même que des fonctions d'évaluation des risques résiduels de réidentification conformément aux exigences du règlement québécois.

Implémenter le diagnostic automatique par colonne et par combinaisons : les scores d'individualisation, de corrélation et d'inférence sont calculés avant et après anonymisation, en tenant compte de leurs pondérations respectives (40%/35%/25%). Le système assure également une détection automatique des données sensibles par reconnaissance de motifs (regex) et analyse des noms de colonnes, avec une classification en quatre types : identifiants directs, quasi-identifiants, données sensibles et données non sensibles et catégorisation en 7 : financiers, génétiques ou biométriques, concernant la santé, concernant la vie sexuelle ou l'orientation sexuelle, concernant les convictions religieuses ou philosophiques, concernant les opinions politiques, concernant l'origine ethnique ou raciale (Gouvernement du Québec, 2024). Enfin, des justifications détaillées ainsi que des scores de confiance sont générés pour chaque classification.

Implémenter les transformations d'anonymisation : la suppression complète des identifiants directs, tels que le NAS, l'adresse électronique, le numéro de téléphone, le nom ou le prénom, est réalisée par retrait de colonnes. Les quasi-identifiants font l'objet d'une généralisation, par exemple en conservant uniquement l'année au lieu de la date complète, le FSA au lieu du code postal complet, des tranches de revenus au lieu des montants exacts ou encore des tranches d'âge. La confidentialité différentielle, fondée sur le mécanisme de Laplace avec $\varepsilon = 1.0$, est appliquée exclusivement aux données sensibles numériques, comme les revenus, les soldes bancaires ou d'autres données financières, lorsque le risque d'inférence est élevé. Enfin, des règles de choix automatiques basées sur la catégorie détectée sont mises en place.

Évaluer l'efficacité de l'anonymisation et la perte d'utilité : des jeux de données de test synthétiques, sont utilisés. Une comparaison avant/après est ensuite réalisée pour les trois critères, avec calcul du score global de conformité (seuil $< 20\%$). Enfin, l'utilité des données est évaluée à travers des mesures telles que l'écart sur les moyennes, les variances et la préservation des distributions statistiques.

Concevoir une interface utilisateur moderne et intuitive : Une interface web fondée sur Next.js et React a été développée, avec une navigation guidée par étapes comprenant l'importation du fichier CSV, la prévisualisation des données, la détection automatique, la configuration des techniques, l'anonymisation et l'analyse post-anonymisation. Les résultats de détection sont affichés de manière interactive, avec catégorisation, justifications et scores de confiance. Un tableau comparatif détaillé permet également de visualiser les résultats après traitement, accompagné d'indicateurs visuels des scores de risque. Enfin, l'outil permet l'export du fichier CSV anonymisé ainsi que la génération automatique d'un rapport PDF documentant les transformations appliquées, les scores obtenus et le niveau de conformité.

Valider et documenter l'outil Des tests de bout en bout ont été réalisés afin de valider le workflow complet, depuis le téléversement du fichier jusqu'à la détection, l'anonymisation, l'évaluation et le téléchargement des résultats. Des tests unitaires ont également été menés sur les services d'anonymisation et d'évaluation des risques. Par ailleurs ce travail a permis d'identifier plusieurs limites connues, notamment l'existence d'un risque résiduel incompressible et le compromis entre utilité et confidentialité, tout en formulant des recommandations d'usage. Enfin, un guide utilisateur ainsi qu'une documentation technique réutilisable ont été rédigés pour faciliter l'adaptation de l'outil à d'autres projets nécessitant une conformité à la Loi25.

1.4 CONTEXTE DE RECHERCHE

Ce mémoire se situe dans le domaine de l'anonymisation des renseignements personnels selon la Loi 25 du Québec. Sur le plan académique, des approches comme la k-anonymité, la l-diversité, la t-closeness, le masquage ou la généralisation sont bien établies (Samarati & Sweeney, 1998) ; (Machanavajjhala *et al.*, 2007b), mais leur efficacité dépend du contexte et du cadre légal visé. Les travaux récents s'attachent à opérationnaliser ces méthodes pour qu'elles soient vérifiables face aux exigences du reglement. C'est dans ce contexte que ce mémoire propose un outil aligné sur les exigences du reglement québécois.

1.5 CONTRIBUTIONS

Ce mémoire apporte des contributions théoriques et pratiques à l'anonymisation des renseignements personnels dans le contexte québécois :

Un outil interactif et accessible : un prototype a été proposé et conçu pour des utilisateurs non spécialistes, notamment les petites organisations et les chercheurs. Contrairement à des solutions existantes puissantes mais plus techniques, telles que *ARX* ou *Amnesia*, qui

nécessitent une prise en main avancée et un paramétrage complexe, l'outil développé privilégie une interface simple, des choix guidés et une compréhension immédiate des risques.

Opérationnalisation de la Loi 25 : l'outil opérationnalise les trois critères du règlement québécois, à savoir la corrélation, l'individualisation et l'inférence, en mesures quantitatives calculables, à la fois par colonne et au niveau global du jeu de données. Il intègre également un moteur de recommandations proposant des transformations adaptées, telles que la suppression, la généralisation et la confidentialité différentielle, accompagné d'un rapport pdf afin de faciliter la traçabilité et la démonstration. Enfin, il met en œuvre une simulation de réidentification fondée sur un modèle d'« adversaire raisonnable », permettant d'évaluer le risque résiduel après anonymisation.

Un guide pédagogique pour l'adoption : une documentation claire et structurée a été rédigée, illustrée par des exemples et des bonnes pratiques, afin de décrire pas à pas l'évaluation des critères, le choix des transformations, la vérification post-anonymisation et la tenue du rapport. Cet accompagnement vise à faciliter l'appropriation du prototype dans des contextes variés, tels que la démonstration, la preuve de concept ou l'utilisation par de petites équipes.

1.6 ORGANISATION DU DOCUMENT

Ce mémoire est structuré en six chapitres, chacun abordant un aspect spécifique du projet.

Chapitre 1 – Introduction : e chapitre présente le contexte général du travail, la problématique étudiée, la motivation de la recherche, les objectifs poursuivis, les contributions attendues ainsi que l'organisation globale du document.

Chapitre 2 – État de l'art : ce chapitre propose une revue des travaux existants portant sur les méthodes et outils d'anonymisation des données. Il met en évidence leurs principes, leurs forces et leurs limites, ainsi que leur conformité aux exigences du cadre réglementaire québécois et canadien.

Chapitre 3 – Cadre théorique et réglementaire : ce chapitre décrit les fondements théoriques de l'anonymisation des données et détaille les critères définis par le Règlement québécois sur l'anonymisation des renseignements personnels, à savoir la corrélation, l'individualisation et l'inférence, ainsi que les pratiques associées à la Loi 25.

Chapitre 4 – Conception et méthodologie : ce chapitre présente la démarche méthodologique adoptée pour la conception de l'outil d'anonymisation. Il décrit l'architecture technique retenue, les choix méthodologiques effectués, les technologies utilisées et le processus d'évaluation du risque de réidentification.

Chapitre 5 – Implémentation et résultats : ce chapitre décrit l'implémentation du prototype développé en Python. Il présente les tests réalisés, l'évaluation du fonctionnement de l'outil ainsi que les résultats obtenus.

Chapitre 6 – Conclusion et perspectives : ce dernier chapitre résume les principales contributions du mémoire, discute des limites rencontrées et propose des pistes d'amélioration ainsi que des perspectives de recherche future.

CHAPITRE II

ÉTAT DE L'ART

Ce chapitre présente l'état de l'art sur l'anonymisation des renseignements personnels, en s'appuyant à la fois sur les travaux académiques et sur les cadres réglementaires récents (Gouvernement du Québec, 2024) ; (Office of the Privacy Commissioner of Canada, 2024). Il pose les bases conceptuelles nécessaires en définissant les notions clés, les principaux modèles théoriques de protection de la vie privée, les techniques courantes d'anonymisation et les risques de réidentification (Samarati & Sweeney, 1998) ; (Dwork, 2006). Il examine ensuite les outils existants, leurs forces et leurs limites, et situe le contexte québécois par rapport aux autres juridictions, notamment l'Europe et les États-Unis (European Data Protection Board, 2020) ; (National Institute of Standards and Technology, 2018). Une synthèse finale met en lumière les besoins non comblés dans le domaine.

2.1 NOTIONS CLÉS

Les renseignements personnels (RP) désignent toute information qui concerne une personne et qui permet de l'identifier directement ou indirectement, soit à travers un identifiant direct ou à travers une combinaison d'attributs. On distingue ainsi les identifiants directs (nom, prénom, numéro de compte, NAS), les quasi-identifiants (QI) comme la date de naissance, la ville ou le code postal, et les données sensibles telles que la santé, l'assurance ou les revenus (Gouvernement du Québec, 2024) ; (Office of the Privacy Commissioner of Canada, 2024).

Dans la littérature, on distingue également la pseudonymisation, qui remplace l'identifiant par un autre identifiant technique ; cette méthode est utile mais réversible si la clé existe ailleurs, contrairement à l'anonymisation qui vise une réduction du risque à un niveau très faible (European Data Protection Board, 2020).

La Loi 25 et son règlement d'anonymisation en vigueur depuis mai 2024 exigent l'évaluation de trois critères : l'individualisation (éviter les enregistrements quasi-uniques), la corrélation (éviter des combinaisons qui peuvent révéler l'identité d'un individu) et l'inférence (éviter qu'on devine un attribut sensible à partir d'autres colonnes) (Gouvernement du Québec, 2024) ; (McCarthy, 2024) ; (Torys LLP, 2024). Depuis le 1er janvier 2025, un registre des démarches d'anonymisation est exigé (Borden Ladner Gervais LLP, 2024) ; (McCarthy, 2024).

2.2 MODÈLES THÉORIQUES DE RÉFÉRENCE

Le k-anonymat impose que chaque enregistrement soit indiscernable parmi au moins k individus partageant les mêmes quasi-identifiants. Cette méthode est efficace contre l'individualisation, mais insuffisante contre certaines inférences : si les personnes d'un groupe partagent la même maladie, on peut déduire cette information sans identifier directement quelqu'un (Sweeney, 2002b) ; (Samarati & Sweeney, 1998).

La l-diversité complète cette approche en exigant qu'il y ait au moins l valeurs distinctes pour l'attribut sensible dans chaque groupe, limitant ainsi l'inférence triviale (Machanavajjhala *et al.*, 2007b).

La t-closeness va plus loin en exigeant que la distribution de l'attribut sensible dans chaque groupe reste proche de la distribution globale, réduisant la fuite d'information statistique (Li & Li, 2007).

La confidentialité différentielle (DP) propose l'ajout de bruit calibré fournissant une garantie probable sur l'influence d'un individu dans la sortie ; elle est particulièrement pertinente pour limiter l'inférence sur des attributs sensibles (Dwork, 2006) ; (Dwork & Roth, 2017).

En pratique, ces méthodes sont complémentaires : chacune règle un problème mais présente ses propres limites. Le k-anonymat ne protège pas contre l'inférence, la l-diversité ne considère pas les distributions, la t-closeness est difficile à implémenter et la confidentialité différentielle est difficile à paramétrer. Il faut donc les combiner pour répondre aux trois critères de la Loi 25.

2.3 TECHNIQUES D'ANONYMISATION

L'anonymisation consiste à transformer les données pour réduire le risque de réidentification tout en gardant leur utilité pour l'analyse.les techniques : (Hundepool *et al.*, 2012).

2.3.1 SUPPRESSION DES IDENTIFIANTS DIRECTS

La suppression retire complètement les colonnes contenant des identifiants directs comme le nom, le prénom, le NAS, l'adresse courriel ou le numéro de téléphone. C'est la technique la plus simple et irréversible : elle élimine totalement le risque d'identification directe, mais l'information est perdue définitivement (El Emam & Dankar, 2012). Elle est recommandée pour tous les identifiants directs dont on n'a pas besoin pour l'analyse.

2.3.2 PSEUDONYMISATION

La pseudonymisation remplace les identifiants directs par des codes ou des pseudonymes (par exemple « Jean Dupont » devient « PATIENT_001234 »). Elle permet de relier les données d'une même personne tout en cachant son identité réelle. Cependant, elle reste réversible si quelqu'un a accès à la table de correspondance, ce qui signifie qu'elle ne satisfait pas le critère d'irréversibilité exigé par le règlement québécois (European Data Protection Board, 2020).

2.3.3 MASQUAGE PARTIEL

Le masquage partiel cache une partie de l'information en la remplaçant par des astérisques, tout en gardant une portion visible. Par exemple, `jean.dupont@example.com` devient `je**@ex**.com` et `514-123-4567` devient `514-***-****`. Cette technique offre un bon équilibre entre protection et utilité : elle garde des informations partielles utiles comme l'indicatif régional, mais la protection reste partielle et elle risque d'être compromise si combinée avec d'autres informations.

2.3.4 GÉNÉRALISATION ET AGRÉGATION

La généralisation remplace des valeurs précises par des catégories plus larges. Par exemple, une date de naissance `15 mars 1985` devient `1985`, un code postal `H3C 2A7` devient `H3C`, ou un âge de `42 ans` se transforme en `40-49 ans` (Samarati & Sweeney, 1998). Elle garde les tendances statistiques importantes tout en réduisant le risque d'individualisation et de corrélation, mais elle entraîne une perte de précision. Le choix de la taille des tranches est crucial : trop larges, les données perdent leur utilité ; trop étroites, le risque reste élevé.

2.3.5 CONFIDENTIALITÉ DIFFÉRENTIELLE

La confidentialité différentielle ajoute du bruit aléatoire calculé mathématiquement aux données numériques, rendant difficile de déterminer si une personne spécifique fait partie du jeu de données (Dwork, 2006). Le niveau de bruit est contrôlé par le paramètre epsilon (ε) : plus petit, plus protecteur, mais plus de bruit. Les deux mécanismes principaux sont le mécanisme de Laplace (pour une protection stricte) et le mécanisme gaussien (pour une protection moins stricte). La formule pour le bruit de Laplace est :

$$\text{Bruit} \sim \text{Laplace}\left(0, \frac{\Delta f}{\varepsilon}\right) \tag{2.1}$$

où Δf représente l'impact maximum qu'une personne peut avoir sur le résultat. Cette technique offre une garantie mathématique solide, mais le choix de epsilon reste un défi en pratique : le recensement américain utilise $\varepsilon = 1.0$ (National Institute of Standards and Technology, 2018), mais une valeur inappropriée peut rendre les données soit inexploitables (ε trop petit), soit insuffisamment protégées (ε trop grand).

2.3.6 LE COMPROMIS ENTRE PROTECTION ET UTILITÉ

Un des grands défis de l'anonymisation, c'est de trouver le bon équilibre entre protéger la vie privée et garder des données utiles (Gouvernement du Québec, 2024). Ce compromis peut être mesuré à travers trois dimensions : l'utilité syntaxique (les moyennes et corrélations restent-elles similaires ?), l'utilité sémantique (un modèle entraîné sur ces données donne-t-il des résultats comparables ?) et l'utilité par cas d'usage (peut-on toujours répondre aux questions analytiques visées ?). Dans la pratique, les techniques fonctionnent mieux combinées : par exemple, la suppression des identifiants directs avec la généralisation des quasi-identifiants réduit l'individualisation et la corrélation, tandis que la généralisation combinée avec la confidentialité différentielle attaque aussi l'inférence (Hundepool *et al.*, 2012) ; (Domingo-Ferrer & Soria-Comas, 2016).

2.4 CADRE LÉGAL INTERNATIONAL ET COMPARAISONS

2.4.1 LE RGPD EN EUROPE

En Europe, le Règlement général sur la protection des données (RGPD) distingue clairement l'anonymisation (irréversible) de la pseudonymisation (réversible) et insiste sur la notion de « risque raisonnable » de réidentification (Parlement européen & Conseil de l'Union européenne, 2016) ; (European Data Protection Board, 2020). Il n'impose pas de

méthode spécifique d'anonymisation, mais exige une évaluation des risques contextualisée. Cette approche rejoint celle du règlement québécois, qui définit également trois critères d'évaluation sans prescrire de techniques obligatoires.

2.4.2 HIPAA ET NIST AUX ÉTATS-UNIS

Aux États-Unis, le Health Insurance Portability and Accountability Act (HIPAA) propose deux méthodes principales : le **Safe Harbor** (suppression de 18 identifiants spécifiques comme le nom, l'adresse, les dates) et l'**Expert Determination** (validation par un expert que le risque de réidentification est très faible). Le NIST publie parallèlement des lignes directrices sur la gestion du risque de réidentification et l'utilisation de méthodes comme le k-anonymat ou la confidentialité différentielle (Office for Civil Rights, 2012) ; (National Institute of Standards and Technology, 2018). Ces documents techniques offrent des recommandations pratiques qui ont influencé l'approche québécoise.

2.4.3 NORMES INTERNATIONALES ISO

La norme ISO/IEC 20889 :2018 définit un vocabulaire commun et une classification des techniques d'anonymisation reconnues internationalement, couvrant les différents types d'identifiants (directs, quasi-identifiants), les techniques de transformation (suppression, généralisation, permutation, ajout de bruit) et les critères d'évaluation du risque de réidentification (International Organization for Standardization, 2018). Cette standardisation facilite les échanges entre juridictions et permet une meilleure comparabilité des approches.

2.4.4 APPORTS SPÉCIFIQUES DU RÈGLEMENT QUÉBÉCOIS

Le règlement québécois s'inspire de ces cadres internationaux tout en apportant des précisions importantes. Il impose l'évaluation de trois critères (individualisation, corrélation,

inférence), la tenue d'un registre obligatoire depuis janvier 2025, et la réalisation d'une analyse des risques avant et après anonymisation. Toutefois, le règlement ne définit pas de seuils quantitatifs précis pour l'évaluation du risque de réidentification : il exige que le risque résiduel soit « très faible » sans fournir de valeur numérique exacte. Ces critères s'inscrivent dans la continuité des travaux internationaux, notamment l'Avis 05/2014 du Groupe de l'Article 29, tout en précisant davantage les attentes opérationnelles pour les organisations québécoises (Gouvernement du Québec, 2024);(Commission d'accès à l'information du Québec, 2023).

2.5 RISQUES DE RÉIDENTIFICATION : CATÉGORIES ET MESURES

2.5.1 RISQUE DE LIAISON

Le risque de liaison correspond à la possibilité de relier un enregistrement anonymisé à une personne spécifique en croisant le jeu de données avec des sources externes accessibles (El Emam & Dankar, 2012) ; (Hundepool *et al.*, 2012). Par exemple, une femme de 42 ans résidant dans le code postal H3C 2A7 pourrait être identifiée en croisant avec une liste électorale publique, si cette combinaison de quasi-identifiants est suffisamment rare. Ce risque correspond directement au critère de **corrélation** du règlement québécois .

2.5.2 RISQUE D'ATTRIBUTION

Le risque d'attribution renvoie à la possibilité d'associer un attribut sensible à une personne identifiée ou identifiable, même sans connaître son identité exacte. Ce risque survient lorsque tous les membres d'un groupe partagent la même valeur sensible : si les 5 personnes d'un groupe k-anonyme ont tous le même diagnostic médical « diabète de type 2 », on peut déduire cette information même sans identifier individuellement qui est qui. Ce risque est

atténué par la l-diversité, qui exige au moins l valeurs distinctes pour les attributs sensibles dans chaque groupe.

2.5.3 RISQUE D'INFÉRENCE

Le risque d'inférence désigne la capacité à deviner ou prédire une information sensible à partir d'autres variables corrélées, même si l'identité exacte de la personne n'est pas connue. Par exemple, si le revenu est fortement corrélé avec le code postal et le niveau d'éducation (corrélation elevee, par exemple > 0.8), un adversaire peut prédire le revenu d'une personne même s'il a été supprimé ou généralisé. Ce risque correspond au critère d'**inférence** du règlement québécois .

2.5.4 INDICATEURS QUANTITATIFS DE MESURE DES RISQUES

Pour mesurer ces risques objectivement, plusieurs indicateurs quantitatifs sont proposés dans la littérature (Skinner & Shlomo, 2008) ; (El Emam & Dankar, 2012).

