ConfidensIA : un système hybride de pseudonymisation *fine-grained* pour les documents médico-sociaux français

Auteur : Patrick Danto[1]

Affiliation : [1] Initiative indépendante, France

Contact : patrick.danto@outlook.fr

Statut : Projet en cours de dépôt à l'INPI (marque ConfidensIA)

Date : Novembre 2025

Résumé

Le secteur médico-social français produit massivement des écrits professionnels contenant des données hautement sensibles. Aucun système de pseudonymisation n'est adapté à son vocabulaire, à ses organisations et aux exigences du Code de l'Action Sociale et des Familles (CASF). Cet article présente ConfidensIA, un système hybride combinant un modèle NER CamemBERT optimisé (distillation à 11 couches, *pruning magnitude-based*, quantization FP16), 338 règles expertes et 25 916 entrées de gazetteers. La contribution principale est une taxonomie *fine-grained* de 100 catégories d'entités identifiantes, couvrant établissements médico-sociaux, organismes publics, identifiants, adresses, associations et unités de service.

Sur un corpus *gold standard* de 330 phrases (448 entités), ConfidensIA obtient un F1 global de 86,1%, et 95,6% sur les entités critiques (NIR, dates de naissance, adresses complètes). Le modèle distillé (*student*) est publié sous licence MIT, construit à partir du modèle *teacher* Jean-Baptiste/camembert-ner [1]. Le pipeline complet demeure privé pour raisons de développement commercial.

Mots-clés : Named Entity Recognition, De-identification, French NLP, Social Services, GDPR Compliance, Hybrid Systems, Fine-grained Taxonomy

1. Introduction

Le secteur médico-social français regroupe des établissements et services accompagnant des publics vulnérables (personnes en situation de handicap, protection de l'enfance, personnes âgées, précarité). Les professionnels produisent quotidiennement des écrits professionnels (rapports sociaux, synthèses de situation, notes d'observation, évaluations) contenant des informations extrêmement sensibles : identité, adresses, situations familiales, parcours de vie, pathologies, mesures de protection.

Le Règlement Général sur la Protection des Données (RGPD) impose une pseudonymisation fiable avant tout partage, analyse ou utilisation à des fins de recherche [2]. Cette obligation se heurte à l'absence d'outils numériques spécialisés dans le secteur médico-social, contrairement au domaine hospitalier qui dispose de solutions dédiées [3, 4].

Les systèmes existants de reconnaissance d'entités nommées (NER) et de dé-identification sont principalement orientés vers : le domaine clinique hospitalier (i2b2, MIMIC) [5, 6], des taxonomies génériques (PER, ORG, LOC), la langue anglaise [7, 8], ou des *frameworks* peu spécialisés (Presidio, Philter) [9, 10].

Aucun travail ne propose de taxonomie *fine-grained* pour les entités médico-sociales françaises (EHPAD, MECS, SESSAD, ASE, MDPH, CCAS, etc.), ni de système de pseudonymisation adapté à ces écrits professionnels et au vocabulaire du Code de l'Action Sociale et des Familles. ConfidensIA vise à combler cette lacune en proposant la première solution de pseudonymisation spécifiquement conçue pour le secteur médico-social français.

2. Travaux connexes

2.1 NER pour le français et le domaine médical

Les modèles de langage pré-entraînés pour le français ont connu des avancées significatives avec CamemBERT [1] et son adaptation au domaine médical avec DrBERT [11]. Ces modèles montrent d'excellentes performances sur des tâches génériques de NER, mais leur taxonomie reste limitée aux catégories standards (PER, ORG, LOC, MISC).

Pour le domaine médical francophone, des efforts spécifiques ont été déployés sur les dossiers cliniques [12, 13], mais ces travaux se concentrent sur la terminologie médicale (diagnostics, traitements, examens) et non sur les organisations et établissements du secteur social.

2.2 Systèmes de dé-identification

La dé-identification de textes médicaux est un domaine mature en anglais [7, 14], avec des *challenges* dédiés (i2b2/UTHealth 2014) [5] et des corpus annotés publics (MIMIC-III) [6]. Les approches dominantes combinent apprentissage profond et règles expertes [15, 16].

