Dé-identification automatisée de courriers médicaux : la méthode FASDIM

Travail publié par :
Emmanuel CHAZARD
Capucine MOURET-KUBIAK
Grégoire FICHEUR
Régis BEUSCART

Introduction

- Utilisation de données personnelles relatives à la santé : respect de la confidentialité du patient (article L1110-4 du Code de la Santé Publique)
- Bien différencier 3 types de données :
  - Nominatives
  - Anonymes
  - Indirectement nominatives
    - Données « banales » [illustration au tableau]
    - Problème des identifiants arbitraires réels [illustration au tableau]
- Opérations sur un document (termes ambigus liés à l'historique):
  - Anonymisation d'un document :
    Retrait du nom et prénoms, insuffisant
  - Dé-identification d'un document :
    Retrait des données directement et indirectement nominatives

Introduction

- Automatisation de la dé-identification :
  - Nécessaire pour traiter un grand nombre de courriers
  - Aucun outil performant et libre en langue française
  - Tous les outils existants nécessitent un temps très élevé avant de dé-identifier le premier courrier
- DEUX familles d'approches classiques :
  - Pattern matching :
    - SOIT par suppression de mots inclus dans des dictionnaires de noms propres
    - SOIT par conservation des seuls mots inclus dans des dictionnaires de noms communs
    - => Nécessite de disposer de ces dictionnaires, +/- de règles grammaticales (conjugaisons, accords)
  - Machine learning
    - Méthode entièrement automatisée
    - Nécessite une base d'apprentissage : des centaines de courriers natifs et dé-identifiés à la main
    - La machine apprend sur cette base comment identifier les termes à supprimer

Objectifs

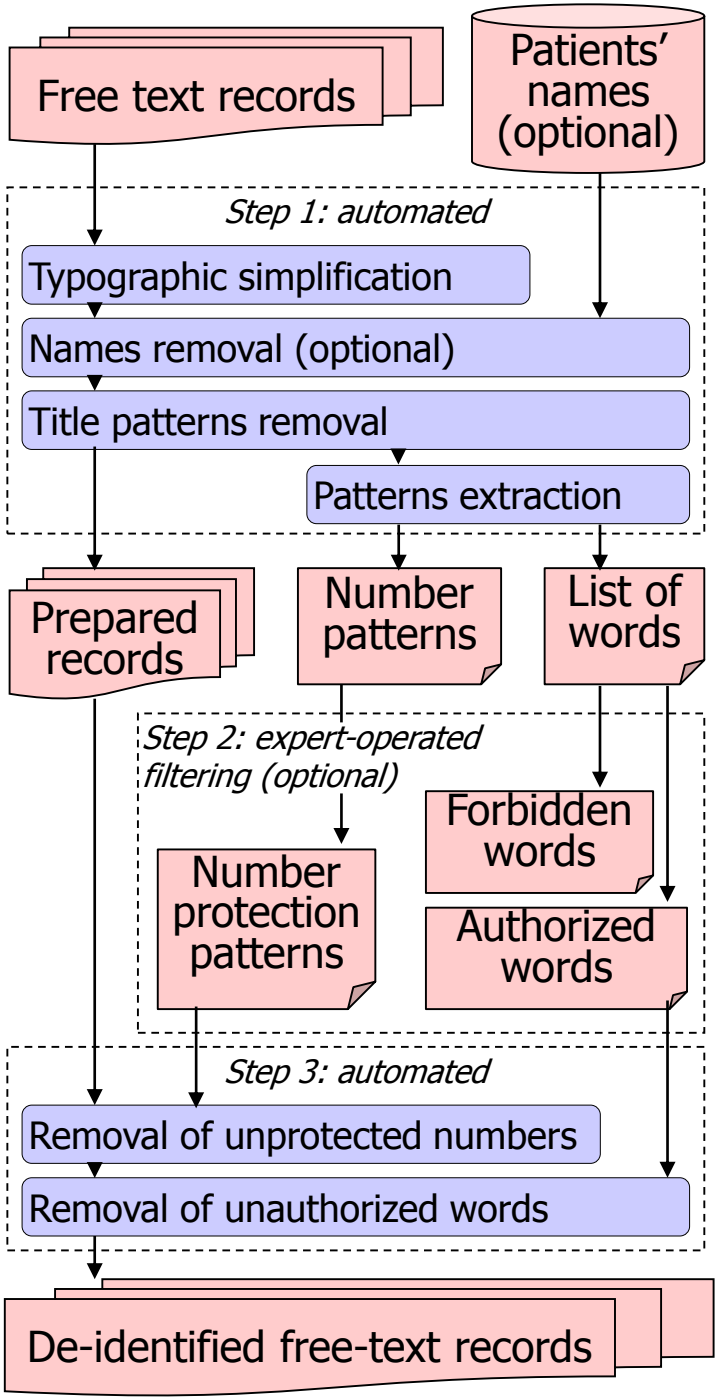
- Élaborer FASDIM, *Fast and Simple De-Identification Method* :
  - Pour courriers de sortie et comptes-rendus non structurés
  - Sans dictionnaire de mots préconçu ni connaissance grammaticale (≠pattern matching)
  - Sans corpus d'entraînement (≠machine learning)
  - => plus rapide en particulier pour des corpus de taille moyenne (1000 à 100 000)
- Evaluer FASDIM :
  - Performance de la méthode
  - Conservation de l'information médicale
  - Temps de travail requis à l'élaboration et la mise à jour

Matériel

- Jeu de 27 540 documents médicaux en texte libre
- Noms et prénoms des patients (SIH, insuffisant)
- Critères définissant une donnée identifiante selon l'HIPAA (*Health Insurrance Portability and Accountability Act*) : les « PHI » (*Protected Health Information*)

Noms
Indications géographiques
Adresses email
Numéros de Sécurité Sociale
Numéros de licence ou de certificat
Numéros d'immatriculation d'un véhicule
Identifiants biométriques
Photographies de face
Tous les éléments de date
Numéros de téléphone
Numéros de fax

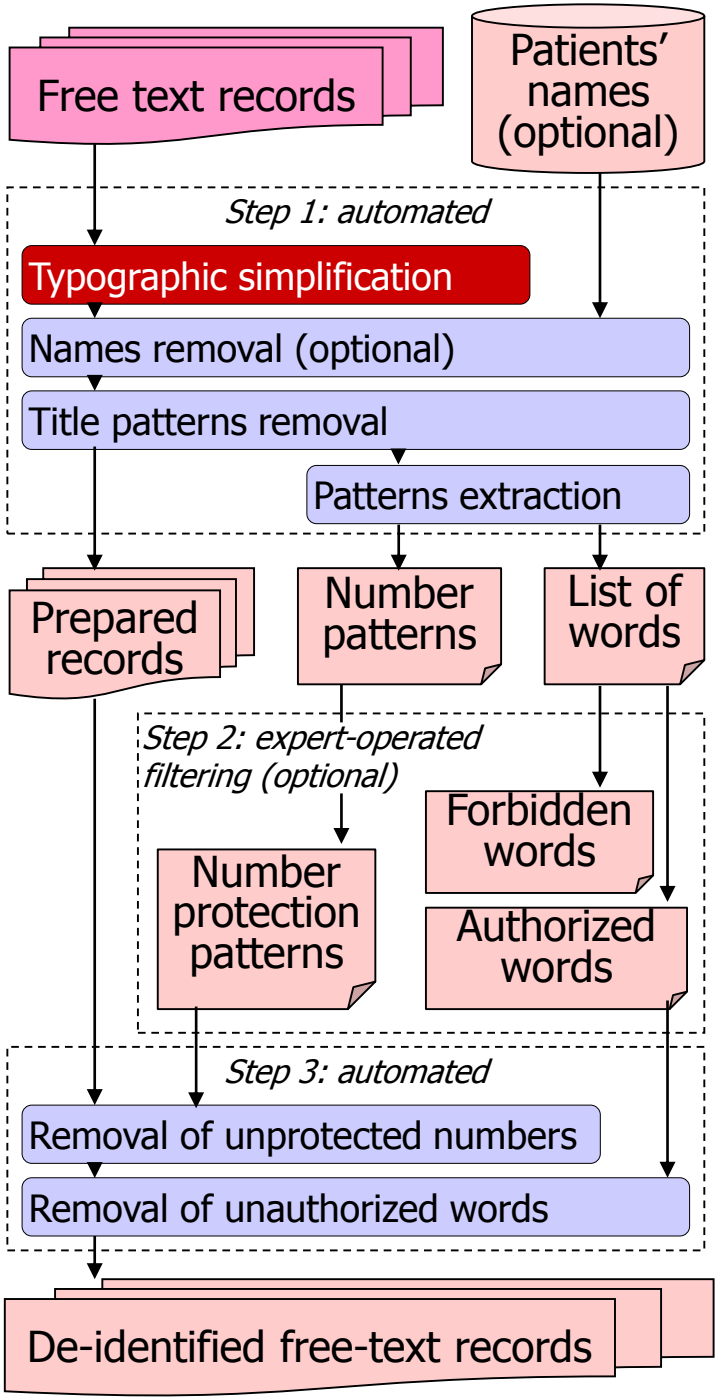
Numéros d'assurance santé
Numéros de compte
URL
Adresses IP
Numéros de dossiers
Numéros d'identification ou de série
Tout autre numéro ou identifiant ou code
+ noms de soignants ou de structure de soins
+ adresses de structures de soins