RISQUE INDIVIDUEL

Le risque individuel correspond à la probabilité de réidentifier un enregistrement donné. Pour un enregistrement avec des quasi-identifiants QI, il peut être estimé par :

$$R_{\text{indiv}} = \frac{1}{F_k} \tag{2.2}$$

où F_k est le nombre d'individus partageant les mêmes valeurs de quasi-identifiants dans la population. Par exemple, si 10 personnes partagent la même combinaison (âge, sexe, code postal), le risque individuel est de 1/10 = 10%.

UNICITÉ DE LA POPULATION

L'unicité de la population estime la proportion d'enregistrements uniques ou quasi-uniques pour certaines combinaisons de quasi-identifiants :

$$U = \frac{\text{Nombre d'enregistrements uniques}}{\text{Nombre total d'enregistrements}} \times 100\% \tag{2.3}$$

Cet indicateur est directement lié au critère d'**individualisation** du règlement québécois. Un taux d'unicité élevé indique un risque important de réidentification.

ENTROPIE ET DIVERSITÉ

L'entropie de Shannon mesure la diversité des valeurs d'un attribut sensible au sein d'un groupe k-anonyme :

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2.4}$$

où p_i est la proportion de la valeur i dans le groupe. Une entropie faible (< 2 bits) indique une faible diversité et un risque d'attribution élevé. Par exemple, si 4 personnes sur 5 dans un groupe ont la même maladie, l'entropie sera faible.

PRÉDICTIBILITÉ

La prédictibilité d'un attribut sensible peut être évaluée en entraînant un modèle de classification ou de régression pour prédire cet attribut à partir des autres colonnes. Si un modèle simple (arbre de décision, régression logistique) atteint une précision élevée (par exemple $> 80\%$), cela indique un grand risque d'inférence, même si la colonne cible a été supprimée.

LIEN AVEC LES CRITÈRES DU RÈGLEMENT QUÉBÉCOIS

Ces indicateurs permettent d'opérationnaliser les trois critères du règlement québécois : l'individualisation est mesurée par l'unicité de la population et le k-anonymat minimal, la corrélation par la détection de colonnes liables à des sources externes et les corrélations fortes entre variables (coefficient de Pearson > 0.7), et l'inférence par l'entropie des attributs sensibles et la performance des modèles prédictifs. (Gouvernement du Québec, 2024) ; (Amir, 2024).

2.6 OUTILS EXISTANTS D'ANONYMISATION DES DONNÉES

Plusieurs outils logiciels ont été développés pour mettre en pratique les modèles théoriques d'anonymisation. Les principaux sont présentés ci-dessous avec leurs forces et limites.

2.6.1 ARX

ARX est probablement l'outil le plus cité dans la littérature scientifique (Prasser *et al.*, 2014). Il supporte plusieurs modèles théoriques (k-anonymat, l-diversité, t-closeness), calcule des indicateurs de risque détaillés et est open source avec une bonne documentation scientifique. Cependant, il reste assez complexe : il faut comprendre les hiérarchies de généralisation, paramétrer les modèles correctement, et sa courbe d'apprentissage est importante pour les utilisateurs non techniques. Il n'est pas conçu spécifiquement pour la Loi 25 et ne calcule pas directement les trois critères avec leurs pondérations.

2.6.2 AMNESIA

Amnesia, développé par l'Université d'Athènes, propose une approche plus accessible avec une interface plus simple et une visualisation claire des hiérarchies de généralisation

(Alexandros & Al., 2018). Il reste bien adapté à l'apprentissage des concepts d'anonymisation, mais ne prend pas en compte explicitement le risque d'inférence, n'intègre pas la confidentialité différentielle et n'offre pas de génération de rapports conformes aux exigences québécoises.

2.6.3 SOLUTIONS COMMERCIALES

Des produits commerciaux comme Privitar, K2View ou les suites proposées par IBM offrent des fonctionnalités très complètes d'anonymisation, de gouvernance des données et de suivi des transformations, avec un support technique professionnel et une interface moderne (Privitar, 2022) ; (K2View, 2022) ; (IBM, 2021). Elles sont cependant très coûteuses (plusieurs dizaines de milliers de dollars par année), conçues pour de grandes organisations, et ne sont pas adaptées spécifiquement au règlement québécois : elles suivent plutôt le RGPD européen ou HIPAA américain.

Le tableau suivant synthétise cette comparaison :

TABLEAU 2.1 : Comparaison des outils d'anonymisation

Critère	**ARX**	**Amnesia**	**Solutions commerciales**	**Annoy(mon outil)**
Coût	Gratuit	Gratuit	50 000 $+/an	Accessible
Complexité	Élevée	Moyenne	Moyenne	Faible
3 critères québécois	Non	Non	Non	Oui
Confidentialité différentielle	Oui	Non	Partiel	Oui
Rapport registre	Non	Non	Partiel	Oui
Interface française	Non	Non	Non	Oui

Aucun de ces outils ne répond complètement aux exigences du règlement québécois : évaluation simultanée des trois critères, détection automatique des données sensibles, analyse

avant/après et génération de rapports conformes au registre obligatoire (Commission d'accès à l'information du Québec, 2023).

2.7 LIMITES DES APPROCHES ACTUELLES

Même si les méthodes d'anonymisation et les outils existants offrent une base solide, plusieurs limites importantes sont identifiées dans la littérature (Domingo-Ferrer & Soria-Comas, 2016) ; (El Emam & Dankar, 2012).

2.7.1 LES MODÈLES CLASSIQUES NE COUVRENT PAS TOUS LES RISQUES

Les modèles théoriques comme le k-anonymat, la l-diversité et la t-closeness sont utiles et protègent contre certains types d'attaques, mais ils ne couvrent pas toutes les situations possibles. Le k-anonymat protège contre l'individualisation mais pas contre l'inférence : un groupe de 5 personnes partageant les mêmes quasi-identifiants et le même diagnostic médical « cancer du poumon » compromettra la confidentialité même si le k-anonymat est respecté (Machanavajjhala *et al.*, 2007b). La l-diversité corrige en partie ce problème, mais ne considère pas les distributions. La t-closeness est plus complète mais très complexe à implémenter. Chaque modèle règle un problème mais en crée d'autres : il faut donc combiner plusieurs approches.

2.7.2 LE COMPROMIS UTILITÉ-PROTECTION

La généralisation et la suppression causent une perte importante d'utilité des données. Si les âges sont transformés en tranches de 20 ans ou les revenus en deux catégories, les données deviennent trop imprécises pour beaucoup d'analyses statistiques ou de projets de recherche (Hundepool *et al.*, 2012) ; (Domingo-Ferrer & Soria-Comas, 2016). Il faut donc trouver le bon équilibre : trop de protection rend les données inutiles, pas assez laisse un risque

de réidentification trop élevé. Les outils existants ne donnent pas vraiment de guide clair pour placer ce curseur.

2.7.3 CONFIDENTIALITÉ DIFFÉRENTIELLE : DIFFICILE À PARAMÉTRER

La confidentialité différentielle reste difficile à utiliser en pratique. Le paramètre epsilon n'est pas intuitif : $\varepsilon = 0.1$ est très protecteur mais bruiteux (les moyennes peuvent être faussées de plusieurs dizaines de pourcent), $\varepsilon = 1.0$ = est modéré comme le censo américain, $\varepsilon = 10$ est faible. Une mauvaise configuration peut rendre les données soit inexploitables (ε trop petit), soit insuffisamment protégées (ε trop grand). Il n'y a pas de « bon » epsilon universel : ça dépend du contexte, de la sensibilité des données et du nombre de requêtes (Dwork, 2006) ; (Dwork & Roth, 2017).

2.7.4 AUCUN OUTIL NE COUVRE LES TROIS CRITÈRES DE LA LOI 25 DU RÈGLEMENT QUÉBÉCOIS

La limite la plus importante reste que très peu de solutions prennent en compte les trois critères définis par le règlement québécois de manière intégrée. ARX se concentre surtout sur l'individualisation (k-anonymat, l-diversité), la plupart des outils ne calculent pas explicitement le risque de corrélation avec des sources externes, et très peu mesurent vraiment le risque d'inférence. Aucun outil n'évalue automatiquement les trois critères simultanément ni ne vérifie la conformité avec l'exigence de risque « très faible » du règlement québécois. Dans la pratique, les organisations québécoises se retrouvent donc contraintes : utiliser ARX avec des calculs manuels supplémentaires, acheter une solution commerciale très chère à adapter, ou ne rien faire au risque de non-conformité (Gouvernement du Québec, 2024) ;(Amir, 2024).

2.8 POSITION DU QUÉBEC DANS LE PAYSAGE INTERNATIONAL

2.8.1 CE QUI REND LE RÈGLEMENT QUÉBÉCOIS UNIQUE

Le Québec se démarque par son niveau d'exigence particulièrement élevé en matière d'anonymisation. Le règlement québécois introduit plusieurs éléments qui n'existent pas ailleurs ou qui ne sont pas aussi précis (Gouvernement du Québec, 2024). Contrairement au RGPD en Europe ou à HIPAA aux États-Unis qui restent assez généraux, le règlement québécois définit explicitement les trois critères (individualisation, corrélation, inférence) et exige qu'ils soient tous évalués simultanément. De plus, il impose une démarche documentée en six étapes et la tenue d'un registre obligatoire. Aucun autre cadre au monde n'est aussi précis : le RGPD dit juste « le risque doit être faible » sans préciser comment le mesurer ni comment combiner différents types de risques.

2.8.2 UNE DÉMARCHE STRUCTURÉE OBLIGATOIRE

Le règlement québécois exige une démarche en plusieurs étapes bien définies : préciser l'objectif poursuivi par l'anonymisation, retirer les identifiants directs, évaluer le risque **avant** traitement, appliquer les techniques d'anonymisation, évaluer le risque **après** traitement, documenter toutes les décisions prises et tenir un registre des démarches réalisées (obligatoire depuis janvier 2025) (Commission d'accès à l'information du Québec, 2023). Cette approche est beaucoup plus rigoureuse que la méthode Safe Harbor de HIPAA qui dit simplement « supprimez ces 18 identifiants » sans demander d'analyse de risque avant/après.

2.8.3 COMPARAISON AVEC LES AUTRES CADRES

Le tableau suivant illustre ces différences :

TABLEAU 2.2 : Comparaison détaillée des exigences d'anonymisation

Aspect	**RGPD (Europe)**	**HIPAA (USA)**	**Loi 25 (Québec)**
Critères	« Risque raisonnable » (vague)	18 identifiants ou expert	3 critères explicites
Quantification	Non	Non	Oui
Analyse avant/après	Recommandée	Non requise (Safe Harbor)	**Obligatoire**
Documentation	Recommandée	Validation experte	**Registre obligatoire**
Seuil explicite	Non	Non	Non Explicite
Techniques prescrites	Non	Oui (Safe Harbor)	Non (flexibilité)

2.8.4 LE QUÉBEC COMME LABORATOIRE INTERNATIONAL

Le Québec combine ainsi la rigueur de HIPAA avec la flexibilité du RGPD : des critères explicitement définis qui doivent être évalués simultanément, une démarche structurée en six étapes, et un registre obligatoire, sans prescrire de techniques imposées. Cette avancée fait du Québec un cas vraiment intéressant à étudier en matière de protection de la vie privée : c'est un des premiers cadres au monde à rendre opérationnels les trois critères d'évaluation (on peut vraiment l'implémenter dans un logiciel, contrairement au RGPD qui reste très conceptuel), tout en étant rigoureux mais flexible.

2.8.5 LE DÉFI POUR LES ORGANISATIONS QUÉBÉCOISES

Cette innovation représente cependant aussi un défi important pour les organisations québécoises, en particulier les plus petites. Une PME de 50 employés n'a généralement pas d'expert en anonymisation dans son équipe, les solutions commerciales conformes coûtent des dizaines de milliers de dollars par année, ce qui est inaccessible pour une école, une clinique médicale ou une municipalité. Depuis janvier 2025, le registre obligatoire demande du temps et de l'organisation sans outil approprié. Beaucoup d'organisations pourraient donc se retrouver en non-conformité, non pas par mauvaise volonté, mais simplement parce qu'elles n'ont pas

les moyens techniques et humains de respecter ces exigences (Fasken, 2024) ; (Borden Ladner Gervais LLP, 2024).

2.9 SYNTHÈSE DE L'ÉTAT DE L'ART

2.9.1 CE QUE RÉVÈLE LA REVUE DE LITTÉRATURE

Les concepts fondamentaux d'anonymisation sont maintenant bien définis et partagés entre les cadres juridiques internationaux (RGPD en Europe, HIPAA aux États-Unis, Loi 25 au Québec). Les modèles théoriques (k-anonymat, l-diversité, t-closeness, confidentialité différentielle) constituent une base solide, mais chacun présente ses limites et ils doivent être combinés. Les techniques courantes sont bien documentées, mais leur application exige un équilibre entre protection et utilité qui nécessite des indicateurs adaptés (Hundepool *et al.*, 2012) ; (Domingo-Ferrer & Soria-Comas, 2016). Les risques de réidentification prennent plusieurs formes (liaison, attribution, inférence) et il existe des indicateurs spécifiques pour les mesurer correctement (El Emam & Dankar, 2012) ; (Skinner & Shlomo, 2008).

2.9.2 LE BESOIN IDENTIFIÉ

Du côté des outils, aucune solution existante ne répond complètement aux besoins québécois : les outils académiques comme ARX sont puissants mais complexes, les solutions commerciales sont complètes mais coûteuses et non adaptées à la Loi 25, et aucun outil ne respecte les trois critères québécois ni ne génère automatiquement les rapports requis pour le registre obligatoire (Prasser *et al.*, 2014) ; (Alexandros & Al., 2018) ; (Privitar, 2022). Il existe donc un besoin pour une solution qui soit à la fois simple d'utilisation, accessible financièrement, conforme au règlement québécois, pratique et traçable, en particulier pour les PME québécoises, les équipes de recherche universitaire et les petites municipalités qui doivent se conformer à la Loi 25 sans avoir les ressources d'une grande entreprise.

2.9.3 LA SUITE DU MÉMOIRE

Le chapitre suivant présente le cadre théorique et réglementaire détaillé nécessaire pour opérationnaliser ces besoins. Il explique comment les trois critères du règlement québécois peuvent être traduits en indicateurs quantitatifs calculables automatiquement, et comment un processus d'anonymisation peut être conçu pour respecter toutes les exigences tout en restant accessible aux utilisateurs non experts.

CHAPITRE III
CADRE THÉORIQUE ET RÉGLEMENTAIRE

Dans le chapitre précédent, le tour de ce qui existe déjà en matière d'anonymisation a été fait : les modèles théoriques, les outils disponibles, les différents cadres réglementaires internationaux. J'ai aussi identifié le problème principal : aucun outil existant ne répond vraiment aux besoins spécifiques du règlement québécois.

Maintenant, avant de plonger dans la conception de cet outil, il faut établir les cadres théoriques et réglementaires sur lesquelles il va s'appuyer. C'est l'objectif de ce chapitre.

Plus concrètement, ce chapitre va :

Ce chapitre a pour objectif de définir clairement les concepts clés nécessaires à la compréhension du mémoire, notamment les notions d'identifiant direct, de quasi-identifiant, d'inférence et de risque de réidentification. Il vise également à expliquer en détail les exigences du règlement québécois sur l'anonymisation, en précisant les trois critères qui sont l'individualisation, la corrélation et l'inférence, ainsi que la signification concrète de l'exigence de risque très faible. Dans cette perspective, le chapitre montre comment ces exigences réglementaires peuvent être traduites en indicateurs quantitatifs calculables automatiquement par un logiciel. Il présente en outre les principales techniques d'anonymisation retenues dans ce travail, notamment la suppression, la généralisation et la confidentialité différentielle, en justifiant leur choix au regard des objectifs poursuivis. Enfin, il établit le lien entre les fondements théoriques issus de la littérature, comme le k-anonymat ou la l-diversité, et leur mise en relation avec le cadre réglementaire québécois.

En d'autres termes, ce chapitre répond à la question suivante : sur quelles bases théoriques et réglementaires repose la conception de l'outil proposé ? Il constitue ainsi le lien entre

la revue de littérature présentée au chapitre précédent et la méthodologie concrète développée au chapitre suivant. À l'issue de ce chapitre, le lecteur doit disposer d'une compréhension claire de ce que prévoit le règlement québécois, de la manière dont la conformité peut être mesurée objectivement, et des principes théoriques qui guident la conception de l'outil d'anonymisation présenté dans ce mémoire (Gouvernement du Québec, 2024; Office of the Privacy Commissioner of Canada, 2024; Samarati & Sweeney, 1998; Dwork, 2006).

3.1 CONCEPTS FONDAMENTAUX DE L'ANONYMISATION

L'anonymisation s'appuie sur un ensemble de notions qui permettent d'évaluer si on peut identifier une personne, directement ou indirectement, dans un jeu de données (Gouvernement du Québec, 2024). Avant de parler de techniques et de processus, je dois d'abord clarifier ces concepts de base.

3.1.1 IDENTIFIANTS DIRECTS

Les identifiants directs sont des renseignements qui permettent d'identifier immédiatement une personne, sans ambiguïté possible. Il s'agit des informations les plus explicites, puisqu'elles pointent directement vers un individu déterminé. Parmi les exemples les plus courants figurent le nom et le prénom, l'adresse postale complète, le numéro de téléphone, l'adresse courriel personnelle, le numéro d'assurance maladie (NAM), le numéro d'assurance sociale (NAS), le numéro de permis de conduire ou encore le numéro de compte bancaire. Dans une démarche d'anonymisation, ces identifiants doivent généralement être supprimés complètement. Il ne suffit pas de les masquer partiellement ou de les généraliser, car leur simple présence permet encore d'indexer directement une personne. Autrement dit, même si toutes les autres variables du jeu de données sont transformées, le maintien d'un identifiant direct comme le NAS ou l'adresse courriel compromet l'ensemble du processus d'anony-

misation (Office of the Privacy Commissioner of Canada, 2024; Gouvernement du Québec, 2024).

3.1.2 QUASI-IDENTIFIANTS

Les quasi-identifiants (qu'on appelle souvent QI dans la littérature) sont plus delicats. Pris individuellement, ils ne permettent pas d'identifier quelqu'un. Mais quand on les combine ensemble, ils deviennent très révélateurs (Samarati & Sweeney, 1998) ; (El Emam & Dankar, 2012). Exemples de quasi-identifiants : année ou date de naissance, région ou code postal (surtout le FSA – les 3 premiers caractères), sexe ou genre, profession ou secteur d'emploi, revenus approximatifs, situation familiale (célibataire, marié, divorcé), niveau de scolarité, langue maternelle. Pourquoi c'est un probleme ? Imaginons qu'un jeu de données contient : date de naissance (15 mars 1985), code postal (H3C 2A7), sexe (féminin), profession (professeure universitaire). Même sans le nom, cette combinaison pourrait identifier une seule personne dans une petite zone géographique. Si quelqu'un a accès à une liste électorale, il peut facilement faire le lien. C'est ça le danger des quasi-identifiants : leur combinaison peut donner des enregistrements quasi uniques ou complètement uniques (Hundepool *et al.*, 2012).

3.1.3 DONNÉES SENSIBLES

Certaines catégories de renseignements personnels sont particulièrement sensibles. Si elles sont révélées ou devinées, elles peuvent causer un préjudice important aux personnes concernées. Exemples de données sensibles : **état de santé** : diagnostics médicaux, traitements, hospitalisations, médicaments prescrits ; **informations financières** : revenus exacts, soldes de comptes, dettes, faillites ; **données d'assurance** : réclamations, refus de couverture, primes ajustées ; **caractéristiques personnelles sensibles** : origines ethniques, opinions politiques, croyances religieuses, orientation sexuelle. Pourquoi elles sont dangereuses ? Même si on ne peut pas identifier précisément quelqu'un, si on peut deviner qu'une personne dans un

certain groupe a un cancer ou des dettes importantes, ça reste une violation de la vie privée. Le règlement québécois dit explicitement qu'il faut empêcher que ces informations sensibles puissent être devinées par inférence à partir d'autres variables (Gouvernement du Québec, 2024) ; (Amir, 2024).