Pour le français, les travaux demeurent limités [17, 18] et ne couvrent pas les spécificités du secteur médico-social : établissements spécialisés (IME, ITEP, MECS, CHRS), organismes départementaux (ASE, MDPH, PCH), dispositifs réglementaires (APA, AAH, AEEH), et acteurs associatifs du secteur.

2.3 Approches hybrides et gazetteers

L'efficacité des architectures hybrides combinant modèles neuronaux, règles expertes et gazetteers a été démontrée dans plusieurs domaines [19, 20]. Ces approches permettent de compenser les limites des modèles purement statistiques face à des entités rares ou des structures syntaxiques spécifiques. ConfidensIA s'inscrit dans cette tradition méthodologique en l'adaptant au contexte réglementaire français et aux spécificités terminologiques du Code de l'Action Sociale et des Familles.

3. Taxonomie fine-grained des 100 catégories

3.1 Construction

La taxonomie a été élaborée sur six mois avec l'appui de professionnels du secteur médico-social (travailleurs sociaux, cadres, directeurs d'établissement). La méthodologie a consisté en :

• Analyse de 50+ documents réels anonymisés (rapports sociaux, synthèses, notes d'évaluation)

• Identification des entités récurrentes et de leurs variations

• Définition d'une première version de taxonomie (62 catégories)

• Validation itérative sur corpus test

• Extension progressive jusqu'à 100 catégories sur 8 itérations

Chaque catégorie a été validée par au moins deux professionnels du secteur pour s'assurer de sa pertinence opérationnelle.

3.2 Organisation

Les 100 catégories sont réparties en huit groupes fonctionnels :

• **Établissements médico-sociaux :** 37 catégories (MECS, IME, ITEP, SESSAD, SAVS, SAMSAH, FAM, MAS, EHPAD, ESAT, etc.)

• **Organismes publics :** 21 catégories (Conseil Départemental, ASE, MDPH, CCAS, CAF, CPAM, Pôle Emploi, Éducation Nationale, etc.)

• **Associations :** 13 catégories (protection de l'enfance, insertion, handicap, logement, aide alimentaire, etc.)

• **Identifiants/Coordonnées :** 11 catégories (NIR, numéro allocataire CAF, téléphone, email, identifiant MDPH, etc.)

• **Adresses/Localisations :** 9 catégories (adresse complète, ville, département, quartier, bailleur social, etc.)

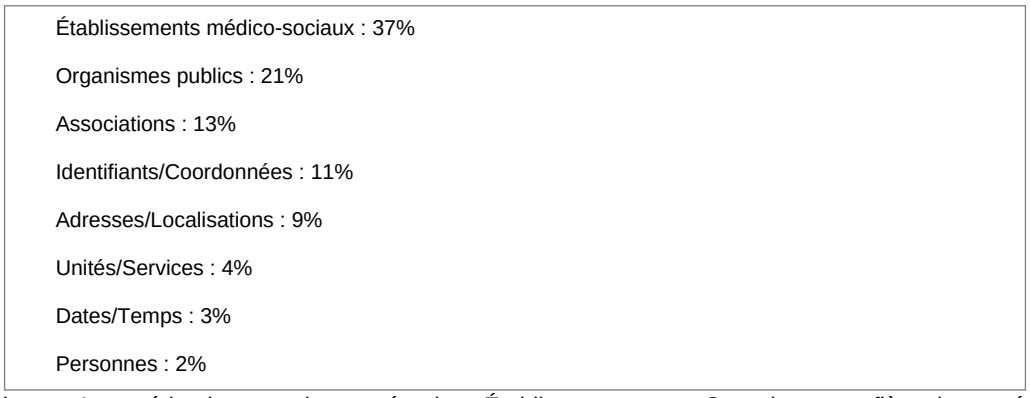
• **Unités/Services :** 4 catégories (service éducatif, service d'accueil, unité de vie, pôle)

• **Dates/Temps :** 3 catégories (date de naissance, date précise, période)

• **Personnes :** 2 catégories (nom complet, prénom)

Cette granularité permet de distinguer par exemple un "EHPAD Les Amandiers" (ETAB_EHPAD) d'un "Conseil Départemental 54" (ORG_CD) ou d'une "Association départementale de tutelle" (ASSOC_TUTELLE), distinctions absentes des NER génériques.