Méthode :

Élaboration de FASDIM

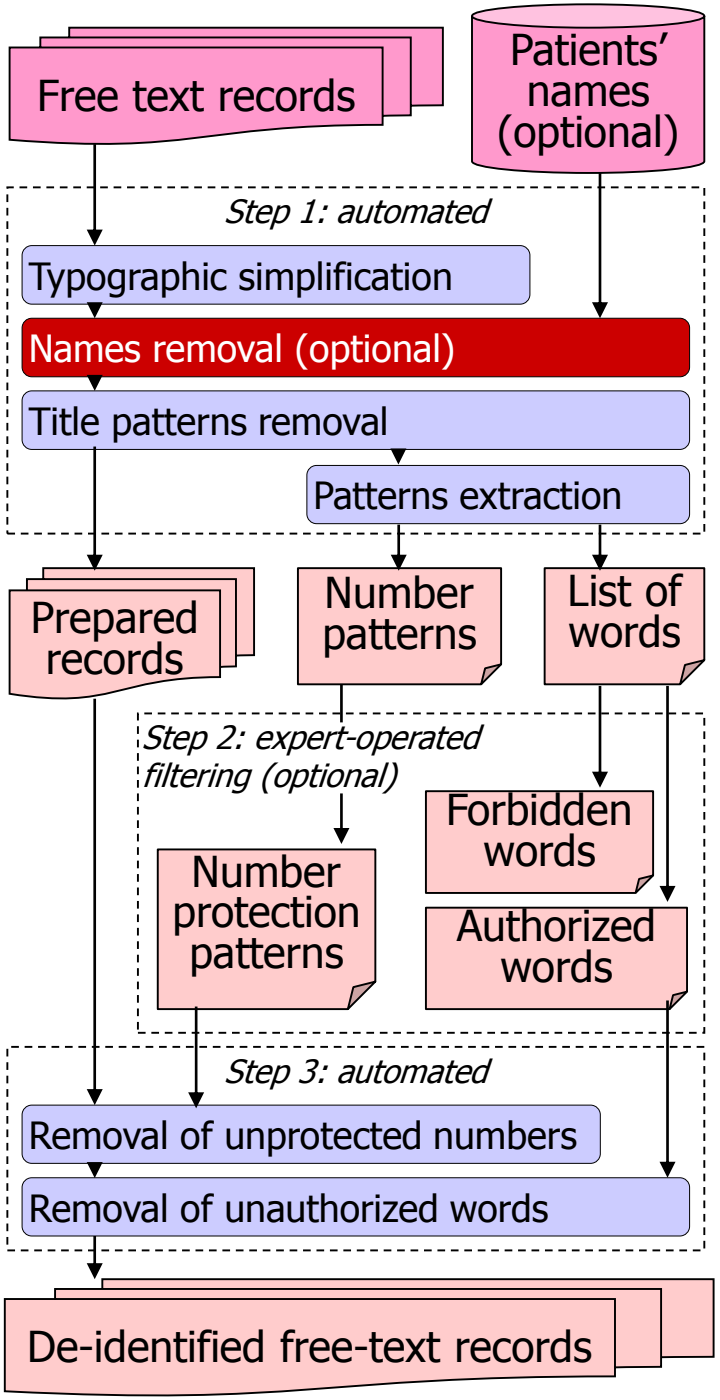
1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

Élaboration de FASDIM

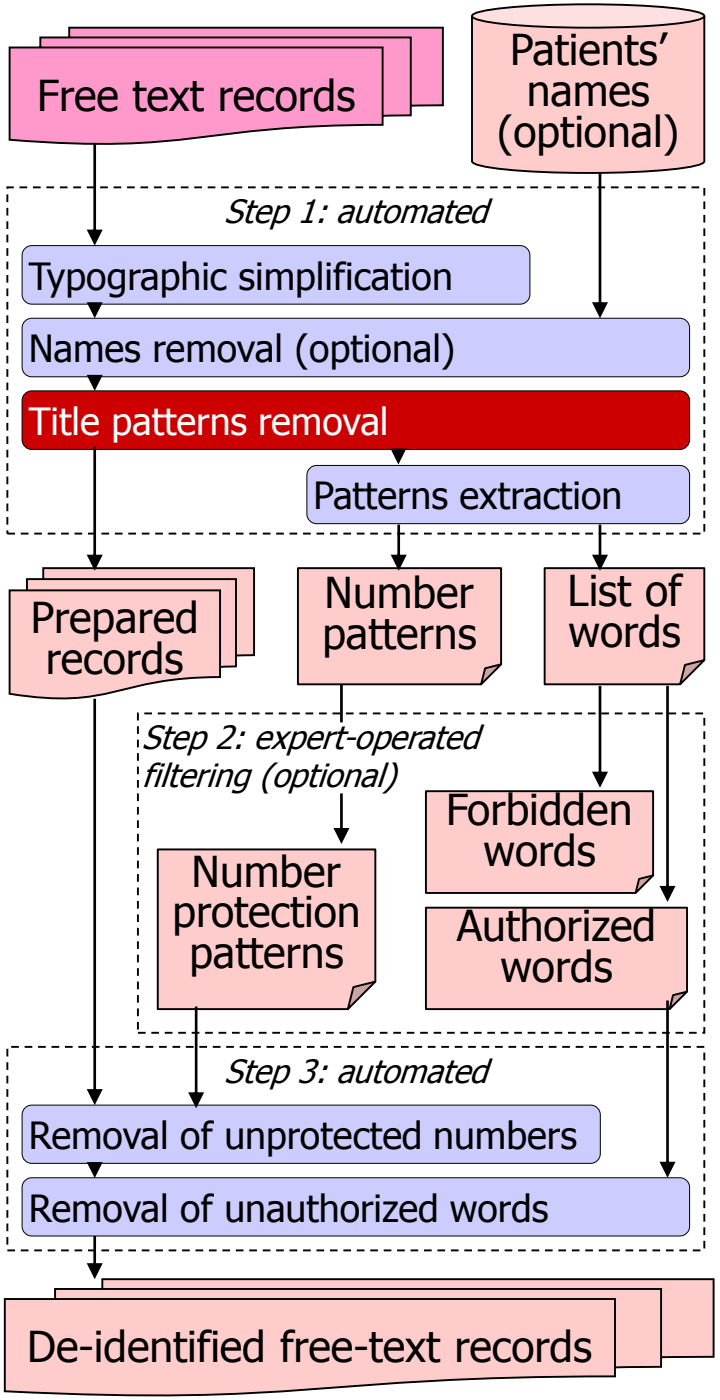
1. Simplification typographique
   - Suppression des accents
     - *« é » remplacé par « e »*
   - Suppression des caractères spéciaux
     - *« œ » remplacé par « oe »*
   - Minuscules
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