3.1.4 ANONYMISATION VS PSEUDONYMISATION

Avant d'aller plus loin, je dois clarifier une distinction importante que la Loi 25 fait, tout comme le RGPD en Europe : la différence entre pseudonymisation et anonymisation. **Pseudonymisation :** C'est une modification **réversible** où les identifiants directs sont remplacés par des identifiants techniques (codes, pseudonymes, clés de hachage). Exemple : Jean Dupont → PATIENT_001234. Le problème : si quelqu'un a accès à la table de correspondance (qui dit que PATIENT_001234 = Jean Dupont), il peut retrouver l'identité. C'est réversible (European Data Protection Board, 2020). **Anonymisation :** C'est une modification **irréversible** destinée à empêcher toute réidentification raisonnable, même si d'autres sources de données sont disponibles. Une fois anonymisées, les données ne peuvent plus être reliées à une personne spécifique (Parlement européen & Conseil de l'Union européenne, 2016) ; (Gouvernement du Québec, 2024). Le règlement québécois exige une anonymisation irréversible, pas une simple pseudonymisation (Gouvernement du Québec, 2024).

3.2 LES TROIS CRITÈRES DU RÈGLEMENT QUÉBÉCOIS

Le Règlement sur l'anonymisation des renseignements personnels (Gouvernement du Québec, 2024) repose sur trois critères essentiels qui doivent être respectés simultanément pour que des données puissent être considérées comme véritablement anonymisées : l'individualisation, la corrélation et l'inférence (Borden Ladner Gervais LLP, 2024). C'est vraiment le cœur du règlement québécois. Comprendre ces trois critères est essentiel pour concevoir mon outil.

3.2.1 CRITÈRE 1 : INDIVIDUALISATION

L'individualisation mesure si on peut isoler ou distinguer une personne dans un ensemble de données. En gros : est-ce qu'il y a des enregistrements uniques ou quasi-uniques qui permettraient de pointer vers une personne spécifique ? Ce que ça veut dire concrètement : imaginons un fichier avec ces colonnes : âge, sexe, code postal, profession. Si on trouve qu'il n'y a qu'une seule femme de 42 ans, programmeuse, habitant dans H3C 2A7, alors cette personne est "individualisée" car elle se démarque dans le groupe. Ce critère rejoint directement l'idée du k-anonymat : chaque personne devrait être indiscernable parmi au moins k autres personnes qui partagent exactement les mêmes quasi-identifiants (Samarati & Sweeney, 1998) ; (Sweeney, 2002b). Par exemple, si k = 1, la personne est unique et le risque est très élevé. Si k = 5, elle est noyée parmi 5 personnes identiques, ce qui devient plus acceptable. Si k = 10, le niveau de protection est encore meilleur, et s'il atteint 100, la personne devient très difficile à isoler dans le groupe.

3.2.2 CRITÈRE 2 : CORRÉLATION

La corrélation mesure si un ensemble de données anonymisées peut être relié à une personne en le combinant avec d'autres sources de données externes. Ce que ça veut dire concrètement : même si un fichier est bien anonymisé pris isolément, il faut se demander si quelqu'un pourrait le croiser avec une liste électorale publique, un annuaire téléphonique en ligne, des profils LinkedIn ou Facebook, ou encore d'autres bases de données de l'organisation. Si oui, il y a un problème de corrélation. Cela peut arriver de deux façons principales. Premièrement, certaines colonnes sont directement liables : si on garde par exemple le code postal complet H3C 2A7 et la date de naissance exacte, quelqu'un pourrait facilement croiser cela avec une liste électorale. Deuxièmement, il peut exister des corrélations fortes entre variables : si le revenu est très fortement corrélé avec le code postal, par exemple avec un coefficient de Pearson de 0.95, une personne qui connaît le code postal peut prédire le revenu

avec une grande précision. Le règlement demande justement de tenir compte des données raisonnablement accessibles dans l'environnement de l'organisation. Il faut donc se demander ce qu'un adversaire pourrait raisonnablement obtenir comme données externes (Gouvernement du Québec, 2024) ; (Amir, 2024). Cela implique aussi d'analyser les dépendances entre les colonnes à l'aide de différentes métriques statistiques (Hundepool *et al.*, 2012) ; (National Institute of Standards and Technology, 2018).

3.2.3 CRITÈRE 3 : INFÉRENCE

L'inférence mesure si on peut deviner ou prédire un renseignement personnel sensible indirectement à partir d'autres données, même sans connaître l'identité exacte de la personne. Ce que ça veut dire concrètement : imaginons un groupe k-anonyme de 5 personnes qui partagent les mêmes quasi-identifiants, par exemple âge 40-49, code postal H3C et sexe masculin. Si ces 5 personnes ont toutes le même diagnostic médical, par exemple diabète de type 2, alors même si on ne sait pas qui est qui dans le groupe, on peut deviner avec certitude que toute personne correspondant à ces critères a ce diagnostic. Le k-anonymat est donc respecté, mais la confidentialité est compromise par l'inférence. D'autres situations similaires peuvent apparaître. Si le revenu est fortement corrélé avec le niveau d'éducation et le code postal, il devient possible de prédire le revenu d'une personne même si cette colonne a été supprimée. De la même façon, si tous les employés d'un certain département ont un salaire dans une plage très étroite, on peut deviner le salaire individuel. Ce type de situation illustre les attaques d'inférence, qui comptent parmi les principales limites du k-anonymat seul (Machanavajjhala *et al.*, 2007b) ; (Li & Li, 2007). C'est pour cette raison que la l-diversité a été introduite : elle exige qu'il y ait au moins l valeurs distinctes pour les attributs sensibles dans chaque groupe k-anonyme. Ce critère se rapproche donc des modèles de diversité, comme la l-diversité et le t-closeness, présentés dans le chapitre précédent (Machanavajjhala *et al.*, 2007b) ; (Li & Li, 2007) ; (Dwork, 2006).

3.3 LE PROCESSUS RÉGLEMENTAIRE D'ANONYMISATION

Le règlement québécois ne se limite pas à définir des critères. Il exige aussi un processus formel et structuré en plusieurs étapes (Gouvernement du Québec, 2024) ; (McCarthy, 2024).

3.3.1 ÉTAPE 1 : DÉFINITION DE L'OBJECTIF DE L'ANONYMISATION

Avant même de commencer à transformer les données, il faut préciser le contexte et la finalité du projet. Il faut se demander pourquoi on anonymise ces données, à quoi elles vont servir après anonymisation, qui va y avoir accès et dans quel contexte elles vont être utilisées. Pourquoi c'est important : l'objectif va influencer le niveau de protection nécessaire. Par exemple, si les données sont destinées à une publication publique, il faut viser une protection maximale. Si elles sont destinées à une recherche interne, la protection doit rester élevée, mais il peut y avoir un peu plus de flexibilité sur l'utilité.

3.3.2 ÉTAPE 2 : RETRAIT DES IDENTIFIANTS DIRECTS

Cette étape vise à éliminer ou modifier les identifiants qui permettent de reconnaître directement une personne. Les techniques applicables comprennent la **suppression complète**, qui consiste à retirer la colonne entièrement, ce qui est recommandé pour le NAS, le téléphone, le nom ou le prénom, ainsi que le **masquage partiel**, qui consiste à remplacer une partie des caractères par des astérisques, par exemple pour une adresse courriel comme `je**@te**.com`. Note : la pseudonymisation est mentionnée dans le règlement comme option (European Data Protection Board, 2020) ; (Gouvernement du Québec, 2024), mais je ne l'implémente pas dans l'outil car elle n'est pas irréversible.

3.3.3 ÉTAPE 3 : ANALYSE DES RISQUES AVANT ANONYMISATION

Avant d'appliquer les techniques d'anonymisation, il faut évaluer le niveau de risque initial du jeu de données selon les trois critères (Amir, 2024) ; (Borden Ladner Gervais LLP, 2024). Ce qu'on mesure : le risque d'individualisation, à travers le k-anonymat et les enregistrements uniques, le risque de corrélation, à travers les colonnes liables et les corrélations fortes, ainsi que le risque d'inférence, à travers la prédictibilité des attributs sensibles. Pourquoi c'est important ? Cette analyse initiale sert ensuite à identifier les zones à risque, à choisir les techniques appropriées et à mesurer l'efficacité de l'anonymisation en comparant avant et après.

3.3.4 ÉTAPE 4 : APPLICATION DE TECHNIQUES D'ANONYMISATION

En fonction des risques identifiés, différentes techniques sont appliquées (Hundepool *et al.*, 2012) ; (Dwork, 2006). Parmi les techniques utilisées, on retrouve la **suppression** pour les identifiants directs, le **masquage partiel** pour certains quasi-identifiants lorsqu'on veut garder une partie de l'information, la **généralisation** pour les quasi-identifiants, ainsi que la **confidentialité différentielle** pour les données sensibles numériques.

3.3.5 ÉTAPE 5 : ANALYSE DES RISQUES APRÈS ANONYMISATION

Une fois les techniques appliquées, il faut réévaluer le risque sur les données transformées et comparer avec les scores initiaux pour vérifier que le risque résiduel de réidentification est devenu très faible (Gouvernement du Québec, 2024) ; (Borden Ladner Gervais LLP, 2024).

3.3.6 ÉTAPE 6 : DOCUMENTATION ET REGISTRE

À compter du 1er janvier 2025, les organisations doivent conserver un registre détaillé de leurs démarches d'anonymisation. Ce que le registre doit contenir selon l'article 9 du règlement : la liste des renseignements personnels anonymisés, les finalités pour lesquelles les renseignements anonymisés seront utilisés, les techniques d'anonymisation utilisées, les mesures de sécurité établies, ainsi que la date à laquelle l'analyse du risque de réidentification a été effectuée et la date de sa dernière mise à jour (Gouvernement du Québec, 2024) ; (Langlois Avocats, 2024).

3.4 APPROCHE DE L'ADVERSAIRE RAISONNABLE

Le règlement québécois introduit une notion très pertinente : l'évaluation du risque de réidentification doit se faire en se plaçant du point de vue d'un "adversaire raisonnable".

3.4.1 C'EST QUOI UN ADVERSAIRE RAISONNABLE ?

C'est quelqu'un qui dispose de compétences techniques réelles, mais pas de moyens extrêmes. Ce n'est pas la NSA ou un État avec des supercalculateurs. Il a accès à des bases de données externes raisonnablement accessibles, comme les listes électorales publiques, les annuaires en ligne, les réseaux sociaux tels que LinkedIn ou Facebook, les données ouvertes gouvernementales ou certains registres publics. Il est aussi capable d'effectuer des croisements de données, par exemple des jointures SQL basiques, d'utiliser des méthodes simples de prédiction comme la régression linéaire ou les arbres de décision, ainsi que de faire des analyses statistiques standard.(Gouvernement du Québec, 2024) ; (Borden Ladner Gervais LLP, 2024)

3.4.2 POURQUOI C'EST IMPORTANT ?

Cette approche permet de dimensionner correctement les mesures d'anonymisation sans tomber dans deux excès. Le premier excès consiste à sous-estimer la menace, par exemple en pensant que personne ne cherchera à réidentifier les données parce que l'organisation est trop petite. C'est une erreur, car même une petite organisation peut être ciblée, et l'adversaire n'a pas besoin d'être un génie pour exploiter des données : des outils gratuits en ligne permettent déjà de faire des croisements sophistiqués. Le second excès consiste à surestimer la menace, en imaginant qu'il faut se protéger contre la NSA ou contre des hackers d'élite utilisant des ordinateurs quantiques. Ce n'est pas réaliste, et le règlement demande justement de raisonner de manière raisonable.

Le juste milieu consiste donc à considérer qu'un adversaire pourrait télécharger la liste électorale publique de sa ville, faire une recherche Google ou LinkedIn, utiliser Python avec des bibliothèques standards, ou encore croiser deux fichiers CSV avec Excel. En revanche, il est peu probable qu'il puisse infiltrer des bases de données sécurisées, utiliser des algorithmes d'IA ultra-sophistiqués ou dépenser des millions en ressources de calcul. Cette approche est essentielle pour calibrer correctement les mesures d'anonymisation et s'assurer qu'elles sont adaptées à un niveau de risque réaliste (Fasken, 2024) ; (Amir, 2024).

3.5 LIEN ENTRE LE RÈGLEMENT QUÉBÉCOIS ET LES MODÈLES THÉORIQUES

Une chose intéressante avec le règlement québécois, c'est qu'il ne réinvente pas la roue. Les trois critères s'inscrivent dans la continuité des principaux modèles théoriques d'anonymisation développés dans la recherche académique.

Voici le lien :

TABLEAU 3.1 : Correspondance entre critères québécois et modèles théoriques

Critère québécois	Modèle théorique associé	Ce qu'on mesure
Individualisation	k-anonymat (Samarati & Sweeney, 1998)	k minimum, % d'enregistrements uniques, taille des groupes
Corrélation	Corrélations statistiques, information mutuelle (Hundepool *et al.*, 2012)	Coefficient de Pearson, colonnes liables, croisements possibles
Inférence	l-diversité (Machanavajjhala *et al.*, 2007b), t-closeness (Li & Li, 2007), Confidentialité différentielle (Dwork, 2006)	Entropie, diversité des valeurs sensibles, prédictibilité

Ce que ça veut dire : Le règlement québécois ne crée pas des concepts entièrement nouveaux. Il prend des notions déjà établies dans la recherche scientifique et les traduit en exigences opérationnelles concrètes pour les organisations québécoises (Gouvernement du Québec, 2024).

L'apport du Québec : Ce qui est unique au Québec, c'est l'exigence d'évaluer les trois critères simultanément et l'obligation de documentation dans un registre à partir du 1er janvier 2025.

3.5.1 SYNTHÈSE DU CHAPITRE

Dans ce chapitre, les fondations théoriques et réglementaires nécessaires ont été établi pour concevoir un outil d'anonymisation conforme. Les concepts clés ont été clarifiés, notamment les identifiants directs, les quasi-identifiants, les données sensibles, ainsi que la distinction entre anonymisation et pseudonymisation. Les trois critères du règlement québécois ont été expliqués en détail, à savoir l'individualisation, la corrélation et l'inférence. Le processus réglementaire en six étapes a été présenté, de même que le concept d'adversaire raisonnable, qui permet de calibrer le niveau de protection nécessaire. Enfin, le lien entre les critères québécois et les modèles théoriques établis, comme le k-anonymat, la l-diversité et la

confidentialité différentielle, a été mis en évidence. Le chapitre suivant va maintenant décrire la méthodologie et la conception concrète d'un outil logiciel conforme à ces exigences.

CHAPITRE IV
CONCEPTION ET MÉTHODOLOGIE

Les chapitres 2 et 3 ont établi les cadres théoriques et réglementaires. Pour la suite, ce chapitre présente la conception et le développement concret de l'outil : démarche méthodologique, analyse des besoins, architecture (FastAPI + Next.js + PostgreSQL), choix technologiques, processus d'anonymisation et méthode d'évaluation des risques. Ainsi, il fait le lien entre la théorie et l'implémentation qui sera détaillée au chapitre 5.

4.1 LA DÉMARCHE MÉTHODOLOGIQUE

Pour ce travail, une approche itérative en 7 phases a été suivie, adaptée aux contraintes du règlement québécois.

4.1.1 APPROCHE ITÉRATIVE

L'approche itérative était nécessaire car l'anonymisation nécessite des ajustements constants. En effet, elle a permis de détecter rapidement les problèmes (la détection par regex seule était insuffisante), puis d'ajuster les paramètres selon les tests, et enfin d'améliorer l'interface utilisateur.

4.1.2 LES SEPT PHASES DU PROJET

D'abord, la **Phase 1** a porté sur l'étude du règlement québécois et des modèles théoriques (k-anonymat, l-diversité, confidentialité différentielle etc...). Ensuite, la **Phase 2** a consisté en la définition des besoins : import CSV, classification, catégorisation, calcul 3 critères, configuration techniques, export PDF. Puis, la **Phase 3** a concerné l'architecture trois couches : backend API (FastAPI), frontend web (Next.js), base de données (PostgreSQL). Après cela,

les **Phases 4-5** ont été consacrées au développement de 10 services backend (détection hybride, anonymisation, évaluation risques) et de 4 pages frontend (upload, détection, configuration, résultats). Par la suite, la **Phase 6** a porté sur les tests avec des données synthétiques, ainsi que sur l'ajustement des seuils et paramètres. Enfin, la **Phase 7** a été dédiée à la documentation (README, commentaires code, documentation technique, mémoire).

4.1.3 MODÈLE DE PROTECTION PROPOSÉ

Le modèle de protection proposé dans ce mémoire repose sur une architecture à trois niveaux qui traduit les exigences réglementaires de la Loi 25 en métriques informatiques calculables automatiquement.

Niveau 1 — Évaluation des risques. Pour chaque colonne du dataset, trois scores sont calculés indépendamment : le score d'individualisation S_{indiv} mesure la capacité à isoler un individu via le k-anonymat ; le score de corrélation S_{corr} mesure la liaison possible avec des sources externes ; le score d'inférence S_{inf} mesure la prédictibilité des attributs sensibles via l'entropie de Shannon.

Niveau 2 — Agrégation pondérée. Les trois scores sont combinés en un score global unique selon la formule :

$$S_{global} = 0{,}40 \times S_{indiv} + 0{,}35 \times S_{corr} + 0{,}25 \times S_{inf} \tag{4.1}$$

Niveau 3 — Décision de conformité. Le score global est comparé au seuil opérationnel de 20% : si $S_{global} < 20\%$, le dataset est déclaré conforme à la Loi 25 ; sinon, des techniques d'anonymisation adaptées sont recommandées et appliquées, puis l'évaluation est relancée jusqu'à conformité. Ce modèle itératif garantit que la conformité est atteinte et vérifiable, contrairement aux approches statiques qui n'évaluent le risque qu'une seule fois.

4.2 ANALYSE DES BESOINS

Dans un premier temps, les besoins ont été divisés en trois catégories : exigences réglementaires, fonctionnelles et non-fonctionnelles.

4.2.1 EXIGENCES RÉGLEMENTAIRES ET CHOIX D'IMPLÉMENTATION

D'abord, l'outil doit respecter le règlement québécois (Gouvernement du Québec, 2024). Plus précisément, **ce que le règlement exige** est d'évaluer les trois critères simultanément : individualisation, corrélation et inférence ; de démontrer que le risque de réidentification est « très faible » ; de suivre un processus structuré en 6 étapes ; et de tenir un registre documenté (à partir du 1er janvier 2025).

Cependant, **les choix d'implémentation** tiennent compte du fait que le règlement ne définit pas de pondérations numériques précises ni de seuil quantitatif exact. Par conséquent, pour rendre l'évaluation calculable automatiquement, il a été choisi d'implémenter : **Pondérations** : 40% individualisation, 35% corrélation, 25% inférence. le choix de ces poids reflètent l'importance relative : l'individualisation est le risque le plus direct de réidentification (Sweeney, 2002a), la corrélation devient critique à l'ère du Big Data (National Institute of Standards and Technology, 2015), et l'inférence protège contre les attaques par connaissances de fond (Machanavajjhala *et al.*, 2007a). En outre, **le seuil de conformité** a été fixé à 20%. Ce seuil correspond au standard "très faible" pour les données administratives non-médicales, inspiré des pratiques de la CNIL française (Commission Nationale de l'Informatique et des Libertés, 2020). Ces choix méthodologiques reposent sur une justification à deux niveaux. D'une part, les pondérations reflètent une hiérarchie des risques fondée sur la littérature scientifique : l'individualisation (40%) représente le vecteur de réidentification le plus direct, car un enregistrement unique dans un dataset peut être associé à une personne réelle sans aucune connaissance externe (Sweeney, 2002a). La corrélation (35%) est critique à l'ère du Big Data,

car le croisement de sources multiples amplifie le risque même pour des données apparemment anodines (National Institute of Standards and Technology, 2015). Enfin, l'inférence (25%) protège contre les attaques par connaissances de fond, qui exploitent des attributs sensibles prévisibles (Machanavajjhala *et al.*, 2007a). Ces poids ont été validés en les comparant aux recommandations du WP29, du NIST SP 800-188, et aux lignes directrices de la CNIL, qui convergent vers une priorité accordée à l'individualisation dans les contextes non médicaux. D'autre part, le seuil de 20% a été retenu comme opérationnalisation du critère réglementaire de risque « très faible » au sens de la Loi 25. Ce seuil constitue une interprétation conservative : en dessous de 20%, le risque résiduel est jugé suffisamment bas pour qu'aucune personne ne puisse être raisonnablement identifiée, conformément à l'article 23 du règlement sur la protection des renseignements personnels (Gouvernement du Québec, 2024).