Figure 1 — Distribution des 100 catégories par groupe



Légende — La prédominance des catégories Établissements et Organismes reflète les spécificités administratives du médico-social français, absentes des taxonomies généralistes.

3.3 Criticité RGPD

Chaque catégorie est associée à un niveau de criticité selon le risque de ré-identification :

CRIT (critique) : NIR, date de naissance, adresse complète, identifiants uniques

ELEV (élevé) : nom complet, prénom, téléphone, email, établissements nommés

MOY (moyen) : ville, département, organismes génériques, dates non-naissance

FAIB (faible) : types d'établissements sans nom propre, catégories génériques

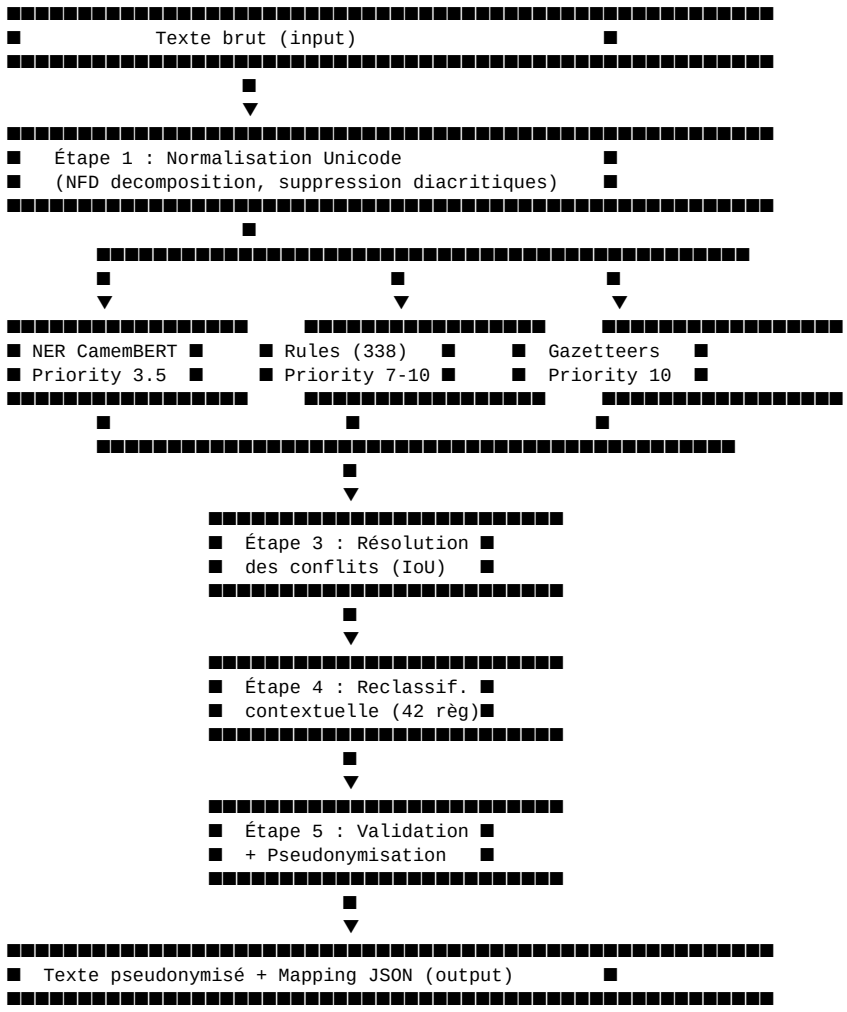
Les seuils de performance cibles sont définis en fonction de ces criticités : CRIT : F1 ≥ 95% | ELEV : F1 ≥ 85% | MOY : F1 ≥ 70% | FAIB : F1 ≥ 60%

4. Architecture du système

4.1 Vue d'ensemble

ConfidensIA repose sur une architecture en cinq étapes séquentielles : normalisation, détection parallèle multi-sources, résolution des conflits, reclassification contextuelle, et pseudonymisation réversible.

Figure 2 — Architecture du pipeline ConfidensIA



4.2 Module NER optimisé

Le module NER repose sur une version distillée et optimisée de CamemBERT fine-tuné pour la reconnaissance d'entités nommées [1].