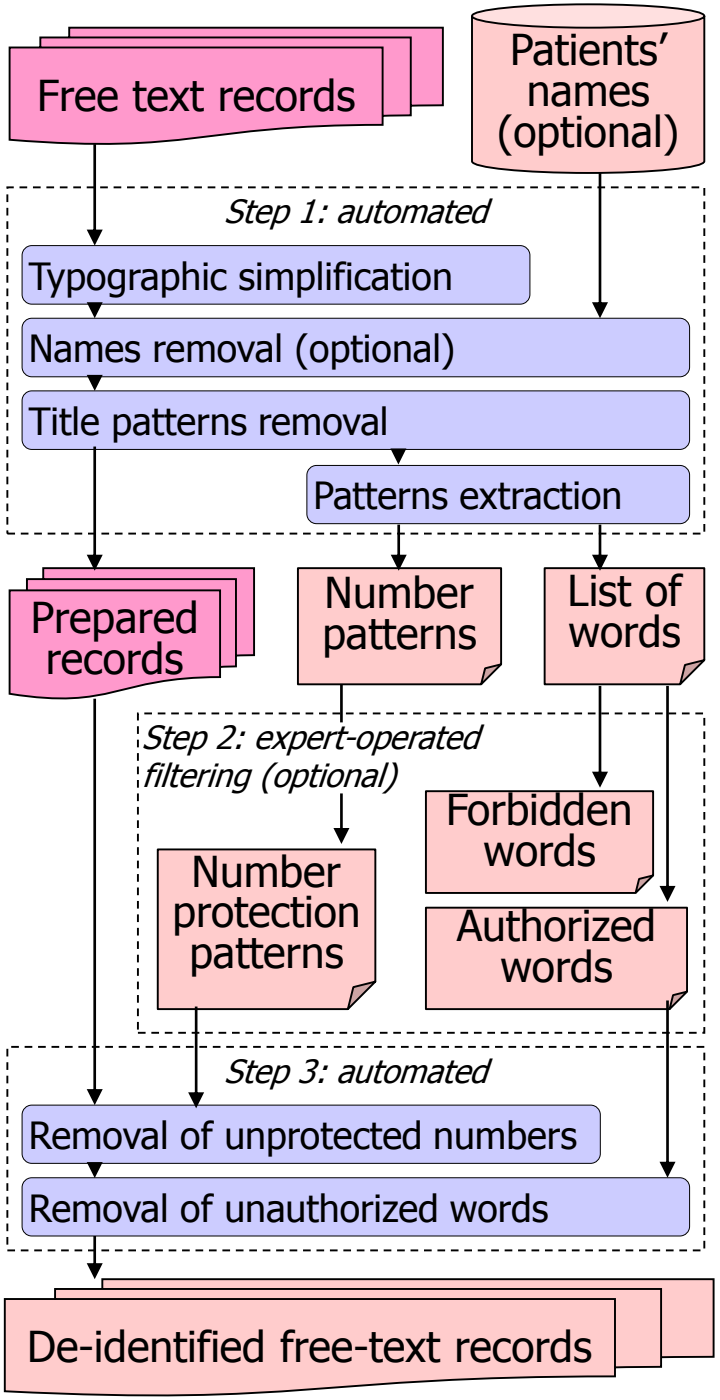
Élaboration de FASDIM

1. Simplification typographique
2. Suppression des noms et prénoms à partir de la base de données du Système d'Information Hospitalier
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode : Élaboration de FASDIM

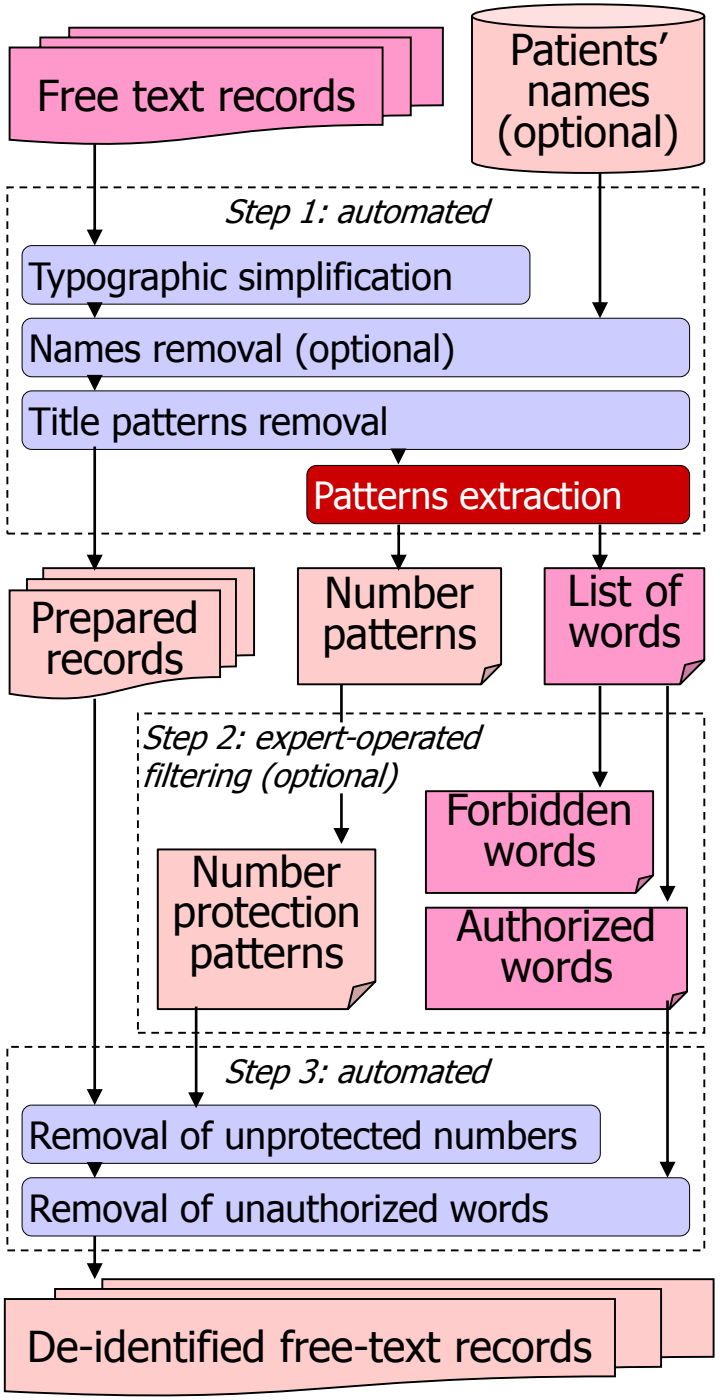
1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
   - « monsieur jean dupont » ➔ « @@@ »
   - « dr marie martin » ➔ « @@@ »
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

Élaboration de FASDIM

1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés

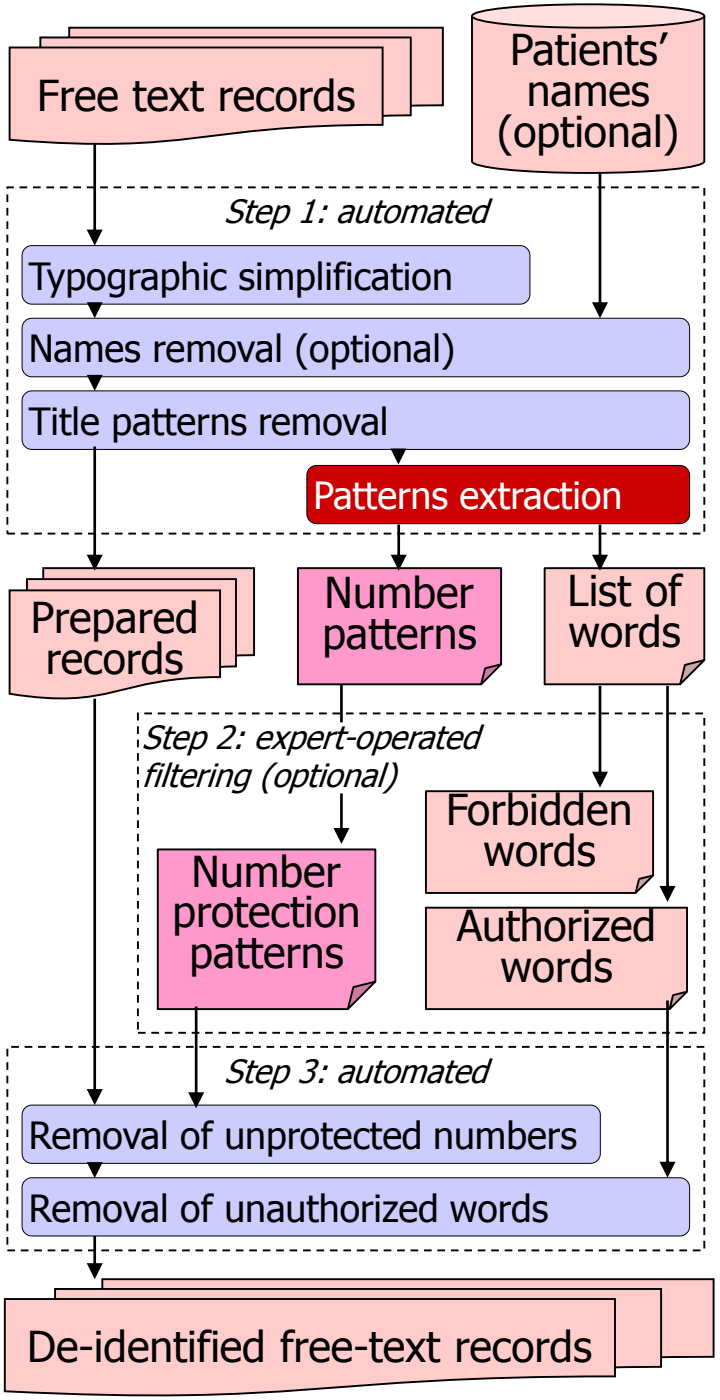


Méthode :

Élaboration de FASDIM

1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
   - 2 listes créées à partir de tous les mots présents dans les courriers
   - Mots interdits :