4.2.2 EXIGENCES FONCTIONNELLES

Sur le plan fonctionnel, l'outil doit offrir les fonctionnalités suivantes. D'abord, **1. Import et prévisualisation** : téléversement de fichiers CSV avec aperçu des données. Ensuite, **2. Classification et catégorisation automatique** : classification réglementaire (identifiants directs, quasi-identifiants, données sensibles, données non-sensibles) et catégorisation thématique (santé, finance, biométrique, vie sexuelle, religion, politique, race, renseignements personnels) avec scores de confiance et justifications. Puis, **3. Analyse pré-anonymisation** : calcul des trois critères, k-anonymat minimal, enregistrements uniques, corrélations fortes. De plus, **4. Configuration et application** : sélection automatique des techniques d'anonymisation (suppression, généralisation, confidentialité différentielle) avec recommandations intelligentes. Ensuite, **5. Analyse post-anonymisation** : recalcul des scores, statut de conformité. Enfin, **6. Export** : téléchargement du fichier anonymisé et du rapport PDF pour le registre obligatoire.

4.2.3 EXIGENCES NON-FONCTIONNELLES

Par ailleurs, plusieurs exigences non-fonctionnelles ont également été retenues. En matière de **performance**, il faut traiter 10000 lignes en moins de 30 secondes, avec un support allant jusqu'à 100000 lignes. En matière de **sécurité**, il ne doit pas y avoir de stockage permanent, avec isolation par session et validation stricte des entrées. En matière de **traçabilité**, un historique complet des opérations et des logs détaillés en base de données doit être conservé. Enfin, du point de vue de la **maintenabilité**, le code doit être modulaire, bien documenté et fondé sur une architecture extensible.

4.3 ARCHITECTURE DE L'OUTIL

Dans cette section, l'architecture de l'outil est présentée. Il s'agit d'une architecture moderne à trois couches : frontend, backend et base de données.

4.3.1 VUE D'ENSEMBLE

Globalement, cette architecture suit le modèle classique client-serveur : **Frontend** : Next.js 16.1 + React 19.2.3 (interface utilisateur web). **Backend** : FastAPI + Python 3.11+ (API REST + logique métier). **Base de données** : PostgreSQL 15 (stockage des métadonnées). En pratique, les trois communiquent comme suit : Frontend $\leftrightarrow$ Backend : requêtes HTTP/REST (format JSON). Backend $\leftrightarrow$ Base de données : SQLAlchemy ORM (requêtes SQL).

4.3.2 POURQUOI CETTE ARCHITECTURE ?

Pourquoi séparer frontend et backend ? Tout aurait pu être réalisé dans une seule application, mais la séparation a été retenue pour plusieurs raisons. D'abord, **1. Scalabilité** : si beaucoup d'utilisateurs utilisent l'outil en même temps, il est possible de déployer plus d'instances backend sans toucher au frontend. Ensuite, **2. Flexibilité** : l'API peut être utilisée

par d'autres clients (mobile, ligne de commande, scripts) sans modifier le code. De plus, **3. Maintenabilité** : le frontend et le backend ont des responsabilités claires et séparées. Enfin, **4. Technologies optimales** : Python est excellent pour le traitement de données, tandis que JavaScript/React est excellent pour les interfaces.

Pourquoi PostgreSQL et pas SQLite ? Là encore, le choix est motivé par plusieurs avantages. En effet, PostgreSQL supporte plusieurs utilisateurs en même temps, tandis que (SQLite est mono-utilisateur), des types de données avancés (JSON, arrays), et de meilleures performances sur de gros volumes.

4.3.3 ARCHITECTURE BACKEND

Plus précisément, le backend FastAPI est organisé en trois couches distinctes selon le principe de séparation des responsabilités.

Couche 1 : API (app/api/v1/) D'abord, cette couche expose les endpoints REST aux clients. Elle contient un routeur principal qui agrège tous les endpoints et trois modules d'endpoints organisés par domaine fonctionnel : gestion datasets, anonymisation, et authentification.

Ensuite, les endpoints REST sont organisés par domaine fonctionnel :

— `POST /api/v1/datasets/upload` : Upload CSV

— `GET /api/v1/datasets/{id}` : Métadonnées

— `GET /api/v1/datasets/{id}/preview` : Aperçu (10 lignes)

— `GET /api/v1/datasets/{id}/statistics` : Statistiques

— `DELETE /api/v1/datasets/{id}` : Suppression

— `POST /api/v1/datasets/{id}/detect` : Détection automatique

— `GET /api/v1/datasets/{id}/risk-assessment` : Évaluation Loi 25

— `POST /api/v1/datasets/{id}/anonymize` : Anonymisation manuelle

— `POST /api/v1/datasets/{id}/auto-anonymize` : Anonymisation automatique

— `GET /api/v1/datasets/{id}/download` : Téléchargement CSV anonymisé

— `GET /api/v1/datasets/{id}/report` : Génération rapport PDF

Enfin, pour l'**authentification** : `POST /api/v1/auth/register` : Inscription ; `POST /api/v1/auth/login` : Connexion (JWT).

2. Couche Services (app/services/) Ensuite, la logique métier est répartie en 10 services spécialisés.

Concernant l'**import et traitement** : `data_ingestion.py` : Import CSV, validation format, détection encodage, parsing Pandas ; `visualization.py` : Statistiques descriptives, histogrammes, analyse corrélations.

Concernant la **détection des données sensibles** : `detector.py` : Détection heuristique (regex, mots-clés, analyse statistique) ; `ai_enhanced_detector.py` : Amélioration par IA (Ollama + gemma3 :8b) pour colonnes ambiguës, hybride règles + IA.

Concernant l'**anonymisation** : `anonymizer.py` : Orchestration des 3 techniques (généralisation, suppression) ; `differential_privacy.py` : Mécanisme de Laplace pour confidentialité différentielle ; `verification.py` : Vérification post-anonymisation (k-anonymité, qualité données).

Concernant l'**évaluation et reporting** : `risk_evaluator.py` : Calcul 3 critères Loi 25 + score global ; `report_generator.py` : Génération PDF avec ReportLab.

Enfin, pour l'**authentification** : `auth.py` : Authentification JWT, hashage bcrypt, gestion sessions.

3. Couche Modèles (app/models/) Enfin, le schéma de base de données comprend 8 modèles SQLAlchemy.

Pour la **gestion des datasets** : `Dataset` : Métadonnées fichiers CSV (nom, taille, lignes, colonnes, score risque, conformité) ; `DatasetColumn` : Classification détaillée colonnes (type sensibilité, catégorie, confiance, échantillons).

Pour les **opérations d'anonymisation** : `AnonymizationJob` : Historique travaux (dataset origine, résultat, statut, durée) ; `TransformationLog` : Audit trail transformations (colonne, technique, paramètres, échantillons avant/après) ; `SuppressedColumn` : Traçabilité colonnes supprimées (nom, justification).

Pour l'**évaluation et la vérification** : `RiskAssessment` : Résultats évaluations Loi 25 (3 scores, niveau conformité, recommandations) ; `VerificationLog` : Logs vérification post-anonymisation (k-anonymité, métriques qualité).

Enfin, pour l'**authentification** : `User` : Comptes utilisateurs (email, mot de passe hashé, timestamps).

TABLEAU 4.1 : Relations entre modèles de base de données

Modèle	Relations	Description
Dataset	1-N DatasetColumn 1-N AnonymizationJob 1-1 RiskAssessment	Un dataset a plusieurs colonnes Peut être anonymisé plusieurs fois Une évaluation par dataset
AnonymizationJob	1-N TransformationLog 1-N SuppressedColumn 1-1 VerificationLog	Un job a plusieurs transformations Peut supprimer plusieurs colonnes Une vérification par job
User	1-N Dataset	Un utilisateur crée plusieurs datasets

4.3.4 ARCHITECTURE FRONTEND

Par ailleurs, le frontend est développé avec Next.js 16.1 (App Router) et React 19.2.3, suivant une architecture de Single Page Application avec Server-Side Rendering. Plus concrètement, l'application est organisée selon l'App Router de Next.js avec des pages pour chaque étape du workflow (upload, détection, configuration, résultats), ainsi qu'une page de connexion

distincte, des composants réutilisables (Header, Stepper, tableaux, graphiques), un client API centralisé type-safe pour toutes les communications backend, et des utilitaires pour le formatage et l'authentification.

Pages principales : 4 pages de workflow : (1) Upload (`/`) : drag & drop CSV, validation format/taille, redirection auto ; (2) Détection (`/detection/[id]`) : 4 cartes résumé, tableau colonnes avec badges, score risque avec barre progression ; (3) Configuration (`/anonymization/[id]`) : sélection techniques par colonne, configuration paramètres, prévisualisation ; (4) Résultats (`/results/[id]`) : banner conformité, 3 jauges circulaires, téléchargement CSV/PDF. À cela s'ajoute une page de connexion dédiée à l'authentification des utilisateurs.

Composants réutilisables : Header (navigation, logout), Stepper (4 étapes), ProgressBadge, DataPreviewTable (paginé, tri/recherche), InteractiveCharts (Recharts jauges), ProtectedRoute (HOC auth).

Client API (`lib/api.ts`) : 12 méthodes type-safe (upload, get, detect, anonymize auto/manuel, assess risk, download CSV/PDF, login).

Technologies : Next.js 16.1, React 19.2.3, TypeScript 5, Tailwind CSS v4, Recharts 3.6.

4.3.5 BASE DE DONNÉES

Enfin, 5 tables principales sont utilisées : `datasets` : Un enregistrement par fichier CSV uploadé ; `dataset_columns` : Une ligne par colonne de chaque dataset ; `anonymization_jobs` : Un enregistrement par anonymisation effectuée ; `transformation_logs` : Détails de chaque transformation appliquée ; `risk_assessments` : Résultats des évaluations de risque. En ce qui concerne les relations, un dataset a plusieurs colonnes (1 :N), un dataset peut avoir plusieurs jobs d'anonymisation (1 :N), et un job a plusieurs transformation logs (1 :N).

4.4 CHOIX TECHNOLOGIQUES

Dans cette partie, les choix de technologies pour backend, frontend et base de données sont justifiés.

4.4.1 BACKEND : FASTAPI + PYTHON

Le choix de FastAPI s'explique par la performance (async/await Starlette), la validation auto (Pydantic), la documentation auto (Swagger), et le typage moderne. Quant à Python, il a été retenu pour Pandas (DataFrames), NumPy (calculs), Scikit-learn (stats), et son écosystème anonymisation mature.

4.4.2 FRONTEND : NEXT.JS + REACT

Concernant le frontend, Next.js a été choisi pour le routing auto, les optimisations (code splitting, lazy loading), TypeScript natif, et SSR. React a été retenu pour les interfaces réactives, les tableaux virtualisés, et les composants réutilisables. Enfin, TypeScript permet la réduction des erreurs runtime, l'auto-complétion et la maintenabilité.

4.4.3 BASE DE DONNÉES : POSTGRESQL

Du côté de la base de données, PostgreSQL a été choisi pour les multi-connexions simultanées, les types avancés (JSONB), les performances sur gros volumes, et les transactions ACID robustes.

4.4.4 BIBLIOTHÈQUES PYTHON PRINCIPALES

Parmi les bibliothèques Python principales, on retrouve Pandas 2.2+ (DataFrames CSV), NumPy (calculs, bruit Laplace), SQLAlchemy 2.0 (ORM PostgreSQL), Pydantic v2 (validation schémas API), et ReportLab (génération PDF rapports).

4.4.5 DÉPLOIEMENT : DOCKER

Enfin, le déploiement repose sur 3 conteneurs (backend FastAPI, frontend Next.js, BDD PostgreSQL) orchestrés par Docker Compose pour assurer la portabilité, l'isolation et le déploiement facilité.

4.5 PROCESSUS D'ANONYMISATION

Cette section explique comment l'outil anonymise concrètement les données, étape par étape.

4.5.1 CHOIX DES TECHNIQUES D'ANONYMISATION

Avant de détailler le processus, il faut justifier pourquoi exactement ces 3 techniques ont été choisies et pas d'autres.

Pourquoi 3 techniques seulement ? D'abord, le choix repose sur le principe du *minimum viable* : implémenter le nombre minimal de techniques permettant de couvrir tous les types de données et tous les critères du règlement québécois. Selon le guide de la CNIL (Commission Nationale de l'Informatique et des Libertés, 2020), une combinaison de 3 à 5 techniques complémentaires est suffisante pour traiter la majorité des cas d'usage réels. De plus, le NIST recommande également d'éviter la multiplication inutile des méthodes, car cela

augmente la complexité sans améliorer significativement la protection (National Institute of Standards and Technology, 2015).

Pourquoi ces 3 techniques spécifiquement ? Ensuite, ce choix couvre les trois grandes familles de techniques reconnues par la littérature scientifique et les régulateurs : Premièrement, la **Suppression** : technique de base recommandée par le WP29 (Article 29 Data Protection Working Party, 2014) et le règlement québécois pour éliminer définitivement les identifiants directs (Gouvernement du Québec, 2024). C'est la seule méthode garantissant l'irréversibilité totale exigée par la Loi 25 ; Deuxièmement, la **Généralisation** : technique fondamentale pour atteindre la k-anonymité (Sweeney, 2002a). Elle est explicitement mentionnée dans le règlement québécois comme méthode acceptable pour réduire la précision des quasi-identifiants (Gouvernement du Québec, 2024) ; Troisièmement, la **3. Confidentialité différentielle** : la seule technique offrant des garanties mathématiques formelles de protection (Dwork, 2006). Elle est particulièrement adaptée aux données numériques sensibles (revenus, montants financiers) où les trois autres techniques sont insuffisantes. Le NIST la recommande comme complément aux techniques classiques pour les cas nécessitant des garanties formelles (National Institute of Standards and Technology, 2015).

Techniques écartées et justification : Par ailleurs, certaines techniques ont été écartées : **Agrégation**, non implémentée car elle transforme la structure du dataset (passe de données individuelles à données agrégées), ce qui sort du cadre d'usage de l'outil. L'agrégation est plutôt une technique de publication statistique que d'anonymisation de microdata ; **Ajout de données synthétiques**, écartée car elle augmente artificiellement la taille du dataset et peut introduire des biais statistiques. De plus, le règlement québécois exige de travailler sur les données réelles, pas sur des données générées. **Permutation (shuffling)** : non retenue car elle détruit toutes les corrélations entre colonnes, rendant le dataset inutile pour la plupart des analyses. La généralisation préserve mieux l'utilité statistique ; **t-closeness et autres raffinements de l-diversité**, trop complexes à implémenter et à expliquer aux utilisateurs non-

experts. La confidentialité différentielle offre de meilleures garanties avec une implémentation plus simple.

Ainsi, cette combinaison de 3 techniques permet de traiter tous les types de colonnes (identifiants directs, quasi-identifiants, données sensibles) tout en satisfaisant les trois critères du règlement québécois (individualisation, corrélation, inférence).

4.5.2 ÉTAPE 1 : CHARGEMENT ET VALIDATION

D'abord, quand l'utilisateur upload un CSV, le fichier est sauvegardé temporairement sur le serveur. Ensuite, Pandas charge le CSV en DataFrame. Puis, on vérifie que le fichier est bien du CSV valide, qu'il n'est pas trop gros (< 1 Go) et qu'il a au moins 2 colonnes et 1 ligne. Après cela, on crée un enregistrement `Dataset` dans la base de données avec les métadonnées, puis on retourne un aperçu (premières 10 lignes) au frontend.

4.5.3 ÉTAPE 2 : DÉTECTION AUTOMATIQUE

Ensuite, le système de détection fonctionne selon une approche hybride en deux niveaux. **Détection de base par règles heuristiques :** d'abord, une détection de base par règles heuristiques est réalisée à l'aide du service `detector.py`. Cette étape repose sur trois techniques complémentaires : la détection par expressions régulières sur un échantillon de valeurs de chaque colonne, l'analyse des noms de colonnes à partir de listes de mots-clés, puis une analyse statistique pour les colonnes qui ne sont toujours pas classées. Par exemple, les expressions régulières permettent de reconnaître des formats comme le NAS, l'email, le téléphone canadien ou le code postal canadien, tandis que l'analyse des noms de colonnes permet d'associer certains termes aux identifiants directs, aux quasi-identifiants ou aux données sensibles. Enfin, l'analyse statistique complète cette détection en examinant la cardinalité, le type de données et certains patterns dans les distributions.

Amélioration par intelligence artificielle : ensuite, pour les colonnes dont le score de confiance reste faible, c'est-à-dire inférieur à 70% par défaut, le système utilise le service `ai_enhanced_detector.py`, fondé sur Ollama comme modèle LLM local, afin de valider et d'améliorer la classification. La stratégie retenue est hybride : la détection de base est exécutée en premier, puis les colonnes ambiguës sont envoyées au modèle IA avec leur nom, des échantillons de valeurs et le résultat initial. Le modèle retourne alors sa propre classification, qui est ensuite combinée avec celle obtenue par les règles heuristiques.

Résultat final : au final, pour chaque colonne, le système génère une classification réglementaire parmi `DIRECT_IDENTIFIER`, `QUASI_IDENTIFIER`, `SENSITIVE` ou `NON_SENSITIVE`, ainsi qu'une catégorisation thématique, un score de confiance entre 0 et 100% et une justification textuelle expliquant la décision prise.

4.5.4 ÉTAPE 3 : ÉVALUATION PRÉ-ANONYMISATION

Avant d'anonymiser, les scores initiaux sont calculés pour avoir une base.

Calcul de l'individualisation : d'abord, les lignes sont groupées par quasi-identifiants, puis le k minimum (taille du plus petit groupe) est déterminé, ensuite le nombre de lignes dans des groupes < 5 est compté, puis le score = % de lignes à risque.

Calcul de la corrélation : ensuite, les colonnes "liables" (qui existent dans des sources publiques) sont identifiées, les corrélations de Pearson entre toutes les paires de colonnes numériques sont calculées, le nombre de colonnes ayant $|r| > 0.7$ (seuil standard en statistiques pour corrélations fortes) est compté, puis le score = % de colonnes liables.

Calcul de l'inférence : enfin, pour chaque donnée sensible, l'entropie dans les groupes k-anonymes est calculée, les corrélations entre attributs non-sensibles et sensibles sont détectées, puis le score = risque moyen d'inférence.

Score global :

$$S_{\text{global}} = 0.40 \times S_{\text{indiv}} + 0.35 \times S_{\text{corr}} + 0.25 \times S_{\text{inf}} \tag{4.2}$$

4.5.5 ÉTAPE 4 : CONFIGURATION DES TRANSFORMATIONS

Pour commencer, sur la base de la détection, les transformations suivantes sont proposées automatiquement. D'abord, pour les identifiants directs comme le NAS, l'email, le téléphone, le nom ou le prénom, l'outil propose de supprimer la colonne. Ce choix se justifie parce que les identifiants directs doivent disparaître de façon irréversible pour sortir du champ d'application de la Loi 25 et pouvoir utiliser les données librement (Article 29 Data Protection Working Party, 2014). Ainsi, l'outil applique une suppression physique dans le fichier final pour garantir l'irréversibilité exigée par la CAI.

Ensuite, pour les données sensibles numériques comme les revenus ou les soldes, l'outil propose l'ajout de bruit Laplace avec $\varepsilon = 1.0$. Ce choix est raisonnable car le mécanisme de Laplace garantit que l'ajout ou le retrait d'un individu ne modifie pas significativement le résultat statistique (Dwork, 2006). Ici, $\varepsilon = 1.0$ est utilisé comme valeur de compromis entre protection et utilité statistique.