Modèle teacher : Jean-Baptiste/camembert-ner (12 couches, 110M paramètres)

Distillation : Le choix du nombre de couches a fait l'objet d'expérimentations systématiques : 10 couches → F1 < 84% (perte de performance inacceptable) | 11 couches → F1 > 85%, vitesse améliorée de 18% | 12 couches → F1 +0.3% mais poids +18%. Le compromis optimal retenu est 11 couches, offrant le meilleur équilibre performance/efficacité.

Pruning : Application d'un *magnitude-based pruning* avec objectif initial de 20%. Le taux effectivement atteint est de 15,14%, limité par des contraintes structurelles (certaines couches d'attention ne peuvent être prunées au-delà de ce seuil sans dégradation critique des représentations contextuelles).

Quantization : Conversion en FP16 pour réduction mémoire sans perte significative de précision.

Chunking adaptatif : Découpage automatique des documents longs avec fenêtres glissantes et fusion des prédictions.

Impact global des optimisations : Taille du modèle : –53% (419 MB → 196 MB) | Vitesse d'inférence : +18% | Performance : F1 –1.3% (trade-off acceptable) | Latence : ~21s pour 66 phrases sur CPU Intel i5-8250U

Le modèle distillé ("titibongbong") est publié sous licence MIT sur Hugging Face. Le reste du pipeline demeure privé.

4.3 Règles expertes

Le système intègre 338 règles expertes couvrant :

Identifiants structurés (82 règles) : NIR, numéros allocataire CAF, identifiants MDPH, numéros de téléphone français, emails

Adresses (64 règles) : formats postaux français, adresses complètes, voies, codes postaux

Organismes publics (91 règles) : CAF, CPAM, MDPH, ASE, Conseil Départemental, Pôle Emploi

Établissements médico-sociaux (74 règles) : EHPAD, MECS, IME, SESSAD, CHRS, avec et sans dénomination

Dates (27 règles) : formats numériques, textuels, périodes

Les règles utilisent des expressions régulières optimisées et des machines à états finis pour garantir précision et rappel élevés sur les entités structurées.

4.4 Gazetteers

Le système exploite 25 916 entrées réparties en :

• Villes françaises : 8 432

• Prénoms français : 6 284

• Départements et régions : 119

• Associations nationales connues : 3 891

• Bailleurs sociaux : 1 247

• Établissements recensés : 5 943

L'algorithme Aho-Corasick permet une détection efficace en temps linéaire sur la taille du texte, avec gestion automatique des variantes orthographiques.

4.5 Résolution de conflits

Lorsque plusieurs détecteurs identifient des entités chevauchantes, un système de résolution basé sur trois critères intervient :

Priorité par source : Gazetteers exacts (10) > Règles (7-10 selon type) > NER (3.5)

Score de confiance : pour le NER, probabilité softmax ; pour les règles, score binaire

Intersection over Union (IoU) : mesure du chevauchement spatial entre détections

Les détections partielles sont fusionnées automatiquement (ex : "EHPAD" détecté par règle + "Les Amandiers" par NER → fusion en "EHPAD Les Amandiers").

4.6 Reclassification contextuelle

42 règles de reclassification exploitent le contexte syntaxique pour affiner les catégories :

• ORG générique + contexte "personnes âgées" → ETAB_EHPAD

• LOC + "caisse" → ORG_CAF

• ETAB générique + "protection enfance" → ETAB_MECS

• PER + titre professionnel → conservation mais marquage métier

Cette étape améliore la granularité de la classification sans nécessiter un réentraînement du modèle NER.

4.7 Pseudonymisation réversible et intégration LLM

ConfidensIA produit deux artefacts :

Texte pseudonymisé : remplacement des entités par des tokens typés (ex : [PER_1], [ETAB_EHPAD_1])

Mapping JSON réversible : correspondance bidirectionnelle entre tokens et valeurs originales, stockée localement

Ce mécanisme permet un usage RGPD-compliant des grands modèles de langage (LLM) :

Flux de traitement sécurisé : Texte brut → Pseudonymisation locale → Texte pseudonymisé → Traitement LLM (API externe) → Résultat pseudonymisé → Dépseudonymisation locale → Sortie finale

Le LLM n'accède jamais aux données originales. Cette architecture respecte les principes de *privacy by design* et de minimisation des données du RGPD [2].