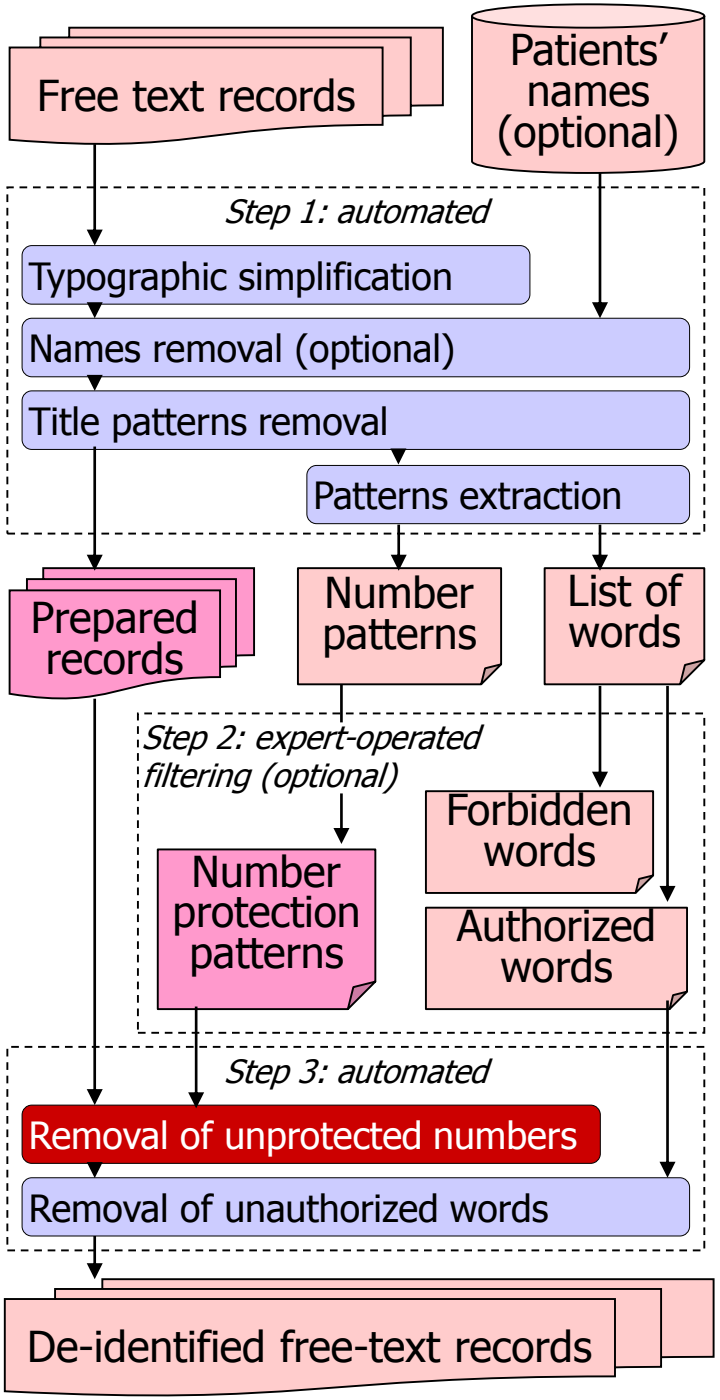
     ex « acacias », « duppont »
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

Élaboration de FASDIM

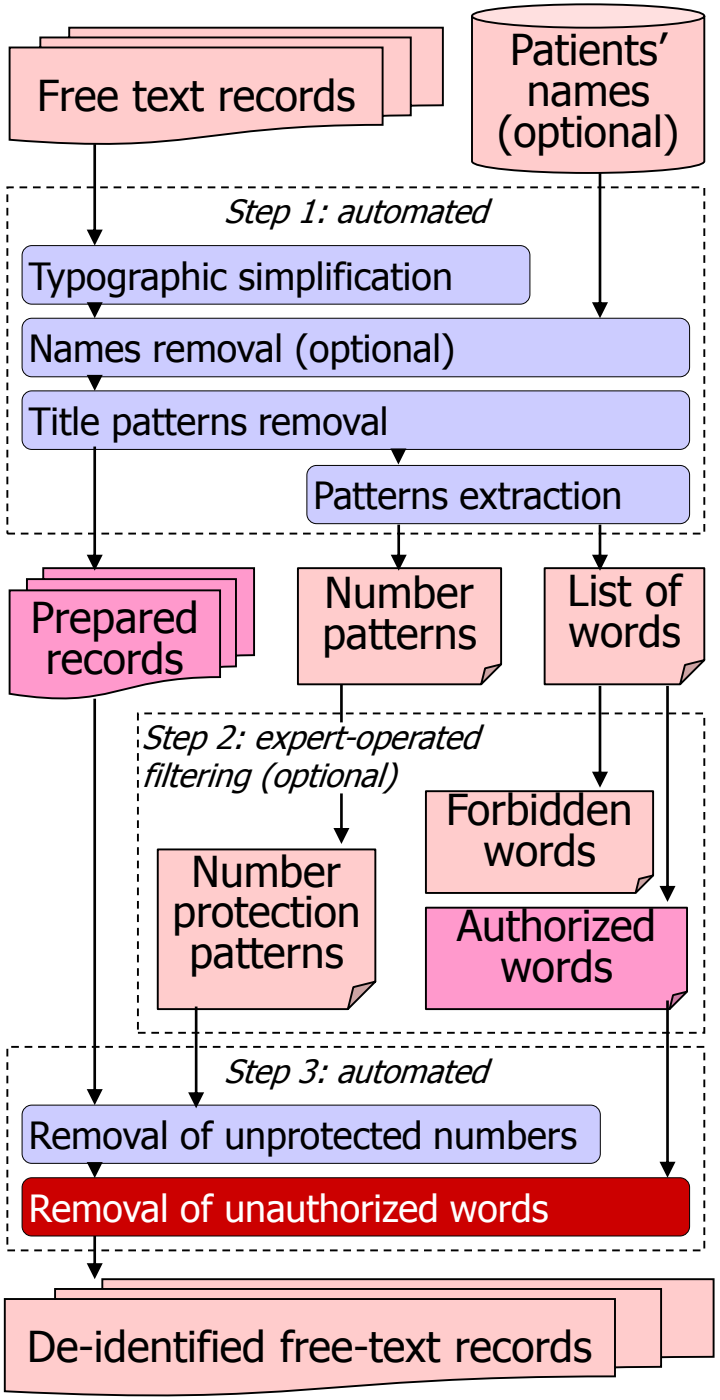
1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
   - Motifs de chiffres non identifiants
     - Ex *« [chiffre] gelules » « [chiffre] g/l »*
   - Protection des chiffres contenus dans ces motifs
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

Élaboration de FASDIM

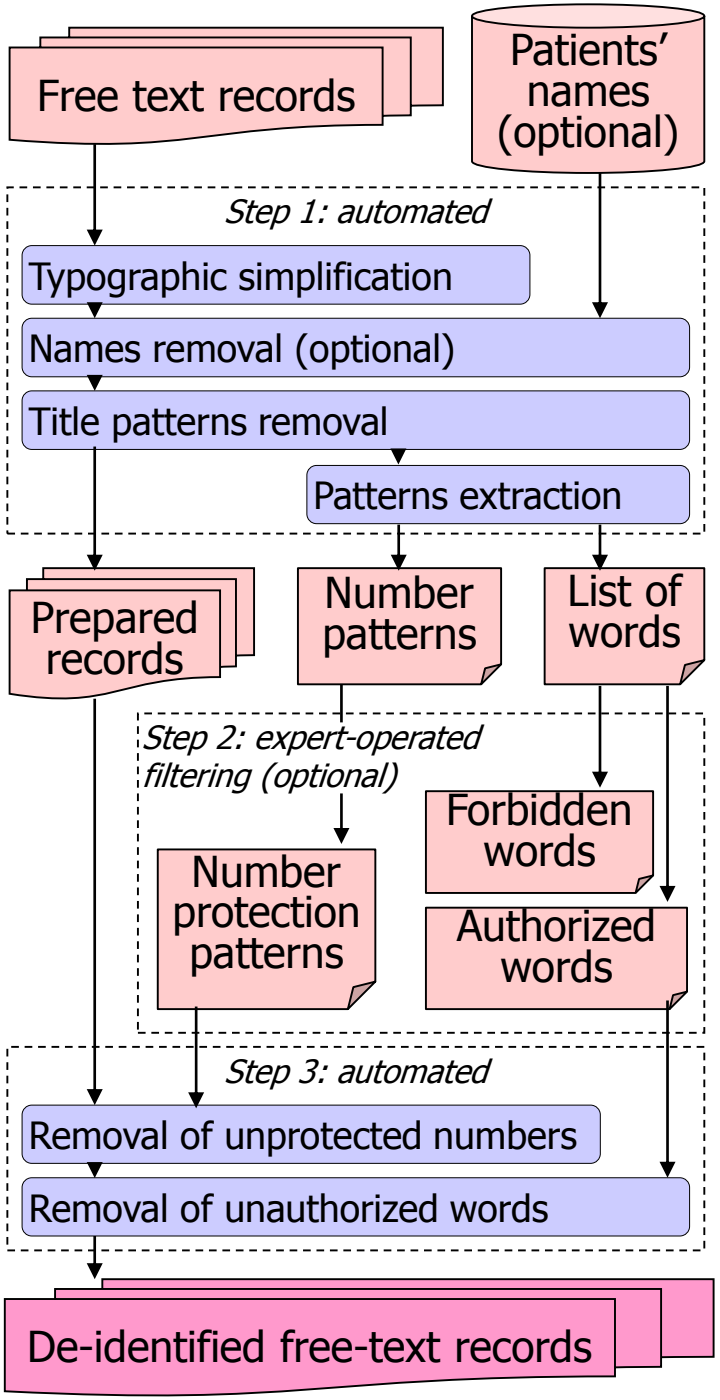
1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :

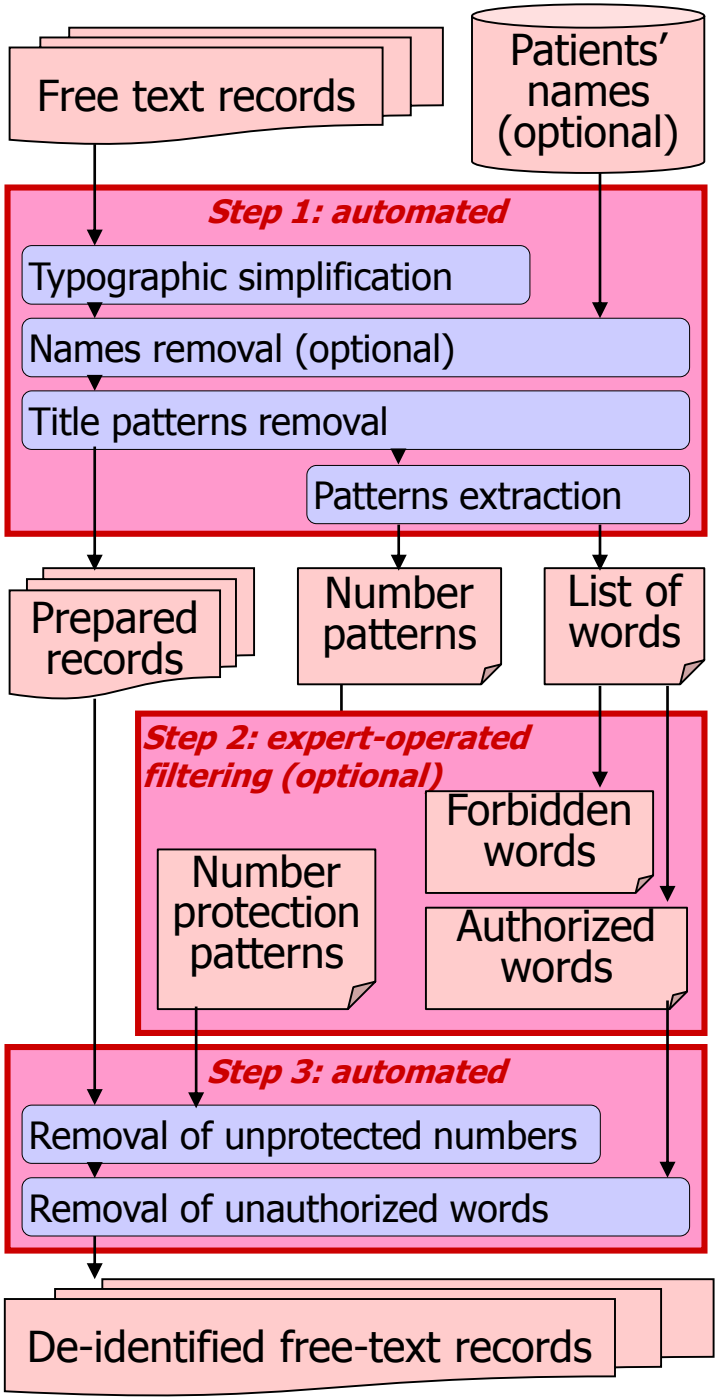
Élaboration de FASDIM

1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode : Élaboration de FASDIM

1. Simplification typographique
2. Suppression des noms et prénoms
3. Suppression des motifs incluant un titre ou une civilité
4. Création de la liste de mots autorisés
5. Création de la liste de motifs de protection des chiffres
6. Suppression des chiffres non protégés
7. Suppression des mots non autorisés



Méthode :
Élaboration de FASDIM

3 phases :

Phase 1 : automatisée, prépare les courriers et les listes à filtrer

Phase 2 : de temps à autre, filtrer et mettre à jour les listes

Phase 3 : automatisée, réalise la dé-identification

Exemple fictif de courrier dé-identifié par FASDIM, focus sur les nombres

@ @
lb.dng.@.@.@
cher confrere,

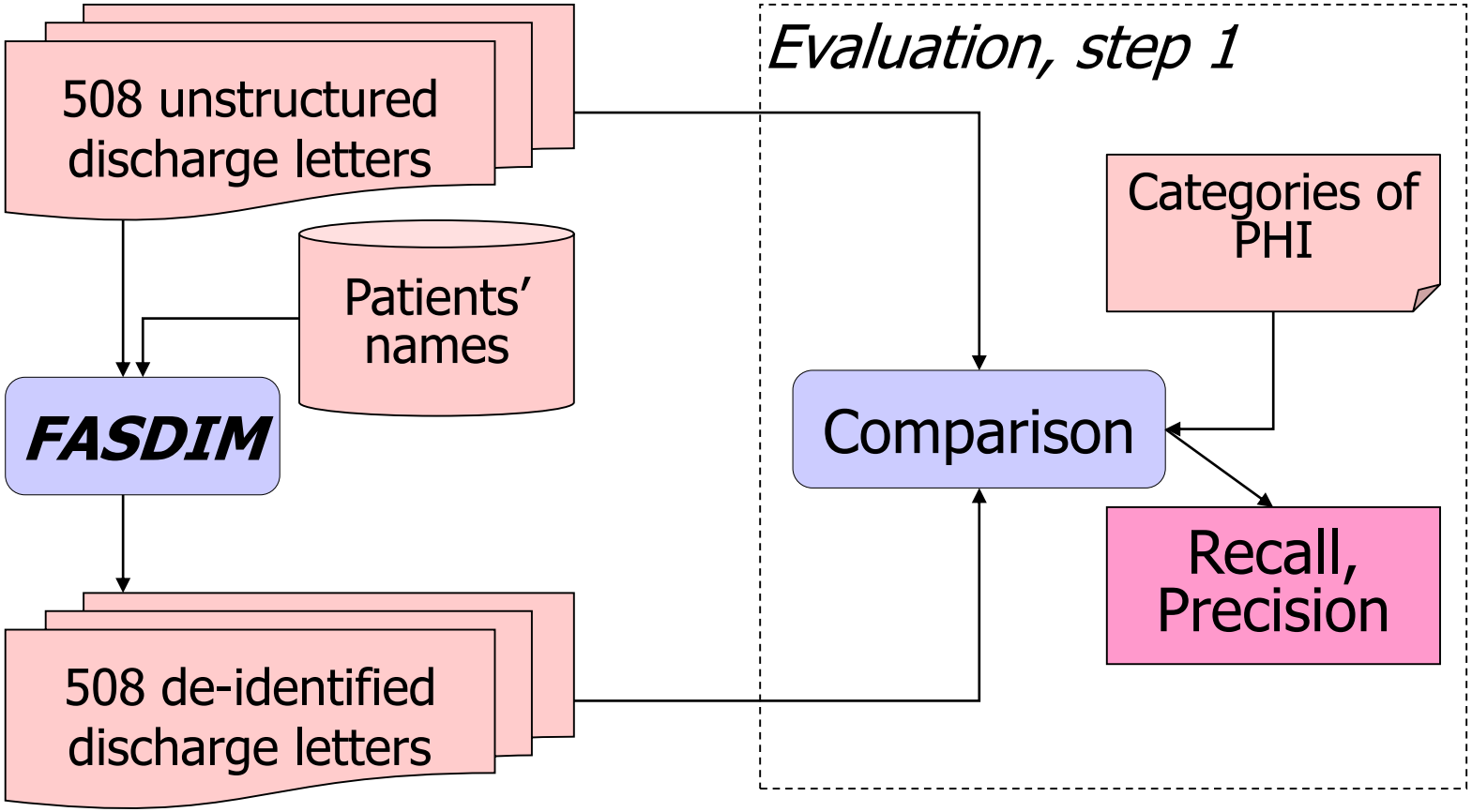
votre patient @ @ @ , age de @ ans a ete admis en chirurgie du @ au @.@.@ pour douleurs abdominales epigastriques en barre accompagnees de sueurs, et de syndromes inflammatoire biologique et infectieux chez ce patient sous previscan.

dans ses antecedents, on note un diabete non insulino-dependant, une hypertension arterielle, une phlebite du membre inferieur gauche, des bronchites a repetition.
l'angio-scanner realise elimine une embolie pulmonaire.
le scanner realise montre une vesicule biliaire de taille normale a contenu liquidien siege de petites structures evoquant de petits calculs.
la paroi vesiculaire est d'epaisseur normale.
deux echo-dopplers veineux realises a une semaine de distance par le @ @ @ la persistance d'une thrombo-phlebite femoro-poplitee droite.
l'examen biologique retrouve une crp inferieure a 3,
en l'absence de douleur abdominale, apres surveillance en milieu chirurgical, l'existence de petits calculs intra-vesiculaires sans signe inflammatoire de la paroi vesiculaire, apres avis du @ @ @, medecin anesthesiste, le traitement chirurgical de cette lithiase vesiculaire est reportee ulterieurement car il existe un fort risque de complications thrombo-emboliques post-operatoires.
j'autorise son retour a domicile avec reprise de son traitement anti coagulant, previscan : 1/2 par jour.
je le reverrai en consultation dans @ mois pour controle clinique et radiologique.
je reste a votre disposition pour le revoir avant cette date en cas de recidive de douleurs abdominales.
bien cordialement.
entree le @.@.@
sortie le @.@.@