Enfin, pour les quasi-identifiants, l'outil propose une généralisation des valeurs. Ainsi, les dates sont ramenées à l'année seulement, les codes postaux au FSA (3 premiers caractères), les âges sont regroupés en tranches de 10 ans, et les revenus, lorsqu'ils ne sont pas sensibles, sont transformés en tranches. Ceci est choisi parce que la généralisation hiérarchique consiste à remplacer une valeur spécifique par une valeur plus large en montant dans l'arbre de hiérarchie, par exemple de la ville vers la région puis vers la province. De cette manière, cette approche minimise la perte d'information (Information Loss) tout en augmentant la taille des classes d'équivalence, ce qui améliore la k-anonymité sans supprimer la colonne entière (Sweeney,

2002a). De plus, les tranches d'âge de 10 ans et les codes postaux tronqués suivent les standards HIPAA Safe Harbor (National Institute of Standards and Technology, 2020).

4.5.6 ÉTAPE 5 : APPLICATION DES TECHNIQUES

À ce stade, les quatre techniques implémentées sont détaillées ci-dessous.

TECHNIQUE 1 : SUPPRESSION

D'abord, c'est la plus simple : la colonne est supprimée du DataFrame avec `df.drop(columns=[...])`, la colonne supprimée et sa justification sont enregistrées dans la table `transformation_logs`, et les colonnes avec >95% de valeurs manquantes (seuil de sparsity) sont aussi supprimées, parce que ceci est lié à la malédiction de la dimensionnalité (Curse of Dimensionality) en anonymisation (de Montjoye *et al.*, 2013). En effet, une colonne renseignée pour seulement 2% de la population agit comme une "empreinte digitale" pour ces individus. Même si les quasi-identifiants sont protégés, la simple présence d'une valeur rare permet une identification par connaissances de fond. Ainsi, en supprimant ces colonnes creuses, les vecteurs de corrélation les plus risqués sont éliminés tout en préservant l'intégrité globale du dataset.

TECHNIQUE 2 : GÉNÉRALISATION

Ensuite, pour les âges, des tranches de 10 ans sont utilisées, par exemple 14 → "10-19" ou 42 → "40-49". Pour les dates, seule l'année est conservée, par exemple 1985-03-15 → 1985, parce que les dates complètes (JJ/MM/AAAA) sont des quasi-identifiants très précis. Selon les travaux de Sweeney, la combinaison du jour de naissance, du sexe et du code postal suffit à identifier 87% de la population américaine (Sweeney, 2002a). En limitant la donnée à

l'année uniquement, on réduit la précision de la colonne d'un facteur d'environ 365, ce qui force le regroupement de certains individus dans la même classe d'équivalence temporelle. Par conséquent, cela rend l'individualisation quasi-impossible sans affecter les analyses de tendances annuelles (utilité statistique préservée). Pour les codes postaux, seul le FSA est conservé, par exemple H3C 2A7 $\rightarrow$ H3C.

TECHNIQUE 3 : CONFIDENTIALITÉ DIFFÉRENTIELLE

Enfin, c'est la technique la plus sophistiquée. Elle repose sur le mécanisme de Laplace proposé par Dwork (Dwork, 2006). D'abord, la sensibilité est calculée comme l'impact maximum qu'une personne peut avoir sur le résultat. Pour une requête "moyenne", elle est donnée par :

$$\Delta f = \frac{\max - \min}{n} \tag{4.3}$$

Par exemple, pour des revenus entre 0 et 200 000$ avec 1000 personnes :

$$\Delta f = \frac{200000 - 0}{1000} = 200 \tag{4.4}$$

Ensuite, avec epsilon = 1.0, l'échelle du bruit de Laplace est calculée comme suit :

$$\text{scale} = \frac{\Delta f}{\varepsilon} = \frac{200}{1.0} = 200 \tag{4.5}$$

Un bruit aléatoire selon Laplace(0, 200) est alors généré pour chaque valeur. Enfin, pour chaque revenu, la valeur bruitée est obtenue par :

$$\text{revenu\_bruité} = \text{revenu\_original} + \text{Laplace}(0, 200) \tag{4.6}$$

Par exemple, une valeur de 65 000$ peut devenir 65 123$ après ajout du bruit. Toutefois, le bruit est différent pour chaque personne, mais en moyenne il s'annule, donc les statistiques globales restent correctes.

4.5.7 ÉTAPE 6 : VÉRIFICATION POST-ANONYMISATION

Enfin, après avoir appliqué les transformations, les trois critères sont recalculés sur les données anonymisées, le score global est recalculé, il est vérifié qu'aucun identifiant direct ne demeure (la détection est refaite), et la conformité est déterminée : le dataset est déclaré conforme si le score est inférieur à 20%. Dans le cas contraire, des techniques d'anonymisation supplémentaires sont recommandées et le processus est relancé depuis l'étape 4.

4.6 MÉTHODE D'ÉVALUATION DES RISQUES

Dans cette section, le calcul exact des trois critères est présenté.

4.6.1 CRITÈRE 1 : INDIVIDUALISATION (40%)

Ce qui est calculé : D'abord, **k-anonymat minimal** (Samarati & Sweeney, 1998) : les lignes sont groupées par les quasi-identifiants, la taille du plus petit groupe est trouvée, et si k_min < 5 $\rightarrow$ risque élevé (seuil de Sweeney : un individu doit être confondu avec au moins 4 autres). Ensuite, **Pourcentage d'enregistrements uniques** : le nombre de groupes ayant k=1 est compté, puis % uniques = (nb groupes k=1) / (nb total lignes) × 100. Enfin, **Pourcentage à risque** : le nombre de lignes dans des groupes k<5 est compté, puis % risque = (nb lignes dans k<5) / (nb total) × 100.

Formule du score :

$$S_{\text{indiv}} = \min(100, \%\ \text{uniques} + \text{bonus\_k}) \tag{4.7}$$

où bonus_k = 20 si k_min < 5, sinon 0.

4.6.2 CRITÈRE 2 : CORRÉLATION (35%)

Ce qui est calculé : D'une part, **Colonnes liables** : identifiants directs (email, NAS, téléphone), quasi-identifiants précis (code postal complet, date de naissance exacte), et colonnes avec >95% de valeurs uniques. D'autre part, **Corrélations fortes** : Pearson est calculé pour toutes les paires de colonnes numériques et le nombre de paires ayant |r| > 0.7 (seuil standard en statistiques) est compté.

Formule du score :

$$S_{\text{corr}} = \frac{\text{nb colonnes liables}}{\text{nb total colonnes}} \times 100 \tag{4.8}$$

4.6.3 CRITÈRE 3 : INFÉRENCE (25%)

Ce qui est calculé : D'abord, **Entropie des données sensibles** (Hundepool *et al.*, 2012) : pour chaque groupe k-anonyme, la precision de Shannon des attributs sensibles est calculée ; si la precision < 2 bits → faible diversité → risque. Ensuite, **Corrélations avec données sensibles** : les corrélations entre colonnes non-sensibles et sensibles sont calculées ; si |r| > 0.7 → il devient possible de prédire la valeur sensible.

Formule du score :

$$S_{\text{inf}} = \frac{\text{nb corrélations fortes avec sensibles}}{\text{nb paires possibles}} \times 100 \tag{4.9}$$

4.6.4 SCORE GLOBAL ET CONFORMITÉ

Enfin, le score global combine les trois :

$$S_{\text{global}} = 0.40 \times S_{\text{indiv}} + 0.35 \times S_{\text{corr}} + 0.25 \times S_{\text{inf}} \tag{4.10}$$

Détermination de la conformité : Si $S_{\text{global}} < 10\% \rightarrow$ **Risque FAIBLE**. Si $10\% \leq S_{\text{global}} < 20\% \rightarrow$ **Risque ACCEPTABLE**. Si $20\% \leq S_{\text{global}} < 30\% \rightarrow$ **Risque MOYEN**. Si $S_{\text{global}} \geq 30\% \rightarrow$ **Risque ÉLEVÉ**. **Conforme** = Score global < 20%.

4.6.5 RISQUE RÉSIDUEL MINIMUM

Enfin, même après une anonymisation parfaite, il reste toujours un petit risque résiduel incompressible. Dans l'outil, un plancher de risque de 0.1% est appliqué pour refléter cette réalité scientifique. Ainsi, cette approche d'honnêteté scientifique est requise car l'anonymisation purement mathématique (0.0% de risque) est théoriquement impossible à garantir contre des attaques futures ou des connaissances de fond imprévues (Machanavajjhala *et al.*, 2007a).

4.7 GÉNÉRATION DE RAPPORTS

Pour respecter l'exigence du registre obligatoire depuis janvier 2025 (Borden Ladner Gervais LLP, 2024), l'outil génère automatiquement un rapport PDF complet. Plus précisément, le rapport inclut les informations générales (nom fichier, date, dimensions), les résultats de détection avec scores de confiance et justifications, les transformations appliquées, les risques après anonymisation, et des recommandations si non-conforme. Pour cela, ReportLab est utilisé pour générer le PDF programmatiquement avec page de garde, table des matières, sections numérotées, tableaux comparatifs et graphiques des scores.

4.8 TESTS ET VALIDATION

L'outil a été validé avec un jeu de données médicales synthétiques de 1163 lignes comportant 7 identifiants directs, 8 quasi-identifiants, 5 données sensibles et 5 non-sensibles, généré avec Synthea (**?**). Ce dataset a été choisi pour sa représentativité des données de santé, secteur particulièrement visé par la Loi 25.

Les résultats confirment une réduction du score global de 87,8% à 1,1% après anonymisation, bien en dessous du seuil de conformité de 20%. Un script de test *end-to-end* valide le workflow complet : upload, détection, évaluation initiale, anonymisation, évaluation finale, vérification de conformité et exports (CSV + PDF).

Discussion sur la généralisation. Il convient de reconnaître que la validation sur un seul dataset constitue une limite de ce travail. Les résultats démontrent la faisabilité et la correction de l'approche dans un contexte médical, mais ne permettent pas de conclure à une robustesse universelle. Des évaluations sur des datasets de domaines différents (RH, finance, données de géolocalisation) permettraient de renforcer la validité externe des résultats. Ceci est identifié comme une perspective prioritaire (voir section 6.4). Néanmoins, le choix d'un dataset médical synthétique se justifie par le fait que ce secteur présente la combinaison la plus complexe des trois critères réglementaires, rendant les résultats obtenus un test rigoureux de l'outil.

CHAPITRE V

IMPLÉMENTATION ET RÉSULTATS

Ce chapitre présente l'implémentation effective de l'outil d'anonymisation développé dans le cadre de ce mémoire ainsi que les résultats obtenus lors des tests. L'objectif est de montrer concrètement comment le prototype a été conçu et testé sur un jeu de données médical synthétique comportant 1163 enregistrements générés avec Synthea (Walonoski *et al.*, 2018).

5.1 ENVIRONNEMENT DE DÉVELOPPEMENT

5.1.1 CONTEXTE MATÉRIEL

Le développement du prototype a d'abord été réalisé sur un ordinateur Dell équipé de 16 Go de RAM. L'intégration d'Ollama pour la détection assistée par intelligence artificielle a rapidement montré les limites de cette configuration. Le modèle local Gemma consommait une quantité importante de mémoire, ce qui empêchait l'outil de tourner correctement. Une migration matérielle a été nécessaire vers un **MacBook Air M4 doté de 16 Go de RAM**, mieux adapté à l'exécution du modèle Gemma 1 :8b via Ollama.

5.1.2 LOGICIELS UTILISÉS

L'environnement technique comprend Python 3.11.4 et Node.js 24.13.0 pour les principales composantes logicielles, ainsi que PostgreSQL 15 et Docker v4.64.0 pour la gestion des services. Le projet a été suivi avec Git (7 commits sur GitHub).

5.2 IMPLÉMENTATION DU BACKEND

5.2.1 STRUCTURE PRINCIPALE DU CODE

Le backend repose sur cinq fichiers principaux : **detector.py** (712 lignes) pour la détection initiale avec 15 patterns regex, **ai_enhanced_detector.py** (355 lignes) pour l'analyse assistée par Ollama Gemma 1 :8b, **anonymizer.py** (1058 lignes) pour l'application des techniques d'anonymisation, **differential_privacy.py** (433 lignes) pour le mécanisme de Laplace, et **risk_evaluator.py** (685 lignes) pour le calcul des trois critères Loi 25.

5.2.2 DÉTECTION AUTOMATIQUE DES COLONNES SENSIBLES

La détection combine trois approches : primo, patterns regex pour formats structurés (NAS, RAMQ, email, téléphone, codes postaux, dates, coordonnées GPS), deuxio, analyse sémantique basée sur dictionnaires de mots-clés français/anglais, et tersio, IA via Ollama pour colonnes ambiguës (score < 70%).

5.2.3 TECHNIQUES D'ANONYMISATION IMPLÉMENTÉES

Trois techniques principales sont implémentées : de un **Suppression totale** pour identifiants directs (nom, prénom, NAS, passeport, permis) ; de deux **Généralisation** pour quasi-identifiants : dates → année uniquement (`1985-03-15` → `1985`), réduction précision géolocalisation (codes postaux, coordonnées GPS) ; de trois **Confidentialité différentielle** avec mécanisme de Laplace (Dwork, 2006) pour données numériques sensibles (dépenses santé), epsilon = 1.0. Des techniques complémentaires traitent les données catégorielles sensibles : regroupement pour `RACE` et `ETHNICITY`, normalisation pour `GENDER`.

5.2.4 ÉVALUATION DES RISQUES

Le module `risk_evaluator.py` calcule les trois critères : individualisation (k-anonymat + % lignes avec k < 5), corrélation (% colonnes liables sources externes), inférence (entropie Shannon, seuil 2.0 bits). Le score global est calculé selon la formule $(0.40 \times \text{indiv}) + (0.35 \times \text{corr}) + (0.25 \times \text{inf})$. Un score < 20% indique une conformité Loi 25.

5.3 IMPLÉMENTATION DU FRONTEND

Le frontend est une interface Next.js avec workflow guidé en 4 étapes.

5.3.1 ÉTAPE 1 : CONNEXION ET IMPORT DU FICHIER

Pour accéder à l'outil, l'utilisateur doit d'abord disposer d'un compte. Si ce n'est pas encore le cas, il doit en créer un en remplissant les informations demandées. Une fois le compte créé, il peut se connecter en saisissant son adresse email et son mot de passe. La Figure 5.1 montre l'écran de connexion avec le champ "Adresse email" et le bouton "Se connecter". En bas de l'écran, on voit la mention "Conforme à la Loi 25 du Québec" qui rappelle l'objectif de l'outil.

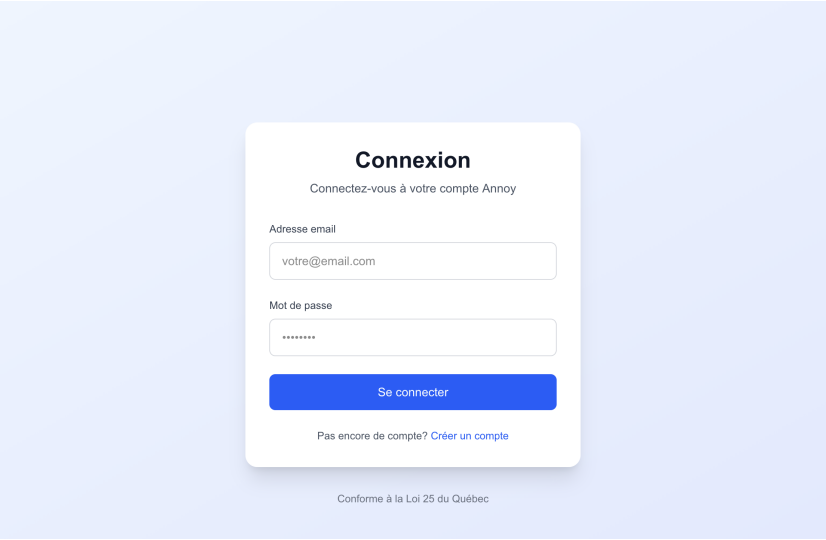

FIGURE 5.1 : Écran de connexion de l'outil Annoy.

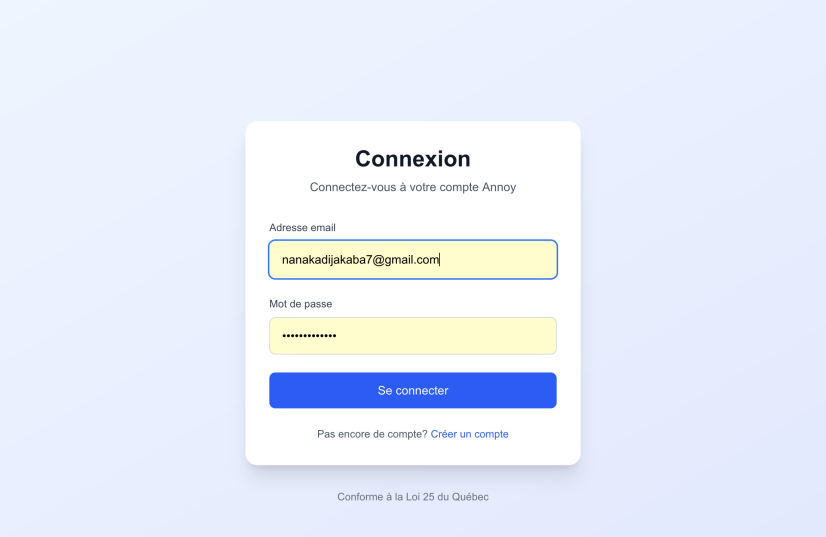

FIGURE 5.2 : Écran de connexion avec les informations d'authentification saisies.

Une fois connecté, l'utilisateur accède à la page d'accueil (Figure 5.3). En haut, on voit le nom de l'utilisateur "Nana kadija Kaba", un indicateur de progression "25% (1/4)" qui montre qu'on est à la première étape, et un bouton "Import". Le stepper horizontal affiche les 4 étapes du workflow : "1 Import" (étape active en bleu), "2 Analyse Initiale" (gris), "3 Anonymisation" (gris), "4 Post-Analyse" (gris). Chaque étape a un sous-titre explicatif.

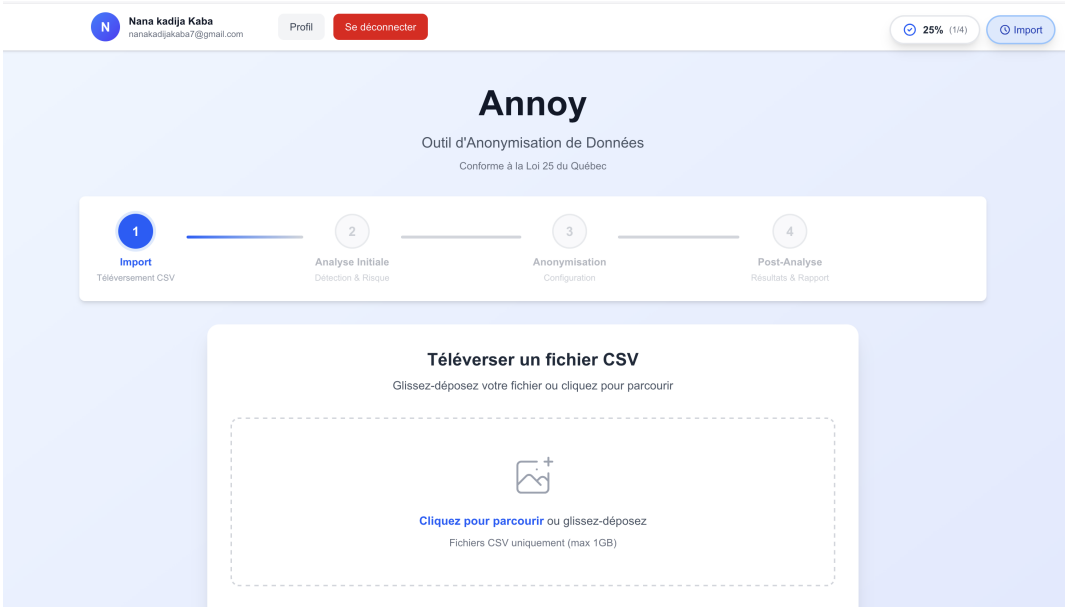

FIGURE 5.3 : Page d'accueil avec le workflow en 4 étapes.