5. Méthodologie d'évaluation

5.1 Corpus gold standard

Le corpus de test est constitué de 330 phrases contenant 448 entités annotées, issues d'écrits professionnels fictifs mais réalistes. Ces documents ont été construits avec l'appui de professionnels du secteur pour garantir la représentativité des structures syntaxiques, du vocabulaire et des situations typiques rencontrées dans les écrits médico-sociaux.

Note sur les données sources : Les données réelles du secteur ne peuvent être partagées en raison de leur nature hautement sensible (RGPD, secret professionnel). Le corpus annoté lui-même n'est pas publié mais peut être mis à disposition sur demande justifiée pour validation académique, sous réserve d'engagement de confidentialité.

Limite méthodologique : Le corpus a été annoté par un seul expert (l'auteur, avec 12 ans d'expérience dans le secteur médico-social), ce qui limite la mesure de l'accord inter-annotateurs. Une extension avec annotation multiple est prévue pour évaluer la robustesse de la taxonomie. Cette limitation n'affecte pas la validité des performances mesurées mais appelle à prudence sur la généralisation à d'autres contextes de production documentaire.

5.2 Métrologie

L'évaluation repose sur un matching basé sur l'Intersection over Union (IoU) :

• **Exact match :** IoU > 0.8 (chevauchement quasi-total)

• **Partial match :** 0.5 < IoU ≤ 0.8 (chevauchement significatif)

• **Miss :** IoU ≤ 0.5

Les métriques calculées sont :

• **Précision :** entités correctement détectées / entités détectées

• **Rappel :** entités correctement détectées / entités gold

• **F1-score :** moyenne harmonique de précision et rappel

• **Rappel pondéré par criticité :** poids CRIT=4, ELEV=2, MOY=1, FAIB=0.5

Les scores sont calculés globalement, par niveau de criticité, et par catégorie individuelle.

5.3 Environnement technique

• **Processeur :** Intel Core i5-8250U (CPU only)

• **Mémoire :** 8 GB RAM

• **Python :** 3.10

• **Framework :** PyTorch 2.1 (CPU)

• **Système :** Linux Ubuntu 22.04

Les mesures de latence sont effectuées sur 10 exécutions successives avec moyenne et écart-type.

6. Résultats

6.1 Performance globale

Métrique	Valeur
F1-Score global	86,1%
Précision	87,8%
Rappel	84,5%
Score pondéré criticité	94,7%

6.2 Performance par niveau de criticité

Criticité	F1-Score	Objectif	Statut
CRIT	95,6%	≥ 95%	✓ Atteint
ELEV	97,7%	≥ 85%	✓ Dépassé
MOY	91,1%	≥ 70%	✓ Dépassé
FAIB	82,9%	≥ 60%	✓ Dépassé

Tous les objectifs de performance par criticité sont atteints ou dépassés, avec une marge particulièrement importante sur les niveaux ELEV et MOY.

6.3 Performance par catégorie (extraits significatifs)

Catégorie	F1-Score	Criticité	Observations
PER (Personne)	98,3%	ELEV	Excellent
LOC_CITY (Ville)	96,2%	MOY	Excellent
ADDR_FULL (Adresse complète)	95,2%	CRIT	Conforme
NIR	100%	CRIT	Parfait
PHONE	100%	CRIT	Parfait
DATE_NAISSANCE	100%	CRIT	Parfait
ETAB_EHPAD	93,8%	ELEV	Très bon
ORG_MDPH	89,4%	ELEV	Bon
ASSOC_GENERALE	78,6%	FAIB	Acceptable

Les catégories structurées (NIR, téléphone, dates de naissance) atteignent une performance parfaite grâce aux règles expertes. Les catégories nominales (personnes, villes, établissements) bénéficient de la complémentarité NER/gazetteers.