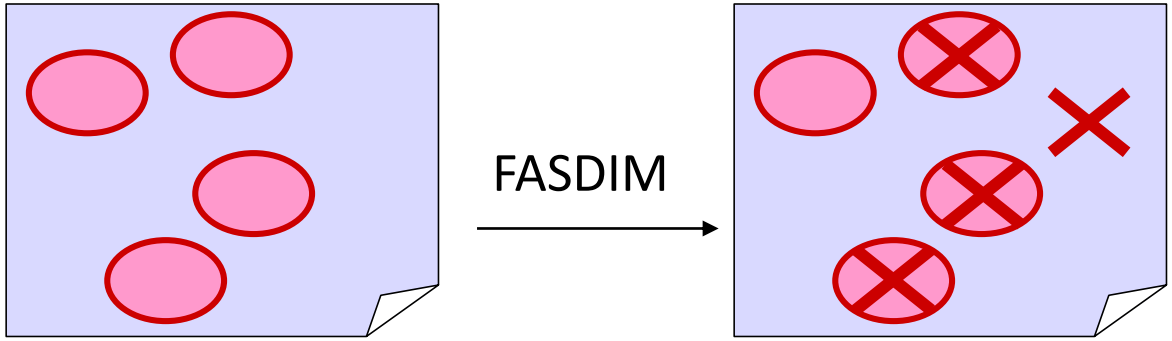
Évaluation : méthode et résultat

- Etape 1 : évaluation de l'efficacité de FASDIM
- Etape 2 : évaluation de la conservation de l'information médicale
- Etape 3 : évaluation de la charge de travail

Etape 1 : évaluation de l'efficacité de FASDIM



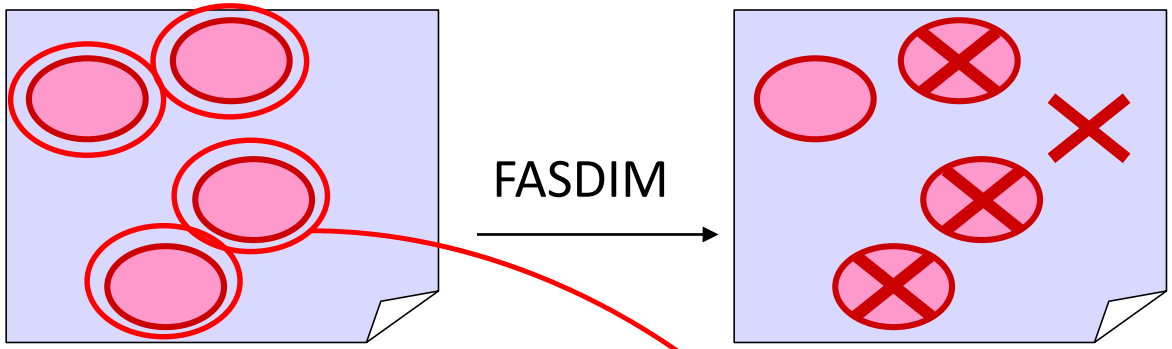
Méthode : Évaluation de l'efficacité
508 courriers de sortie tirés au sort



Courrier avec données identifiantes

Courrier dé-identifié

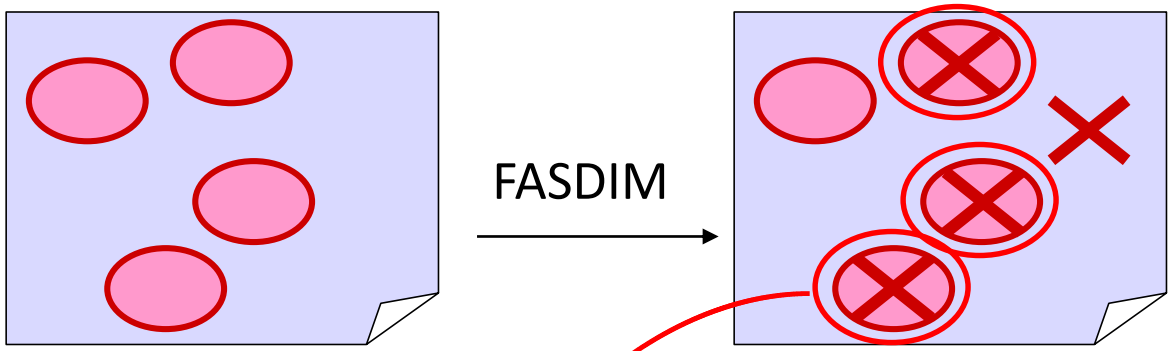
Méthode : Évaluation de l'efficacité 508 courriers de sortie tirés au sort



Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI			Nombre total de PHI
Non PHI			
Total			

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. Décompte de PHI restants dans les courriers dé-identifiés
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

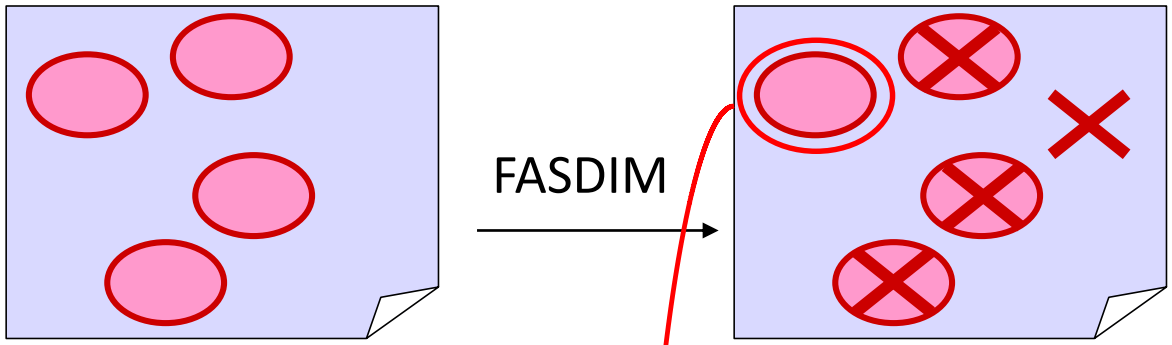
Méthode : Évaluation de l'efficacité 508 courriers de sortie tirés au sort



Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI	Vrais positifs		Nombre total de PHI
Non PHI			
Total			

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. Décompte de PHI restants dans les courriers dé-identifiés
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

Méthode : Évaluation de l'efficacité 508 courriers de sortie tirés au sort



Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI	Vrais positifs	Faux négatifs	Nombre total de PHI
Non PHI			
Total			

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. **Décompte de PHI restants dans les courriers dé-identifiés**
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

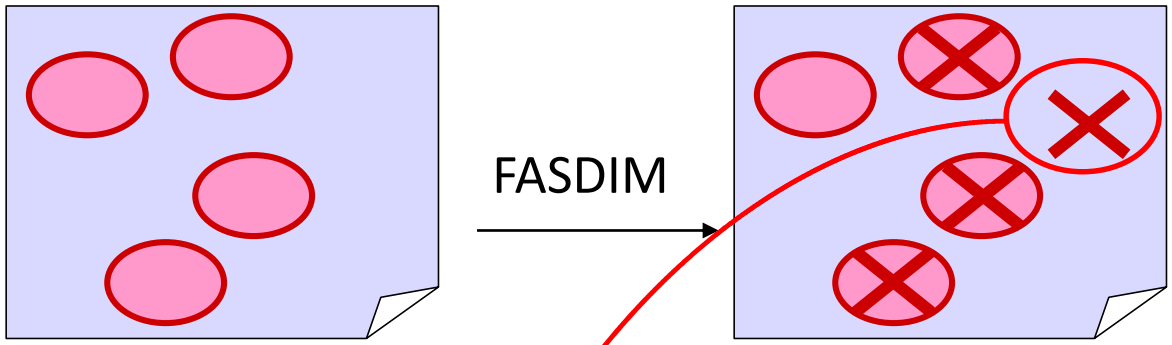
Méthode : Évaluation de l'efficacité 508 courriers de sortie tirés au sort

$$rappel = R = \frac{VP}{\#\ phi} = \frac{VP}{VP + FN}$$

Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI	Vrais positifs	Faux négatifs	Nombre total de PHI
Non PHI			
Total			

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. Décompte de PHI restants dans les courriers dé-identifiés
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

Méthode : Évaluation de l'efficacité
508 courriers de sortie tirés au sort



Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI	Vrais positifs	Faux négatifs	Nombre total de PHI
Non PHI	Faux positifs		
Total	# mots sup.		