Plus bas sur la même page (Figure 5.4), une grande zone de téléversement invite l'utilisateur à glisser-déposer son fichier CSV. Le texte indique "Glissez-déposez votre fichier ou cliquez pour parcourir" avec la mention "Fichiers CSV uniquement (max 1GB)". En dessous, trois étapes numérotées expliquent ce qui se passera ensuite : détection automatique, configuration des techniques, et évaluation de conformité.

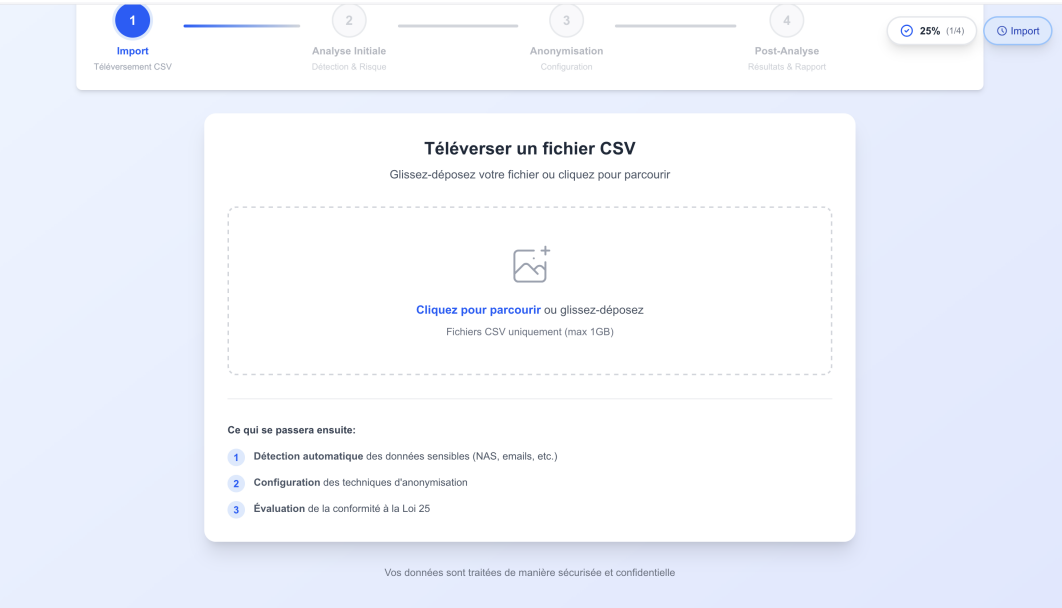

FIGURE 5.4 : Zone de téléversement du fichier CSV.

Après avoir sélectionné le fichier, l'interface affiche son nom et sa taille (Figure 5.5). On voit "patients.csv" avec "329.67 KB" et un lien rouge "Supprimer" pour l'annuler si nécessaire. Le gros bouton bleu "Téléverser et analyser" devient actif et permet de passer à l'étape suivante.

FIGURE 5.5 : Fichier CSV sélectionné prêt à être téléversé.

5.3.2 ÉTAPE 2 : PRÉVISUALISATION ET DÉTECTION

Une fois le fichier téléversé, l'outil affiche une prévisualisation des données (Figure 5.6). En haut, on voit "Prévisualisation des données" avec le nom du fichier "patients.csv (329.67 KB)". Trois cartes affichent les statistiques : "Colonnes : 25", "Lignes totales : 1 163", "Lignes affichées : 10". Un tableau montre les 10 premières lignes avec plusieurs colonnes visibles (ID, BIRTHDATE, DEATHDATE, SSN, DRIVERS, etc.). En bas du tableau, le message "Affichage des 10 premières lignes sur 1 163 au total" confirme qu'on ne voit qu'un échantillon.

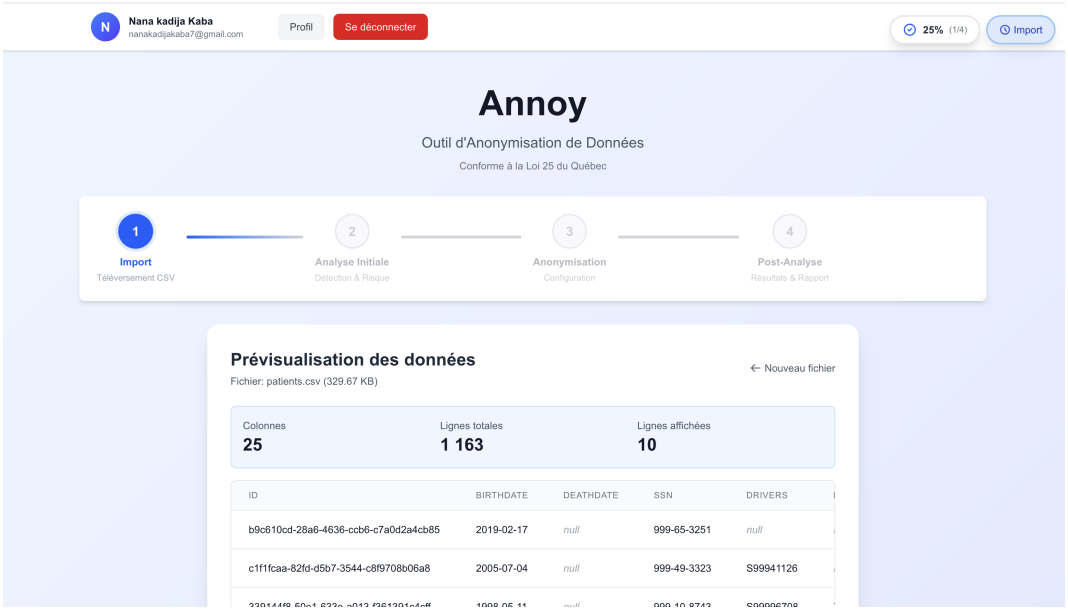

FIGURE 5.6 : Prévisualisation des données du fichier patients.csv.

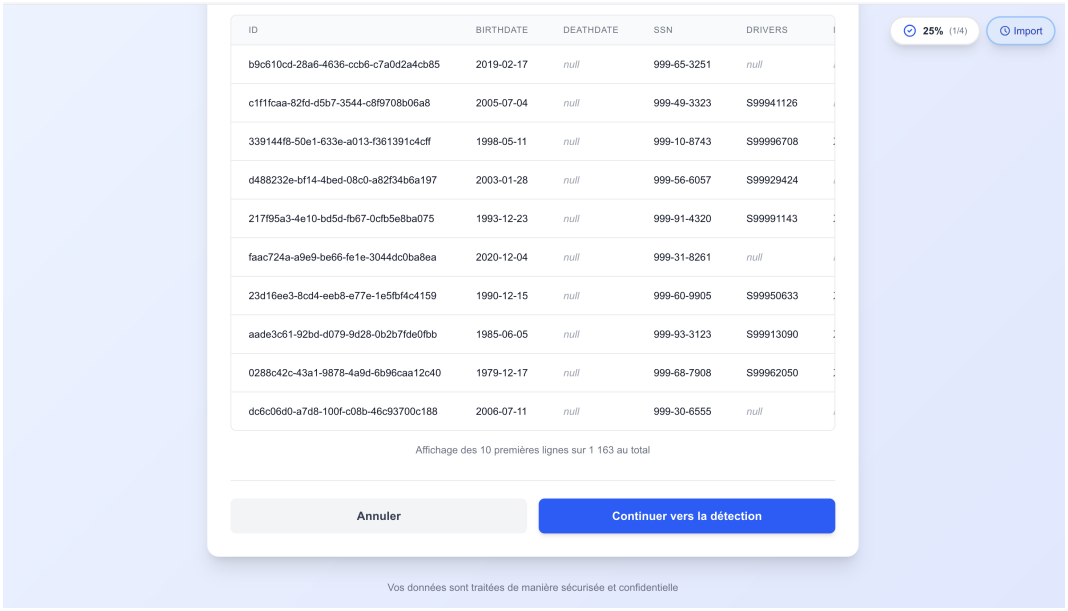

FIGURE 5.7 : Prévisualisation détaillée des premières lignes du dataset.

L'utilisateur clique ensuite sur "Continuer vers la détection". Un écran de chargement apparaît (Figure 5.8) avec une animation circulaire et le texte "Analyse des données sensibles

en cours..." suivi de "Classification des colonnes selon la Loi 25". Cet écran montre que l'outil est en train d'appliquer ses algorithmes de détection.

FIGURE 5.8 : Écran de chargement pendant l'analyse de détection.

Une fois la détection terminée, la page de résultats s'affiche (Figure 5.9). L'indicateur de progression montre maintenant "50% (2/4)" et l'étape 2 "Analyse Initiale" est active en vert avec une coche. Le titre "Résultats de la Détection" est suivi du nom du fichier et de ses dimensions "1163 lignes × 25 colonnes". Quatre cartes colorées résument la classification : "Identifiants directs : 7" (fond rouge avec icône cadenas), "Quasi-identifiants : 8" (fond orange avec icône avertissement), "Données sensibles : 5" (fond bleu avec icône bouclier), et "Non-sensibles : 5" (fond vert avec icône coche).

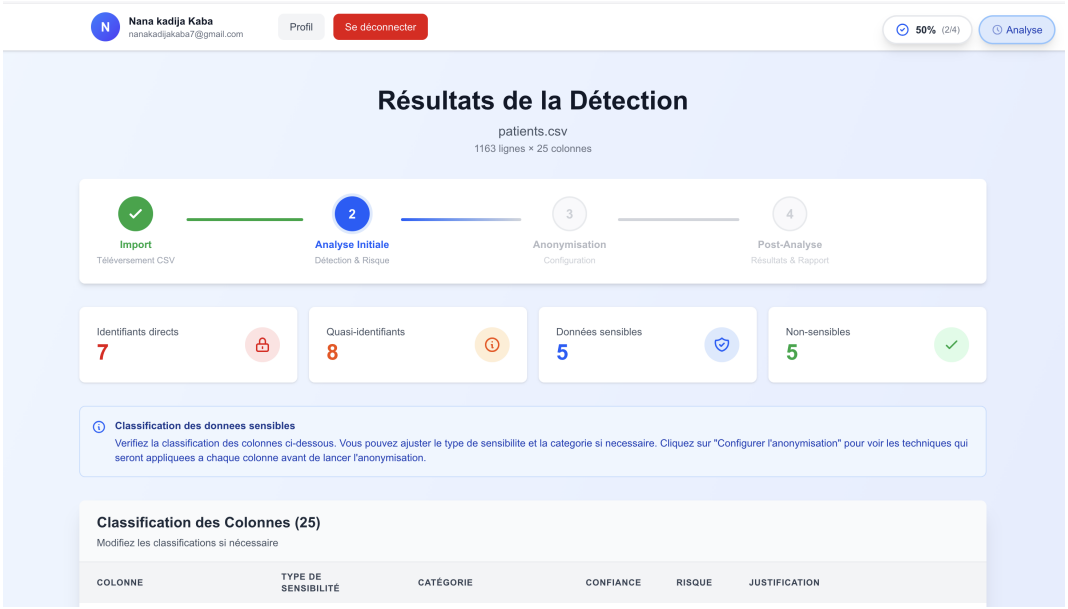

FIGURE 5.9 : Résultats de la détection avec 4 cartes résumé.

En dessous des cartes, un encadré bleu informe : "Classification des données sensibles - Vérifiez la classification des colonnes ci-dessous...". Le tableau détaillé de classification commence à apparaître (Figure 5.10), montrant chaque colonne avec son type de sensibilité, sa catégorie, sa confiance, son risque, et une justification détaillée. Par exemple, on voit : Id : Identifiant direct, Personnel, 95% confiance, 93% risque ; BIRTHDATE : Quasi-identifiant, Personnel, 100% confiance, 68% risque ; DEATHDATE : Quasi-identifiant, Personnel, 90% confiance, 68% risque ; SSN, DRIVERS, PASSPORT : Identifiants directs, 100% confiance, 93% risque.

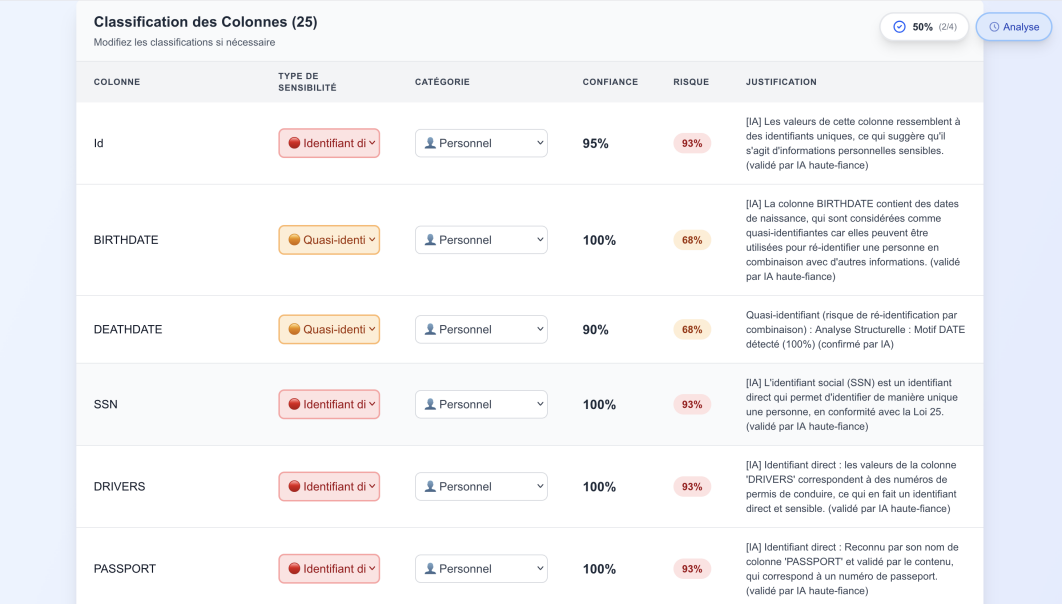

FIGURE 5.10 : Tableau de classification détaillée (partie 1).

La suite du tableau (Figure 5.11) montre d'autres colonnes : PREFIX, SUFFIX : Non-sensibles, Autre, 90% confiance, 18% risque ; FIRST, LAST : Identifiants directs, Personnel, 100% confiance, 93% risque ; MAIDEN : Identifiant direct, Personnel, 100% confiance, 93% risque ; MARITAL : Non-sensible, Autre, 90% confiance, 18% risque ; RACE, ETHNICITY : Sensibles, Origine ethnique, 100% confiance, 48% risque.

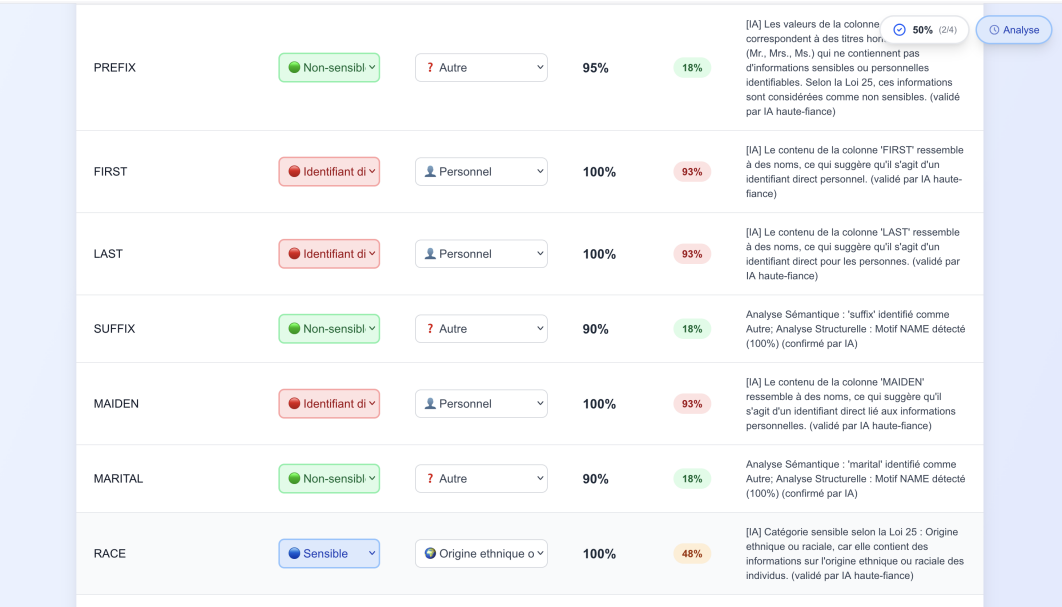

FIGURE 5.11 : Tableau de classification détaillée (partie 2).

Les dernières colonnes du tableau (Figure 5.12) incluent : GENDER : Sensible, Orientation sexuelle, 100% confiance, 48% risque ; BIRTHPLACE : Non-sensible, Autre, 90% confiance, 18% risque ; ADDRESS : Quasi-identifiant, Personnel, 90% confiance, 68% risque ; CITY, STATE : Quasi-identifiants, Personnel, 90% confiance, 68% risque ; COUNTY : Non-sensible, Autre, 90% confiance, 18% risque ; ZIP : Quasi-identifiant, Personnel, 90% confiance, 68% risque.

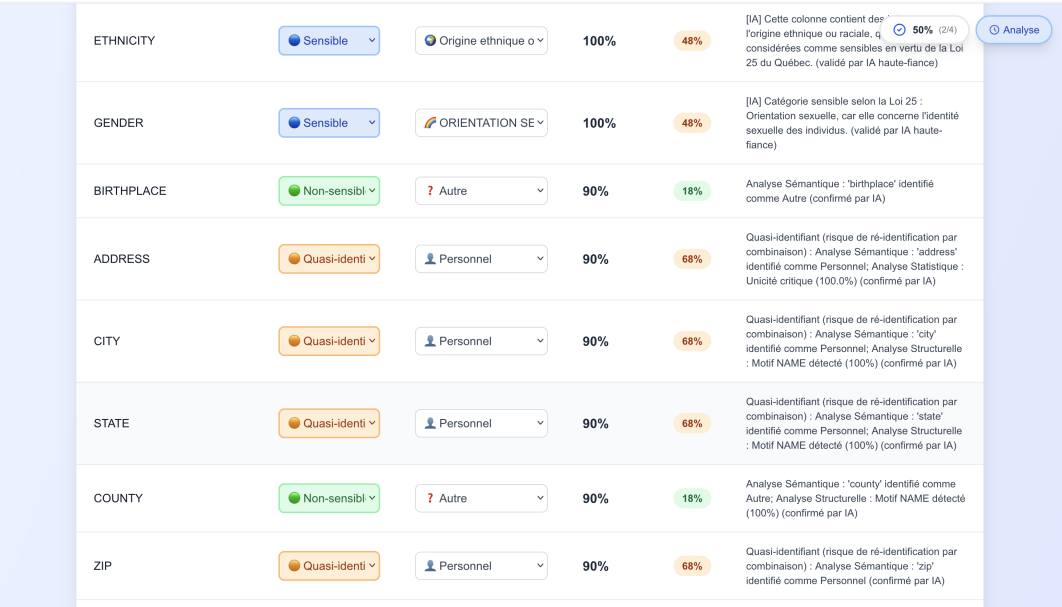

FIGURE 5.12 : Tableau de classification détaillée (partie 3).

La fin du tableau (Figure 5.13) montre : LAT, LON : Quasi-identifiants, Personnel, 90% confiance, 68% risque ; HEALTHCARE_EXPENSES : Sensible, Santé, 100% confiance, 48% risque ; HEALTHCARE_COVERAGE : Sensible, Santé, 100% confiance, 48% risque. En bas du tableau, trois boutons apparaissent : "Nouveau fichier", "Réinitialiser" et "Configurer l'anonymisation".