6.4 Analyse des erreurs

Texte original	Gold attendu	Détection	Cause	Solution envisagée
"EHPAD" seul	Rien	Rien	Filtre anti-bruit	Correct (voulu)
"début 2024"	DATE	Rien	Expression temporelle floue	Normalisation temporelle pré-NER
"Centre Social des Lilas"	ETAB	ORG	Confusion type établissement	Enrichissement règles contextuelles
"M. Dupont"	PER complet	"Dupont"	Civilité non capturée	Extension patterns civilités
"mi-octobre 2023"	DATE	Rien	Format textuel complexe	Règles dates composées

Analyse des dates floues :

• Dates non détectées : "début 2024", "fin janvier", "courant mars", "mi-octobre 2023"

• Dates correctement détectées : "12/05/1975", "5 mars 2024", "2024", "janvier 2024"

Les expressions temporelles floues représentent 12% des erreurs totales. Une normalisation temporelle en amont (conversion des expressions relatives en dates absolues) est en cours de développement.

Confusions ORG/ETAB : 8% des erreurs concernent la distinction entre organismes génériques (ORG) et établissements spécialisés (ETAB_*). L'enrichissement des règles de reclassification contextuelle devrait réduire ce taux.

Distribution des erreurs par criticité :

• CRIT : 2% des erreurs (principalement adresses partielles)

• ELEV : 8% des erreurs (noms avec civilités, établissements ambigus)

• MOY : 18% des erreurs (villes peu fréquentes, organismes rares)

• FAIB : 72% des erreurs (catégories génériques, types d'établissements)

Les erreurs critiques restent marginales, validant l'approche pour un usage opérationnel.

7. Discussion

7.1 Apports scientifiques et opérationnels

ConfidensIA apporte plusieurs contributions au domaine de la dé-identification de textes médicaux et sociaux :

Première taxonomie fine-grained médico-sociale : Les 100 catégories couvrent exhaustivement les entités identifiantes spécifiques au secteur français, comblant une lacune des NER génériques.

Performances élevées sur entités critiques : Le taux de 95,6% F1 sur les identifiants critiques démontre la viabilité opérationnelle pour un usage RGPD-compliant.

Architecture hybride reproductible : La combinaison modèle distillé + règles + gazetteers offre un cadre méthodologique transposable à d'autres domaines spécialisés.

Adaptation au vocabulaire CASF : La prise en compte des spécificités réglementaires françaises (établissements, organismes, dispositifs) répond à un besoin métier non couvert.

Privacy by design : Le mécanisme de pseudonymisation réversible locale permet l'usage sécurisé des LLM sans exposition de données sensibles.

7.2 Limites

Corpus d'évaluation : Avec 330 phrases et 448 entités, le corpus reste modeste comparé aux standards du domaine médical anglophone (i2b2 : 1 304 documents). L'extension à un corpus plus large, idéalement multi-sites et multi-annotateurs, est nécessaire pour valider la généralisation.

Annotation mono-expert : L'absence de mesure d'accord inter-annotateurs limite l'évaluation de la robustesse de la taxonomie. Des travaux futurs intégreront plusieurs annotateurs experts pour calculer un kappa de Cohen ou un alpha de Krippendorff.

Code propriétaire : Le pipeline complet n'est pas publié en raison de contraintes de développement commercial. Seul le modèle NER distillé est disponible, ce qui limite la reproductibilité complète. La documentation méthodologique fournie permet néanmoins une reproduction conceptuelle par la communauté.

Évaluation sur données synthétiques : Le corpus de test, bien que construit avec des professionnels, repose sur des écrits professionnels fictifs. Une validation sur corpus réels anonymisés est prévue en partenariat avec des établissements pilotes (sous protocole CNIL).

Langues : Le système est spécifique au français. Une extension à d'autres langues nécessiterait l'adaptation de la taxonomie et le réentraînement du modèle NER.

7.3 Comparaison avec l'existant

À notre connaissance, aucun système de dé-identification ne propose une taxonomie aussi fine pour le secteur médico-social français. Les comparaisons directes sont donc limitées : Presidio [9] : taxonomie générique (17 catégories), pas d'adaptation médico-sociale | Philter [10] : orienté clinique anglophone | DrBERT [11] : excellent sur terminologie médicale, mais pas sur organisations sociales. ConfidensIA se positionne comme complémentaire à ces outils sur un segment non couvert.