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. Décompte de PHI restants dans les courriers dé-identifiés
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

Méthode : Évaluation de l'efficacité 508 courriers de sortie tirés au sort

$$precision = P = \frac{VP}{\#retirés} = \frac{VP}{VP + FP}$$

$$F - measure = F = \left(\frac{R^{-1} + P^{-1}}{2}\right)^{-1}$$

Mots	Supprimés par FASDIM	Ignorés par FASDIM	Total
PHI	Vrais positifs	Faux négatifs	Nombre total de PHI
Non PHI	Faux positifs		
Total	# mots sup		

1. Décompte du nombre de données identifiantes « PHI » dans les courriers non dé-identifiés
2. Décompte de PHI supprimés dans les courriers dé-identifiés
3. Décompte de PHI restants dans les courriers dé-identifiés
4. Calcul du rappel
5. Décompte des mots non PHI supprimés dans les courriers dé-identifiés
6. Calcul de la précision

Résultats
Évaluation de l'efficacité

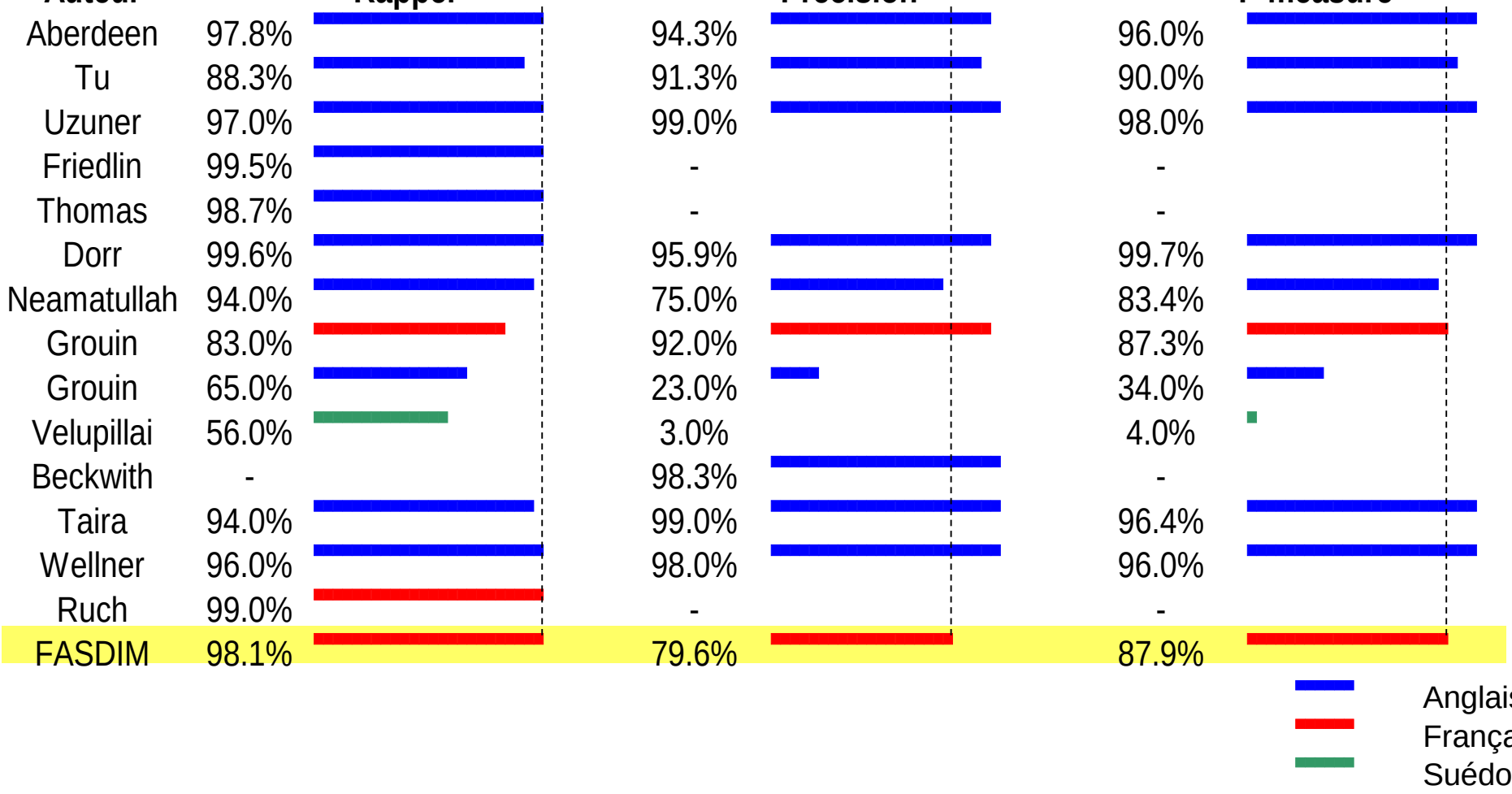
Mesure	Valeur
Nombre de courriers	508
Nombre moyen de mots par courrier	510
Nombre moyen de PHI par courrier	20
Rappel (R) proportion de PHI correctement supprimés	98,1 %
Précision (P) proportion de PHI parmi tous les mots supprimés	79,6 %
F-measure Moyenne harmonique de R et de P	87,9 %

Résultats
Détail des faux négatifs (PHI « oubliés »)

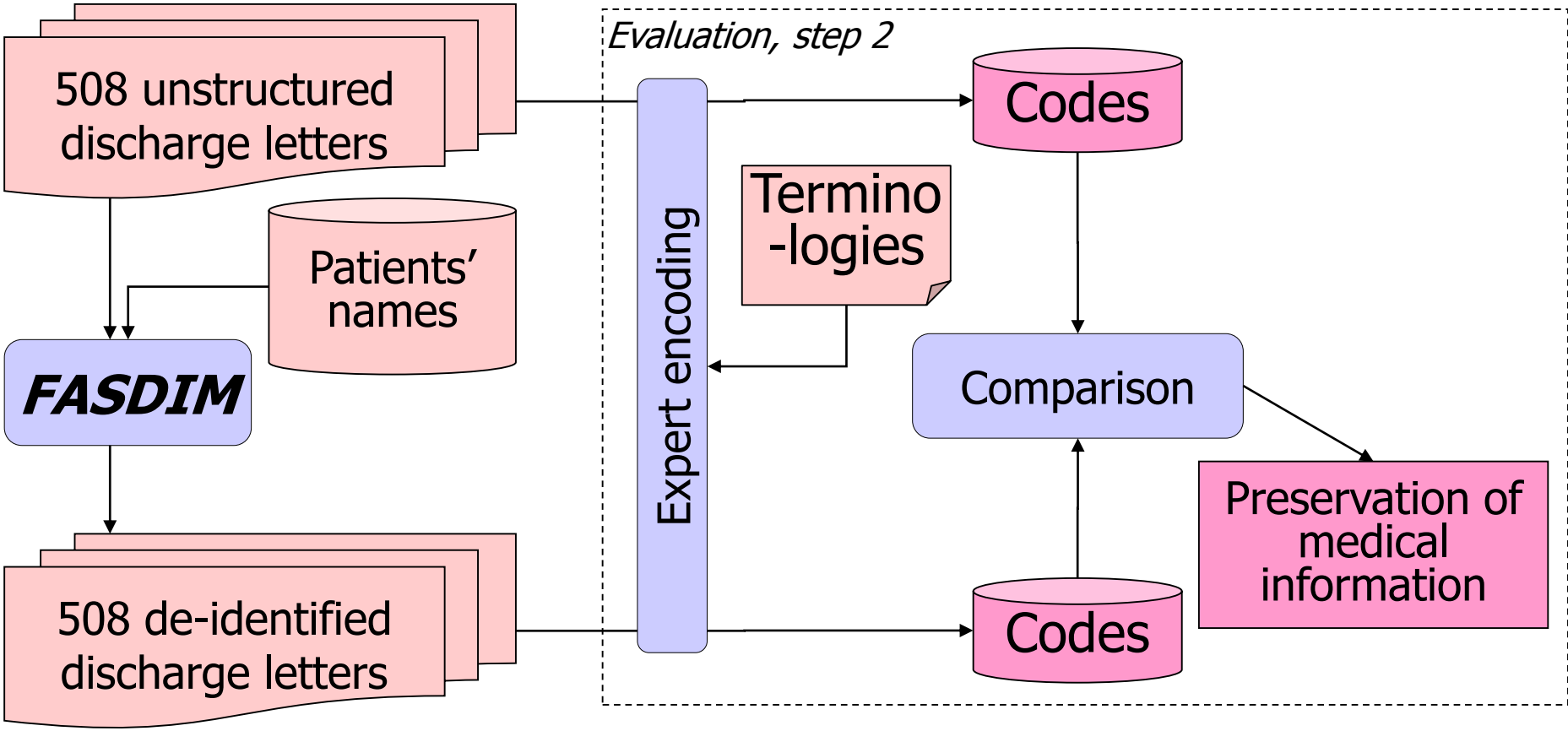
Catégorie de PHI non supprimés à tort	Proportion
Information partielle sur un lieu	63,7 %
Noms de soignant	23 %
Données biométriques (poids uniquement)	5,5 %
Portion de date ou d'âge	4,4 %
Noms de structure de soins	3,3 %
Noms de patient	0 %
Autres nombres	0 %