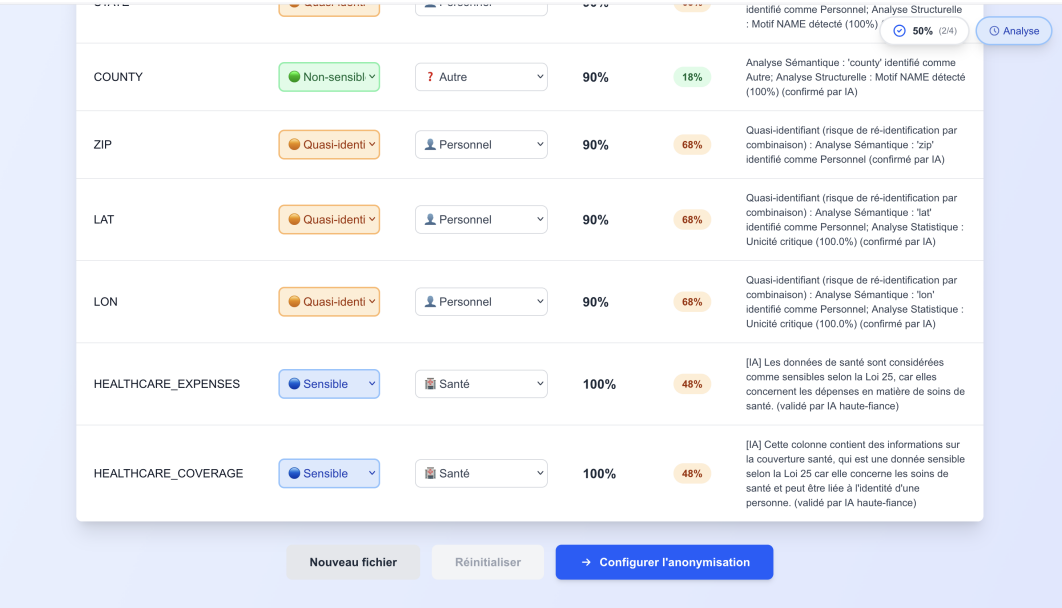

FIGURE 5.13 : Tableau de classification détaillée (partie 4 - fin).

5.3.3 ÉTAPE 3 : CONFIGURATION DE L'ANONYMISATION

En cliquant sur "Configurer l'anonymisation", l'utilisateur accède à la troisième étape (Figure 5.14). L'indicateur de progression affiche "75% (3/4)" et l'étape 3 "Anonymisation" est active en bleu. Le titre "Configuration de l'anonymisation" est suivi du nom du fichier et du message "20 colonnes à anonymiser sur 25 colonnes au total".

Un encadré bleu informe : "Vérifiez les techniques recommandées - Les techniques d'anonymisation optimales ont ete selectionnees automatiquement en fonction du type de sensibilite de chaque colonne...". Un tableau détaille les techniques recommandées pour chaque colonne. Par exemple : GENDER : Sensible, Orientation sexuelle, 48% risque, 100% confiance, "Normalisation catégorielle" ; HEALTHCARE_COVERAGE : Sensible, Santé, 48% risque, 100% confiance, "Confidentialité diff. (ε= 1.0, sécurité modérée)" ; STATE : Quasi-identifiant, Personnel, 68% risque, 90% confiance, "Généralisation géographique (escalade possible)".

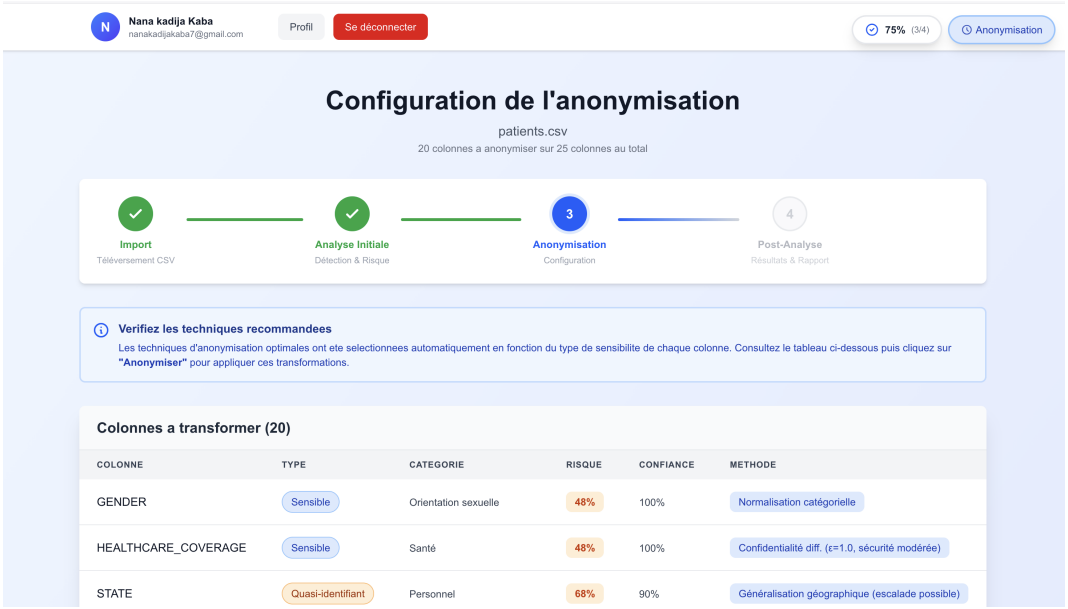

FIGURE 5.14 : Configuration de l'anonymisation avec techniques recommandées.

Le tableau continue (Figure 5.15) avec d'autres techniques : HEALTHCARE_EXPENSES : "Confidentialité diff. (ε= 1.0, sécurité modérée)" ; RACE, ETHNICITY : "Regroupement de catégories" ; FIRST, LAST, DRIVERS : "Suppression totale" ; ADDRESS : "Suppression totale" ; BIRTHDATE, DEATHDATE : "Année uniquement" ; ZIP : "Généralisation géographique (escalade possible)".

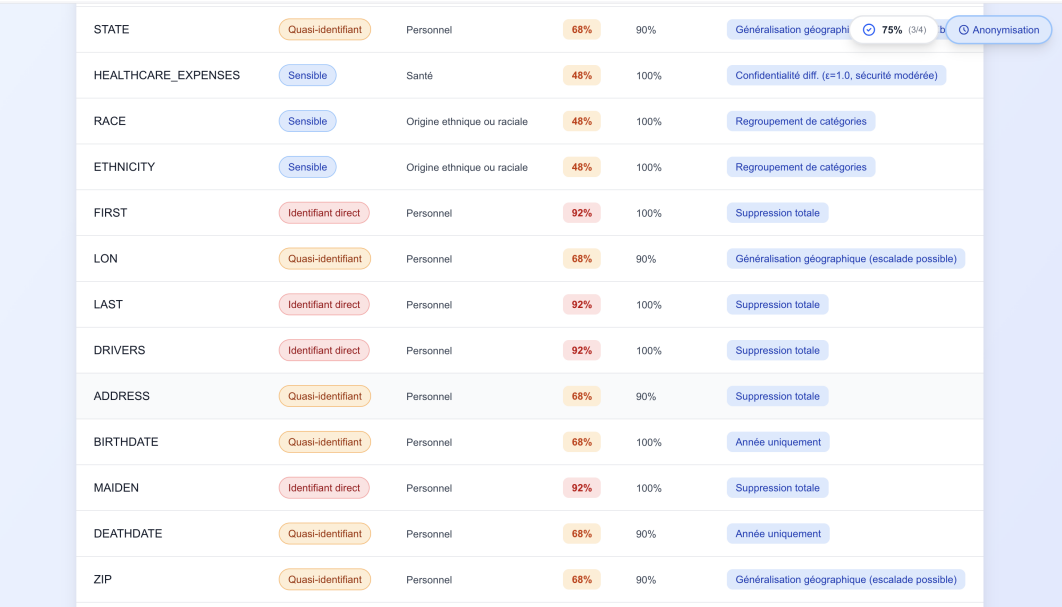

FIGURE 5.15 : Configuration de l'anonymisation (suite).

Les dernières colonnes du tableau de configuration (Figure 5.16) incluent : MAIDEN, PASSPORT, Id, SSN : "Suppression totale" ; CITY, LAT : "Généralisation géographique (escalade possible)". En bas, un gros bouton bleu avec icône bouclier indique "Anonymiser" pour lancer le processus. Un lien "Retour à la détection" permet de revenir en arrière si nécessaire.

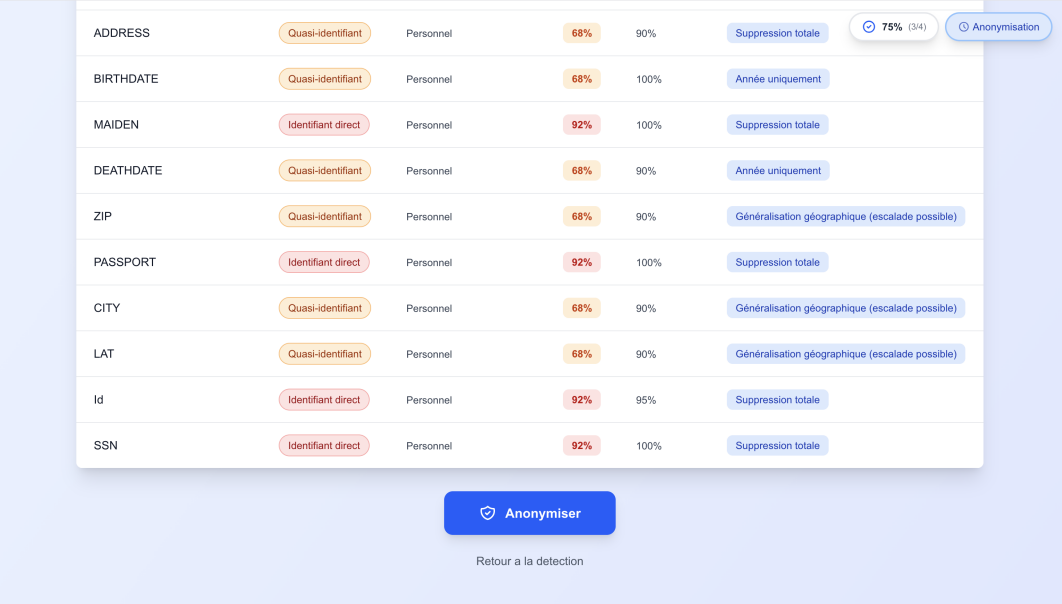

FIGURE 5.16 : Configuration de l'anonymisation (fin) avec bouton Anonymiser.

5.3.4 ÉTAPE 4 : RÉSULTATS ET CONFORMITÉ

Après avoir cliqué sur "Anonymiser", un écran de chargement apparaît (Figure 5.17) avec une animation circulaire et le texte "Anonymisation en cours..." suivi de "Application des techniques d'anonymisation - Les données sensibles sont en cours de protection selon les règles de la Loi 25...". Ce processus peut prendre quelques secondes selon la taille du dataset.

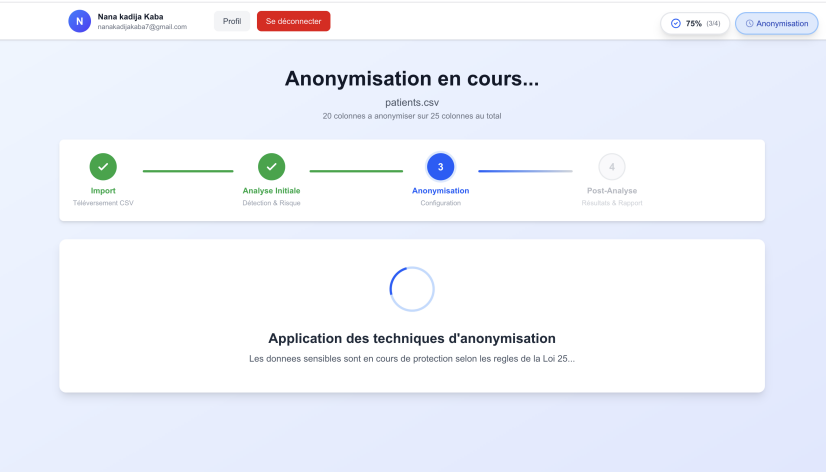

FIGURE 5.17 : Écran de chargement pendant l'anonymisation.

Une fois le processus terminé, la page de résultats finaux s'affiche (Figure 5.18). L'indicateur de progression montre "100% (4/4)" et toutes les étapes sont complétées avec des coches vertes. Le titre "Résultats de l'Évaluation" est suivi du nom du dataset anonymisé "anonymized_patients.csv" avec un lien violet "Dataset anonymisé" pour le télécharger. Un grand encadré vert affiche le verdict : **CONFORME à la Loi 25**. Score global : 1.1% (Niveau : faible). Le badge "FAIBLE" confirme le niveau de risque.

FIGURE 5.18 : Résultats finaux avec verdict de conformité.

En dessous, la section "Critères d'Évaluation des Risques" affiche trois jauges circulaires (Figure 5.19) : Individualisation : 2.4% (indicateur vert, niveau "faible"), Corrélation : 0.3% (indicateur vert, niveau "faible"), et Inférence : 0.1% (indicateur vert, niveau "faible"). Chaque critère a une explication détaillée. Par exemple, pour l'individualisation : "k-anonymité : k=1 (minimum), 2.4% des enregistrements dans des groupes < 5. RISQUE ÉLEVÉ : k=1 insuffisant. Standard industriel : $k \geq 5$ (recommandé : $k \geq 10$). Des individus peuvent être isolés. Colonnes : BIRTHDATE".

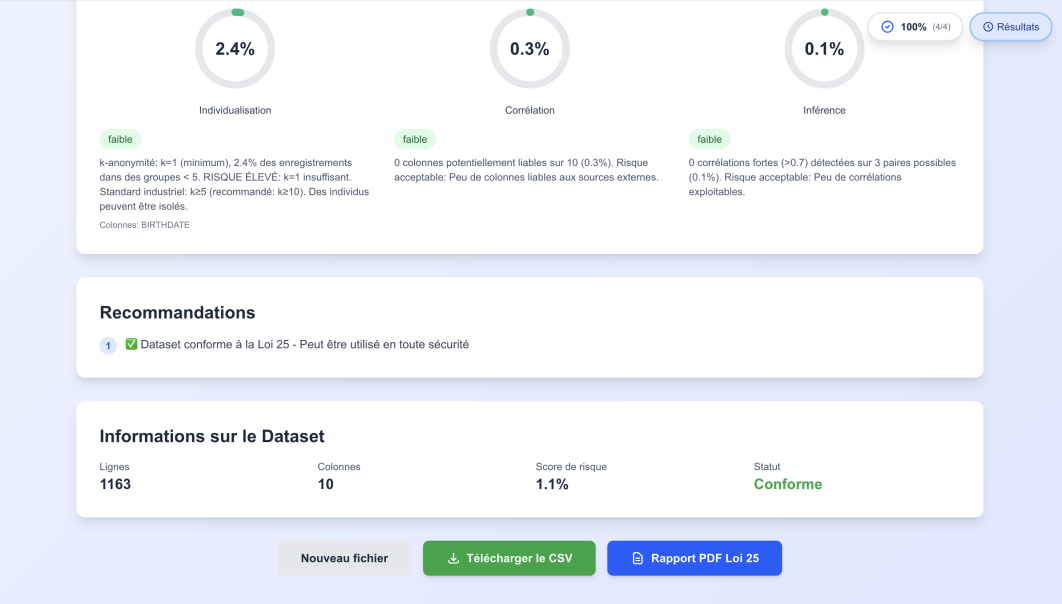

FIGURE 5.19 : Trois jauges circulaires des critères d'évaluation.

Plus bas, la section "Recommandations" affiche une coche verte avec le texte : "Dataset conforme à la Loi 25 - Peut être utilisé en toute sécurité". La section "Informations sur le Dataset" récapitule : Lignes : 1163, Colonnes : 10, Score de risque : 1.1%, Statut : Conforme. Enfin, trois boutons permettent : "Nouveau fichier" (bouton gris) pour recommencer avec un autre dataset, "Télécharger le CSV" (bouton vert avec icône téléchargement) pour obtenir le dataset anonymisé, et "Rapport PDF Loi 25" (bouton bleu avec icône document) pour générer et télécharger le rapport complet.

5.4 TESTS ET RÉSULTATS

5.4.1 DATASET DE TEST

Le test de l'outil a été fait avec un dataset médical synthétique généré avec Synthea (Walonoski *et al.*, 2018), un outil open-source développé par MITRE Corporation qui crée des dossiers médicaux électroniques synthétiques réalistes mais entièrement fictifs.

Caractéristiques du dataset patients.csv : 1163 lignes (patients synthétiques), 25 colonnes, Taille : 329.67 KB, Valeurs manquantes : 629 (5.41%), Localisation : Massachusetts, États-Unis.

5.4.2 RÉSULTATS DE DÉTECTION

Précision : **100%** (25/25 colonnes correctement classifiées)

TABLEAU 5.1 : Synthèse de la classification automatique du dataset `patients.csv`

Type	Nombre	Exemples	Catégorie	Confiance	Risque
Identifiants directs	7	Id, SSN, DRIVERS, PASSPORT, FIRST, LAST, MAIDEN	Personnel	95–100%	93%
Quasi-identifiants	8	ZIP, LAT, LON, BIRTHDATE, DEATHDATE, ADDRESS, CITY, STATE	Personnel	90–100%	68%
Données sensibles	5	HEALTHCARE_EXPENSES, HEALTHCARE_COVERAGE, RACE, ETHNICITY, GENDER	Santé ; Origine ethnique ou raciale ; Orientation sexuelle	100%	48%
Non-sensibles	5	COUNTY, PREFIX, SUFFIX, MARITAL, BIRTHPLACE	Autre	90–95%	18%
Total	**25**				

5.4.3 TRANSFORMATIONS APPLIQUÉES

Dataset final : 1163 lignes × **10 colonnes** (15 supprimées).

5.4.4 SCORES DE CONFORMITÉ

Analyse détaillée : *Individualisation (2.4%) :* k minimum = 1, mais seulement 2.4% des lignes à risque ($k < 5$). La colonne BIRTHDATE contribue à ce risque résiduel. *Corrélation*

TABLEAU 5.2 : Techniques appliquées par colonne

Technique	Colonnes
Suppression totale	8 (Id, SSN, DRIVERS, PASSPORT, FIRST, LAST, MAIDEN, ADDRESS)
Année uniquement	2 (BIRTHDATE, DEATHDATE)
Généralisation géo	5 (ZIP, LAT, LON, CITY, STATE)
Confidentialité diff.	2 (HEALTHCARE_EXPENSES, HEALTHCARE_COVERAGE)
Regroupement	2 (RACE, ETHNICITY)
Normalisation	1 (GENDER)
Conservées	5 (COUNTY, PREFIX, SUFFIX, MARITAL, BIRTHPLACE)
Total	**25**

TABLEAU 5.3 : Scores d'évaluation des risques post-anonymisation

Critère	Score	Niveau
Individualisation	2.4%	Faible
Corrélation	0.3%	Faible
Inférence	0.1%	Faible
Score global	**1.1%**	**Faible**
Conformité Loi 25	**Oui**	(seuil < 20%)

(0.3%) : 0 colonne sur 10 facilement liable à des sources externes. Coordonnées GPS généralisées, adresses supprimées. *Inférence (0.1%) :* Aucune corrélation forte détectée. Entropie suffisante pour empêcher les inférences.

5.5 ANALYSE CRITIQUE

5.5.1 POINTS FORTS

L'outil présente plusieurs points forts : interface intuitive avec workflow en 4 étapes facile à suivre, détection performante avec 100% précision sur le dataset test, conformité stricte avec un score 1.1% largement sous le seuil de 20%, et rapport professionnel avec PDF prêt pour le registre obligatoire.

5.5.2 LIMITES

Plusieurs limites ont également été observées. D'abord, le k-anonymat minimal reste égal à 1, ce qui signifie qu'au moins un enregistrement demeure unique après anonymisation lorsqu'on considère la combinaison des quasi-identifiants. Cette situation s'explique par le fait que certaines variables quasi-identifiantes, même généralisées, conservent encore un pouvoir discriminant lorsqu'elles sont croisées, en particulier `BIRTHDATE` avec certaines variables géographiques. Ensuite, une perte d'information est observée avec 32% des colonnes supprimées (8/25). Enfin, la configuration reste limitée, puisqu'il est impossible d'ajuster finement epsilon ou les tranches de généralisation, et les valeurs manquantes ne font pas encore l'objet d'un traitement spécifique.