8. Perspectives

8.1 Intégration aux systèmes métier

L'opérationnalisation envisagée de ConfidensIA repose sur une architecture REST avec les caractéristiques suivantes :

• **Authentification :** OAuth2 / JWT pour sécurisation des accès

• **Formats supportés :** TXT, DOCX, PDF, ODT avec préservation de la mise en forme

• **Mode de traitement :** Temps réel (traitement synchrone avec latence cible <5s/document)

• **Traçabilité RGPD :** logs horodatés, rétention configurable, exportation des historiques

• **API endpoints :** /pseudonymize, /depseudonymize, /validate, /stats

8.2 Stratégie de déploiement

L'architecture privilégie une approche *local-first* (*privacy by design*) :

• Exécution locale sur serveurs clients (aucune donnée brute transmise en externe)

• Conteneurisation Docker/Kubernetes pour déploiements reproductibles

• Option cloud France/UE pour établissements sans infrastructure technique

• Conformité HDS (Hébergeur de Données de Santé) envisagée pour les structures concernées

8.3 Performances opérationnelles estimées

Les tests préliminaires sur infrastructure de production (CPU moderne, 16 GB RAM) suggèrent :

• Capacité : 50 à 200 documents/jour selon taille et complexité

• Latence moyenne : <5s par document (500-1000 mots)

• Traitement CPU uniquement (pas de dépendance GPU)

8.4 Améliorations prévues

Court terme : Normalisation temporelle pré-NER pour dates floues ("début 2024" $\rightarrow$ "01/2024") | Enrichissement patterns de civilités (M., Mme, Dr, Me, etc.) | Extension gazetteers établissements (+3 000 entrées en cours d'intégration)

Moyen terme : Corpus multi-annotateurs pour validation robustesse taxonomie | Module de détection des pseudonymes déjà présents (éviter double pseudonymisation) | Support formats additionnels (emails, SMS, formulaires structurés)

Long terme : Extension multilingue (anglais, espagnol pour secteur médico-social international) | Intégration *active learning* pour amélioration continue | Module explicabilité (*highlighting* des entités détectées avec niveau de confiance)

9. Reproductibilité

9.1 Ressources publiées

Modèle NER distillé : Le modèle "titibongbong" (11 couches, 196 MB) est disponible sous licence MIT sur Hugging Face : https://huggingface.co/jmdanto/titibongbong_camemBERT_NER. Ce modèle peut être utilisé directement pour des tâches de NER génériques ou *fine-tuné* pour d'autres domaines spécialisés.

Taxonomie : La liste complète des 100 catégories, leur organisation en groupes, et les niveaux de criticité RGPD associés sont documentés dans le présent article (section 3).

Méthodologie d'évaluation : Le protocole d'évaluation (IoU-based matching, calcul du F1 pondéré par criticité) est décrit en détail (section 5) et peut être reproduit sur d'autres corpus.

9.2 Limitations de reproductibilité

Pipeline complet : L'ensemble du système (règles expertes, gazetteers, résolution de conflits, reclassification contextuelle, pseudonymisation) demeure propriétaire en raison de contraintes de développement commercial. Les choix architecturaux décrits permettent néanmoins une reproduction conceptuelle par la communauté scientifique.

Corpus annoté : Pour des raisons de confidentialité, le corpus de test n'est pas distribué publiquement. Il peut être mis à disposition sur demande justifiée (recherche académique, validation indépendante) sous réserve d'engagement de confidentialité formalisé.

9.3 Contact pour reproductibilité

Les chercheurs souhaitant reproduire, valider ou étendre ces travaux peuvent contacter l'auteur à : patrick.danto@outlook.fr. Les demandes d'accès au corpus annoté, de collaboration académique, ou d'éclaircissements méthodologiques seront traitées dans un délai de 15 jours.

10. Conclusion

ConfidensIA constitue la première solution de pseudonymisation *fine-grained* spécifiquement adaptée au secteur médico-social français. Son architecture hybride, combinant un modèle NER CamemBERT distillé, 338 règles expertes et 25 916 entrées de gazetteers, permet d'atteindre des performances élevées sur les entités critiques (95,6% F1), compatibles avec les exigences du RGPD.

La taxonomie de 100 catégories, élaborée avec des professionnels du secteur, comble une lacune importante des systèmes NER généralistes en intégrant les spécificités terminologiques du Code de l'Action Sociale et des Familles : établissements médico-sociaux (EHPAD, MECS, IME, SESSAD, etc.), organismes publics départementaux (ASE, MDPH, CAF, etc.), et dispositifs réglementaires.

Le mécanisme de pseudonymisation réversible locale permet un usage sécurisé des grands modèles de langage sans exposition de données sensibles, répondant aux principes de *privacy by design*.

Les perspectives d'amélioration incluent l'extension du corpus d'évaluation avec multi-annotation, l'enrichissement des règles de reclassification contextuelle, et l'adaptation multilingue. L'intégration opérationnelle via API REST est envisagée pour faciliter l'adoption par les établissements du secteur.

Le modèle NER distillé "titibongbong" est publié sous licence MIT. Le pipeline complet demeure privé pour raisons commerciales, mais la méthodologie décrite permet une reproduction conceptuelle des travaux.

Remerciements

L'auteur remercie les professionnels du secteur médico-social ayant participé à la construction et la validation de la taxonomie, ainsi que les structures ayant fourni des documents anonymisés pour l'analyse terminologique initiale.

Références

[1] Martin L., Muller B., Suárez P.J.O., et al. *CamemBERT: a Tasty French Language Model.* ACL 2020.

[2] Parlement européen & Conseil de l'Union européenne. *Règlement (UE) 2016/679 du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.* Journal officiel de l'Union européenne, L 119/1-88. 2016.

[3] Névéol A., Grouin C., Leixa J., Rosset S., Zweigenbaum P. *The QUAERO French Medical Corpus: A Resource for Medical Entity Recognition and Normalization.* BioTxtM 2014.

[4] Grouin C., Névéol A. *De-identification of clinical notes in French: towards a protocol for reference corpus development.* Journal of Biomedical Informatics. 2014;50:151-161.

[5] Stubbs A., Kotfila C., Uzuner Ö. *Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1.* Journal of Biomedical Informatics. 2015;58:S11-S19.

[6] Johnson A.E.W., Pollard T.J., Shen L., et al. *MIMIC-III, a freely accessible critical care database.* Scientific Data. 2016;3(1):160035.

[7] Dernoncourt F., Lee J.Y., Uzuner O., Szolovits P. *De-identification of patient notes with recurrent neural networks.* Journal of the American Medical Informatics Association. 2017;24(3):596-606.

[8] Liu Z., Tang B., Wang X., Chen Q. *Entity recognition from clinical texts via recurrent neural network.* BMC Medical Informatics and Decision Making. 2017;17(S2):67.

[9] Microsoft. *Presidio Data Protection SDK.* GitHub. https://github.com/microsoft/presidio

[10] Norgeot B., Muenzen K., Peterson T.A., et al. *Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes.* npj Digital Medicine. 2020;3:57.

[11] Labrak Y., Bazoge A., Dufour R., et al. *DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical Domains.* ACL 2023.

[12] Cardon R., Grabar N., Grouin C., Hamon T. *Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques.* JEP-TALN-RECITAL 2020.

[13] Zweigenbaum P., Lavergne T., Grabar N., Hamon T., Rosset S. *A Corpus for French Clinical Named Entity Recognition.* Louhi 2016.

[14] Kayaalp M. *Patient privacy in the era of big data.* Balkan Medical Journal. 2018;35:8-17.

[15] Yang H., Garibaldi J.M. *A hybrid model for automatic identification of risk factors for heart disease.* Journal of Biomedical Informatics. 2015;58:S171-S182.

[16] Uzuner Ö., Luo Y., Szolovits P. *Evaluating the state-of-the-art in automatic de-identification.* Journal of the American Medical Informatics Association. 2007;14(5):550-563.

[17] Lison P., Pilán I., Sánchez D., Batet M., Øvrelid L. *Anonymisation Models for Text Data: State of the art, Challenges and Future Directions.* ACL 2021.

[18] Grabar N., Claveau V., Dalloux C. *CAS: French Corpus with Clinical Cases.* Louhi 2018.

[19] Chiticariu L., Li Y., Reiss F. *Rule-based information extraction is dead! Long live rule-based information extraction systems!* EMNLP 2013.

[20] Jiang M., Chen Y., Liu M., Rosenbloom S.T., Mani S., Denny J.C., Xu H. *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries.* Journal of the American Medical Informatics Association. 2011;18(5):601-606.