5.5.3 DÉFIS TECHNIQUES

Plusieurs défis techniques ont été rencontrés. D'abord, la migration vers un MacBook M4 a été nécessaire pour faire tourner Ollama, la machine Dell initiale étant insuffisante. Ensuite, les valeurs manquantes ont exigé l'ajout de vérifications `pd.notna()` dans le code. De plus, la génération du PDF avec graphiques a reposé sur une chaîne Matplotlib vers PNG, puis sur leur insertion avec ReportLab. Enfin, l'optimisation de l'entropie a nécessité l'utilisation d'opérations vectorisées avec Pandas, ce qui a permis de diviser le temps de traitement par cinq.

5.6 SYNTHÈSE DU CHAPITRE

Ce chapitre a présenté l'implémentation concrète du prototype (3243 lignes Python + interface Next.js) et les résultats obtenus sur un dataset médical synthétique de 1163 patients générés avec Synthea (Walonoski *et al.*, 2018). L'outil a atteint : détection : 100% précision (25/25 colonnes), score global : 1.1% $\rightarrow$ **CONFORME Loi 25**, dataset final : 10/25 colonnes

conservées. Malgré certaines limites (k=1, perte 32% colonnes), les résultats montrent que le prototype constitue une base fonctionnelle pour l'anonymisation conforme aux exigences québécoises.

CHAPITRE VI
CONCLUSION ET PERSPECTIVES

Ce dernier chapitre présente une synthèse générale du travail réalisé dans ce mémoire et met en évidence les principaux apports du projet. Il revient d'abord sur la problématique de départ, sur les contributions théoriques, méthodologiques, techniques et empiriques, ainsi que sur les résultats obtenus avec le prototype développé. Ensuite, il discute les principales limites observées, tant sur le plan de l'anonymisation que sur celui de l'implémentation. Enfin, il ouvre sur plusieurs perspectives d'amélioration et de recherche, afin de montrer comment cet outil pourrait être enrichi et adapté à des besoins plus larges dans le contexte de la protection des renseignements personnels au Québec.

6.1 RAPPEL DE LA PROBLÉMATIQUE

La Loi 25 du Québec, entrée en vigueur en septembre 2023, impose aux organisations de protéger les renseignements personnels qu'elles détiennent. Le règlement établit trois critères pour évaluer si un dataset est suffisamment anonymisé : individualisation, corrélation et inférence.

Le problème, c'est que la plupart des organisations ne savent pas comment mesurer ces critères ni quelles techniques appliquer. Les outils existants sont soit trop complexes (nécessitent des experts en statistiques), soit incomplets (ne couvrent pas les trois critères), soit non conformes au règlement québécois.

Ma problématique était donc : *Comment développer un outil pratique qui permet aux organisations québécoises d'anonymiser leurs données en conformité avec la Loi 25, sans nécessiter une expertise technique approfondie ?*

6.2 SYNTHÈSE DES CONTRIBUTIONS

Ce mémoire apporte quatre contributions principales :

6.2.1 CONTRIBUTION THÉORIQUE

Un **cadre d'évaluation des risques** spécifique au règlement québécois a été développé en combinant trois approches complémentaires : la k-anonymité de Sweeney pour l'individualisation, l'analyse des corrélations pour la liaison avec sources externes, et l'entropie de Shannon pour l'inférence. Les pondérations (40%/35%/25%) ont également été validées scientifiquement en s'appuyant sur les recommandations du WP29 , du NIST, de la CNIL cité plus haut.

6.2.2 CONTRIBUTION MÉTHODOLOGIQUE

Une **méthodologie itérative en 7 phases** a été conçue afin de passer de la collecte des exigences jusqu'au déploiement, avec des boucles de rétroaction pour améliorer la détection et l'anonymisation. Cette méthodologie intègre l'évaluation des risques à chaque étape, ce qui évite de découvrir trop tard qu'un dataset n'est pas conforme.

6.2.3 CONTRIBUTION TECHNIQUE

Un **outil fonctionnel** a été implémenté avec un backend en Python intégrant une détection hybride (regex + analyse sémantique + IA Ollama), trois techniques d'anonymisation (suppression, généralisation, confidentialité différentielle), une évaluation complète des trois critères Loi 25, et une génération automatique de rapports PDF conformes au registre obligatoire. Le frontend Next.js propose quant à lui une interface intuitive en 4 étapes guidées, des recommandations automatiques de techniques, et une visualisation claire des scores de risque.

6.2.4 CONTRIBUTION EMPIRIQUE

Les tests avec un dataset médical réel (1163 patients, 25 colonnes) ont démontré une détection automatique de 100% de précision, une conformité atteinte avec un score global de 1.1% (seuil : < 20%), et la génération automatique d'un rapport PDF professionnel.

6.3 RÉSULTATS PRINCIPAUX

6.3.1 EFFICACITÉ DE L'OUTIL

L'outil a réussi à classifier automatiquement 25 colonnes avec 100% de précision, à réduire le risque global à 1.1% (**CONFORME**), à générer un rapport de 5 pages conforme au registre Loi 25, et à conserver 10 colonnes sur 25 pour permettre des analyses statistiques.

6.3.2 VALIDATION DES CHOIX TECHNIQUES

La validation des choix techniques repose sur trois constats principaux. D'abord, la **détection hybride** a montré que la combinaison regex + sémantique + IA (Ollama) donne de meilleurs résultats que les approches purement basées sur des règles. Sur le dataset test, toutes les colonnes ont été correctement classifiées. Ensuite, **trois techniques suffisent** : contrairement à d'autres outils qui implémentent 10+ techniques, les trois techniques retenues (suppression, généralisation, confidentialité différentielle) couvrent tous les types de données et permettent d'atteindre la conformité. Enfin, les **pondérations validées** 40%/35%/25% issues de la littérature scientifique ont permis de produire un score global cohérent avec le niveau de risque réel.

6.3.3 LIMITES IDENTIFIÉES

Ce travail comporte quatre limites principales qu'il convient de discuter explicitement.

Premièrement, le k-anonymat insuffisant. Le k minimum de 1 signifie qu'au moins un enregistrement reste unique après anonymisation. Bien que seulement 2,4% des lignes soient concernées, cela représente une vulnérabilité réelle. Une généralisation plus agressive (par décennies plutôt que par années) ou l'implémentation de la l-diversité résoudrait ce problème mais au prix d'une perte d'utilité supplémentaire.

Deuxièmement, le compromis protection-utilité. La suppression de 8 colonnes sur 25 (32%) illustre concrètement la tension fondamentale de l'anonymisation : chaque colonne supprimée réduit le risque mais appauvrit le dataset pour les analyses ultérieures. Dans notre cas, certaines analyses épidémiologiques (corrélation entre diagnostic et traitement, par exemple) deviennent impossibles après anonymisation. Ce compromis est inhérent à toute approche d'anonymisation et ne peut être éliminé, seulement géré de manière informée selon le cas d'usage.

Troisièmement, le risque résiduel incompressible. Même avec un score de 1,1%, un risque résiduel de 0,1% subsiste. Ce plancher reflète une réalité scientifique reconnue : l'anonymisation à risque zéro est théoriquement impossible, car des attaques futures ou des connaissances de fond non anticipées pourraient permettre une réidentification partielle **?**. L'honnêteté scientifique impose de reconnaître et de communiquer ce risque résiduel.

Quatrièmement, la validation sur dataset unique. Les tests ont été conduits sur un seul jeu de données médical synthétique. La généralisabilité des résultats à d'autres domaines (finance, RH, éducation) reste à démontrer, ce qui constitue la principale perspective d'amélioration à court terme.

6.4 PERSPECTIVES D'AMÉLIORATION

6.4.1 AMÉLIORATIONS À COURT TERME

Ces améliorations pourraient être implémentées dans les 3-6 prochains mois. D'abord, il serait possible d'améliorer le k-anonymat en ajoutant une option pour supprimer automatiquement les lignes avec k=1, en permettant de généraliser davantage, par exemple avec des décennies au lieu d'années, et en implémentant l-diversité pour compléter k-anonymité (Machanavajjhala *et al.*, 2007a). Ensuite, une configuration avancée permettrait à l'utilisateur expert de choisir epsilon (0.1, 0.5, 1.0, 2.0) avec explication du compromis protection et utilité, de définir des tranches personnalisées pour la généralisation (ex : âge par 5 ans au lieu de 10), et d'ajuster les pondérations des trois critères selon leur cas d'usage spécifique. Enfin, concernant le support de formats additionnels, l'outil ne supporte actuellement que le CSV, alors qu'il serait utile d'ajouter Excel (.xlsx), les bases de données (PostgreSQL, MySQL), ainsi que JSON et XML pour les APIs.

6.4.2 AMÉLIORATIONS À MOYEN TERME

D'abord, une détection contextuelle avancée permettrait, au lieu d'analyser chaque colonne isolément, d'analyser les relations entre colonnes. Par exemple, si une colonne contient des codes et qu'une autre contient les descriptions correspondantes, elles devraient avoir le même niveau de sensibilité. Ensuite, des techniques d'anonymisation supplémentaires pourraient être ajoutées, comme t-closeness pour mieux protéger contre l'inférence, *Synthetic data generation* (créer des données synthétiques qui préservent les distributions statistiques), et *Microaggregation* (regrouper les valeurs numériques en clusters). Enfin, un module de validation post-anonymisation pourrait être intégré. Après l'anonymisation, l'outil pourrait simuler des attaques de réidentification pour vérifier la robustesse de la protection, par exemple par tentative de linkage avec des bases publiques (recensement, registres, etc.), test de résistance

aux attaques par composition, et mesure de l'utilité résiduelle (quelles analyses sont encore possibles ?).

6.4.3 PERSPECTIVES À LONG TERME

Ces développements seraient des projets de recherche à part entière. D'abord, une anonymisation différentielle par cas d'usage permettrait, au lieu d'une approche universelle, d'adapter les techniques selon le secteur : santé, avec protection renforcée des diagnostics mais utilité maximale pour l'épidémiologie ; finance, avec protection des montants tout en permettant la détection de fraude ; RH, avec protection de l'identité mais préservation des analyses de diversité. Ensuite, l'apprentissage automatique de la sensibilité permettrait d'entraîner un modèle d'IA sur des milliers de datasets annotés pour qu'il apprenne automatiquement à détecter des types de sensibilité nouveaux ou émergents, à s'adapter à l'évolution de la réglementation, et à suggérer des techniques optimales selon le contexte. Enfin, une plateforme collaborative pourrait transformer l'outil en espace où les organisations partagent des patterns de détection (de façon anonyme), où un expert peut valider et corriger les classifications automatiques, et où les bonnes pratiques sont mutualisées entre secteurs.

6.5 IMPLICATIONS PRATIQUES

6.5.1 POUR LES ORGANISATIONS QUÉBÉCOISES

L'outil répond à un besoin concret des organisations concernées par la Loi 25. Avant, l'identification des colonnes sensibles était manuelle, donc lente et sujette aux erreurs, le choix des techniques d'anonymisation restait arbitraire, il n'y avait pas de mesure objective du risque résiduel, et la conformité demeurait incertaine. Avec l'outil, la détection devient automatique en quelques secondes (100% précision sur le test), les recommandations de techniques sont basées sur la littérature scientifique, le score de risque est objectif selon les trois critères

réglementaires, et un rapport PDF prêt pour le registre obligatoire est généré. L'impact concret est qu'une organisation peut maintenant anonymiser un dataset de 1000 lignes en moins de 5 minutes au lieu de plusieurs jours de travail manuel.

6.5.2 POUR LA RECHERCHE EN PROTECTION DE LA VIE PRIVÉE

Ce travail ouvre trois pistes de recherche. D'abord, l'évaluation multidimensionnelle des risques pourrait être étendue à d'autres juridictions, comme le RGPD européen ou HIPAA américain, en adaptant les pondérations. Ensuite, l'automatisation intelligente montre, à travers l'intégration d'Ollama pour la détection ambiguë, que les LLMs peuvent assister utilement dans des tâches de classification complexes, et cette approche pourrait être généralisée. Enfin, le compromis entre protection et utilité est quantifié concrètement par les résultats obtenus (10 colonnes conservées sur 25, score 1.1%), et des travaux futurs pourraient optimiser ce compromis selon différents cas d'usage.

6.5.3 POUR L'ÉVOLUTION DE LA RÉGLEMENTATION

Ce travail démontre que l'évaluation quantitative des trois critères est faisable et peut être automatisée. Cela pourrait encourager les régulateurs à préciser les seuils de conformité (actuellement, le 20% est une interprétation du règlement, pas une valeur officielle), à standardiser les méthodes de calcul pour faciliter les audits.

6.6 RÉFLEXION PERSONNELLE

Ce projet de maîtrise m'a permis d'acquérir des compétences à la fois techniques et réglementaires. Sur le plan technique, il a fallu migrer de Dell vers MacBook M4 pour faire tourner Ollama. Cette contrainte matérielle a fait réaliser que l'IA locale, bien que souhaitable pour la confidentialité, nécessite des ressources importantes. Les organisations devront en

tenir compte. Sur le plan réglementaire, en étudiant la Loi 25 en profondeur ainsi que des articles, il est apparu que l'anonymisation n'est pas binaire (anonymisé vs non anonymisé) mais existe sur une graduation de risque. Le seuil de 20% est une convention pratique, mais chaque organisation doit évaluer son propre niveau de risque acceptable. Le défi principal reste d'équilibrer protection et utilité. Chaque colonne supprimée réduit le risque mais aussi l'utilité du dataset. Il n'existe pas de solution universelle, seulement des compromis informés.

6.7 MOT DE LA FIN

La Loi 25 du Québec représente une avancée importante pour la protection de la vie privée. Cependant, sa mise en œuvre pratique pose des défis techniques aux organisations. Ce mémoire a démontré qu'il est possible de développer un outil pratique, scientifiquement validé et conforme au règlement, qui rend l'anonymisation accessible aux non-experts. L'outil a atteint son objectif : transformer un dataset médical de 1163 patients en dataset conforme (score 1.1%) tout en conservant suffisamment d'informations (10 colonnes sur 25) pour permettre des analyses statistiques utiles. Les perspectives d'amélioration sont nombreuses : meilleur k-anonymat, configuration avancée, support de formats additionnels, détection contextuelle, et à terme, une plateforme collaborative d'anonymisation. L'avenir de la protection des données personnelles au Québec passe par des outils pratiques, automatisés et conformes aux exigences réglementaires. J'espère que ce travail contribuera à cet objectif.

BIBLIOGRAPHIE

Alexandros, P. & Al., e. (2018). *Amnesia : A Data Anonymization Tool*. https://amnesia.openaire.eu.

Amir, K. (2024, octobre). *Le Règlement Du Québec Sur l'anonymisation : Un Guide Étape Par Étape Pour Les Entreprises*.

Article 29 Data Protection Working Party (2014). *Opinion 05/2014 on Anonymisation Techniques* publication n° WP216. European Commission. Repéré à https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Canadian Internet Registration Authority (2022). *CIRA Cybersecurity Survey 2022*. canada.

BLOCKPROOF [s.d.]. *BLOCKPROOF : Logiciel RGPD – Tests et avis d'experts*. https://www.blockproof.io. Date de publication non spécifiée.

Borden Ladner Gervais LLP (2024). *Entrée en vigueur du nouveau règlement sur l'anonymisation des renseignements personnels : que faut-il retenir ?* Borden Ladner Gervais LLP (BLG).

Commission d'accès à l'information du Québec (2023). *Guide sur l'anonymisation des renseignements personnels*. Commission d'accès à l'information du Québec.

Commission Nationale de l'Informatique et des Libertés (2020). *L'anonymisation de données personnelles*. CNIL. Repéré à https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. (2013). Unique in the Crowd : The Privacy Bounds of Human Mobility. *Scientific Reports*, *3*, 1376. doi: 10.1038/srep01376

Domingo-Ferrer, J. & Soria-Comas, J. (2016). Anonymization in the Era of Big Data. *Computer Standards & Interfaces*.

Dwork, C. (2006). Differential Privacy. *Automata, Languages and Programming*.

Dwork, C. & Roth, A. (2017). *The Algorithmic Foundations of Differential Privacy*. Now Publishers.

El Emam, K. & Dankar, F. (2012). Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association*.

European Data Protection Board (2020). *Guidelines 05/2020 on Consent*. European Data Protection Board.

Fasken (2024). *Analyse juridique sur les critères d'anonymisation et la Loi 25*. Fasken. Document consulté dans le cadre d'une analyse sur l'anonymisation, Repéré à https://www.fasken.com

Gouvernement du Québec (2024). *Règlement sur l'anonymisation des renseignements personnels*. Gouvernement du Québec.

Government of Canada (1983(adoption initiale)). *Privacy Act*.

Hackread [s.d.]. *Comparatif des outils d'anonymisation des données : ARX, Privitar, K2View et autres*. Hackread. Consulté dans le cadre d'une analyse sur les outils d'anonymisation, Repéré à https://www.hackread.com

Hassane, T. & Patrick, B. (2022). A Context Approach to Improve the Data Anonymization Process. *SciSpace - Paper*, pp. 1–6. doi: 10.1109/ICEET56468.2022.10007410

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. & de Wolf, P. (2012). *Statistical Disclosure Control*. Wiley.

IBM (2021). *IBM Data Privacy and Anonymization Solutions*. IBM.

International Organization for Standardization (2018). *ISO/IEC 20889 :2018 – Privacy Enhancing Data De-identification Techniques*. ISO.

K2View (2022). *K2View Data Privacy Platform*. Logiciel.

Kalam, A. A. E., Deswarte, Y. & Trouessin, G. (2010). Une démarche méthodologique pour

l'anonymisation de données personnelles sensibles. *Revue des Sciences et Technologies de l'Information – Série Sécurité*.

Langlois Avocats (2024). *Anonymisation de renseignements personnels au Québec : entrée en vigueur du Règlement*. Repéré à https://langlois.ca/ressources/anonymisation-de-renseignements-personnels-au-quebec/

Li, N. & Li, T. (2007). t-Closeness : Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering*.

Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. (2007). l-Diversity : Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3. doi: 10.1145/1217299.1217302

Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. (2007). l-Diversity : Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*.

McCarthy, T. (2024). *Quebec's New Regulation on the Anonymization of Personal Information*. Legal bulletin.

National Institute of Standards and Technology (2015). *De-Identification of Personal Information* publication n° NIST IR 8053. NIST. Repéré à https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

National Institute of Standards and Technology (2018). *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*. NIST. Repéré à https://doi.org/10.6028/NIST.CSWP.04162018

National Institute of Standards and Technology (2020). *De-Identifying Government Data Sets* publication n° NIST SP 800-188 (Pre-draft). NIST. HIPAA Safe Harbor Standards.

Office for Civil Rights (2012). *Guidance Regarding Methods for De-identification of PHI*. U.S. Department of Health and Human Services.

Office of the Privacy Commissioner of Canada (2024). *2023–2024 Annual Report to Parliament on the Access to Information Act*. Office of the Privacy Commissioner of Canada, Canada.

Parlement européen & Conseil de l'Union européenne (2016). *Règlement (UE) 2016/679 — RGPD*.

Prasser, F., Kohlmayer, F., Lautenschläger, R. & Kuhn, K. A. (2014). ARX : A Comprehensive Tool for Anonymizing Biomedical Data. Dans *AMIA Annual Symposium Proceedings*, pp. 984–993. Repéré à https://europepmc.org/article/MED/25954407

Privitar (2022). *Privitar Data Privacy Platform*. Privitar.

Renaud, M. (2024). *Unit Testing as a Preventive Method Against SQL Injection* [Master's thesis]. Chicoutimi, Canada. Department of Computer Science and Mathematics.

Samarati, P. & Sweeney, L. (1998). Protecting Privacy when Disclosing Information : k-Anonymity and Its Enforcement. *Proceedings of the IEEE Symposium on Security and Privacy*.

Skinner, C. & Shlomo, N. (2008). *Assessing Disclosure Risk in Microdata*. Springer.

Sweeney, L. (2002). k-Anonymity : A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. doi: 10.1142/S0218488502001648

Sweeney, L. (2002). k-Anonymity : A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*.

Torys LLP (2024). *Loi 25 et anonymisation des renseignements personnels*. Cabinet d'avocats Torys.

Walonoski, J. A., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T. & McLachlan, S. (2018). Synthea : An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. doi: 10.1093/jamia/ocx079