Résultats comparatifs

Auteur	Rappel	Précision	F-measure
Aberdeen	97.8%	94.3%	96.0%
Tu	88.3%	91.3%	90.0%
Uzuner	97.0%	99.0%	98.0%
Friedlin	99.5%	-	-
Thomas	98.7%	-	-
Dorr	99.6%	95.9%	99.7%
Neamatullah	94.0%	75.0%	83.4%
Grouin	83.0%	92.0%	87.3%
Grouin	65.0%	23.0%	34.0%
Velupillai	56.0%	3.0%	4.0%
Beckwith	-	98.3%	-
Taira	94.0%	99.0%	96.4%
Wellner	96.0%	98.0%	96.0%
Ruch	99.0%	-	-
FASDIM	98.1%	79.6%	87.9%



Etape 2 : évaluation de la conservation de l'information médicale



Résultats

Évaluation de la conservation de l'information médicale

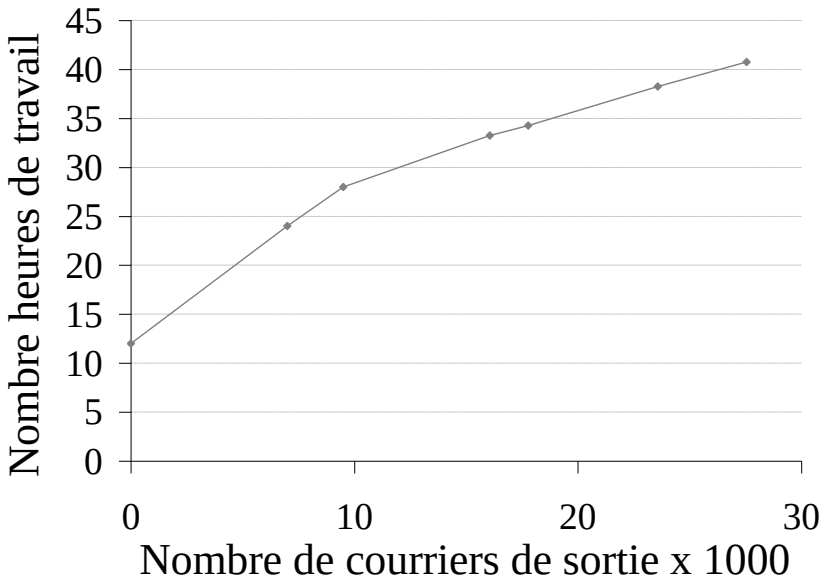
en moyenne 30,4 codes par courrier

Catégorie d'information médicale	Taux conservation
Toutes catégories	99,0 %
CCAM : actes	99,7 %
CIM10 :	
- Maladies, symptômes, motifs d'accès aux soins	99,5 %
- Actes	98,9 %
- Résultats anormaux de biologie	97,0 %
ATC (médicaments)	98,8 %

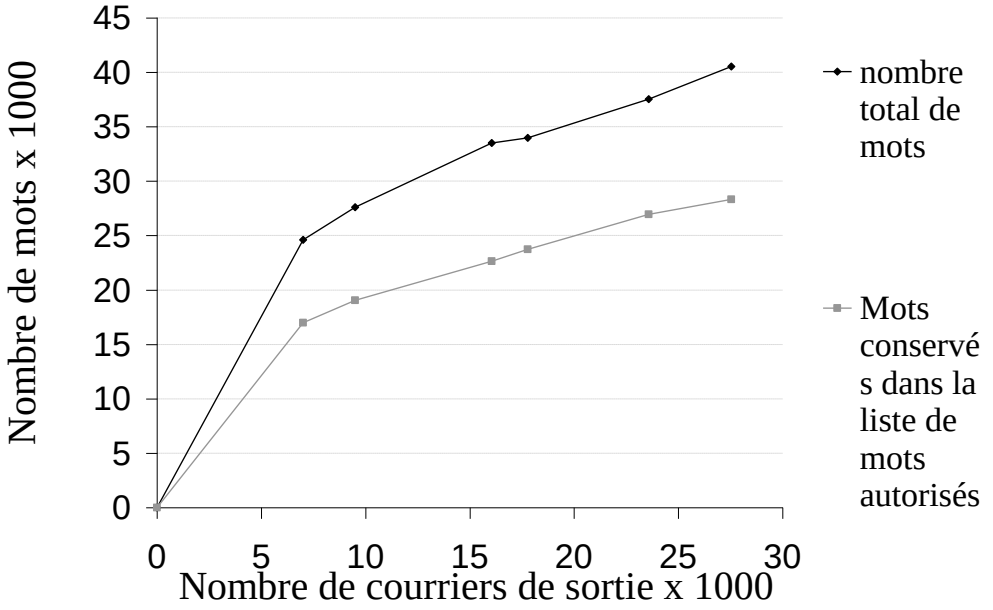
Etape 3 : évaluation de la charge de travail

- Mesure du temps de travail :
  - Implémentation initiale de la méthode
  - Mise à jour des listes de mots autorisés
  - Mise à jour des motifs de nombres
- Mesure du nombre de mots rencontrés/conservés

Résultats
Évaluation de la charge de travail



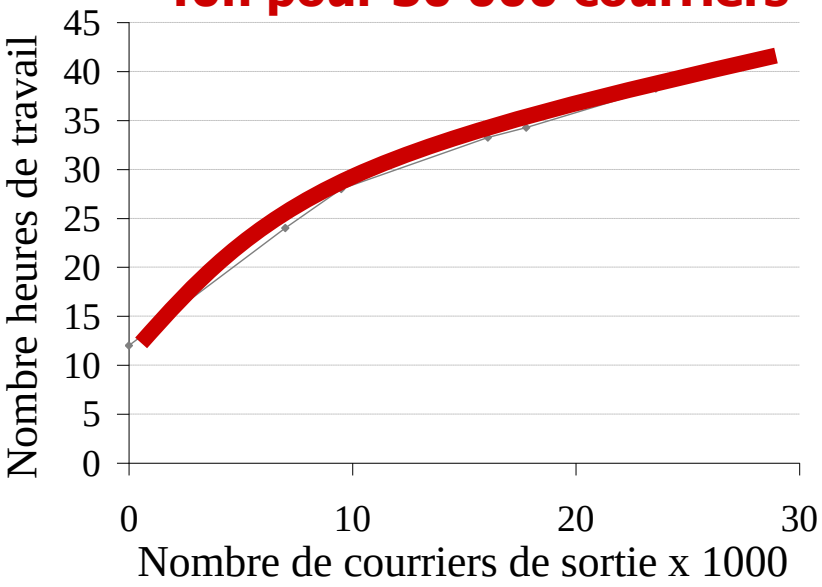
Nombre d'heures de travail en fonction du nombre de courriers de sortie à dé-identifier



Nombre de mots à trier en fonction du nombre de courriers de sortie à dé-identifier

Résultats
Évaluation de la charge de travail

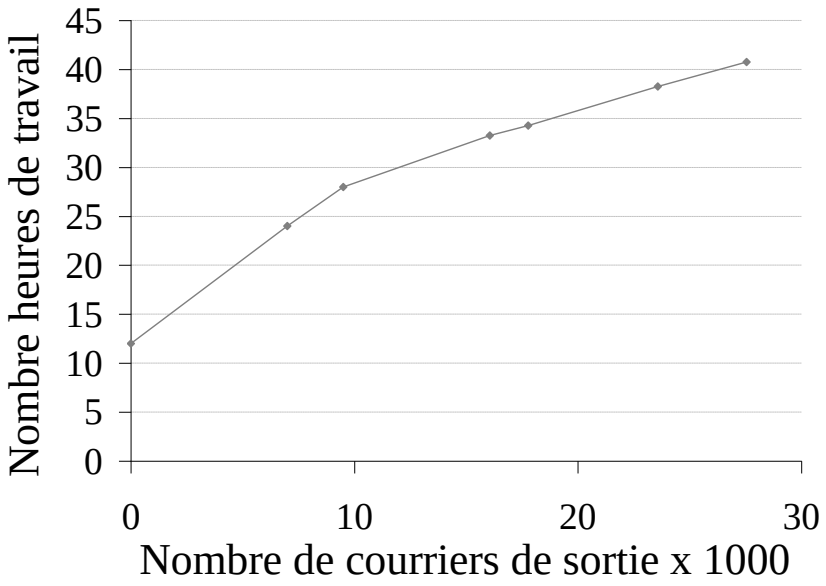
12h pour 0 courriers,
40h pour 30 000 courriers



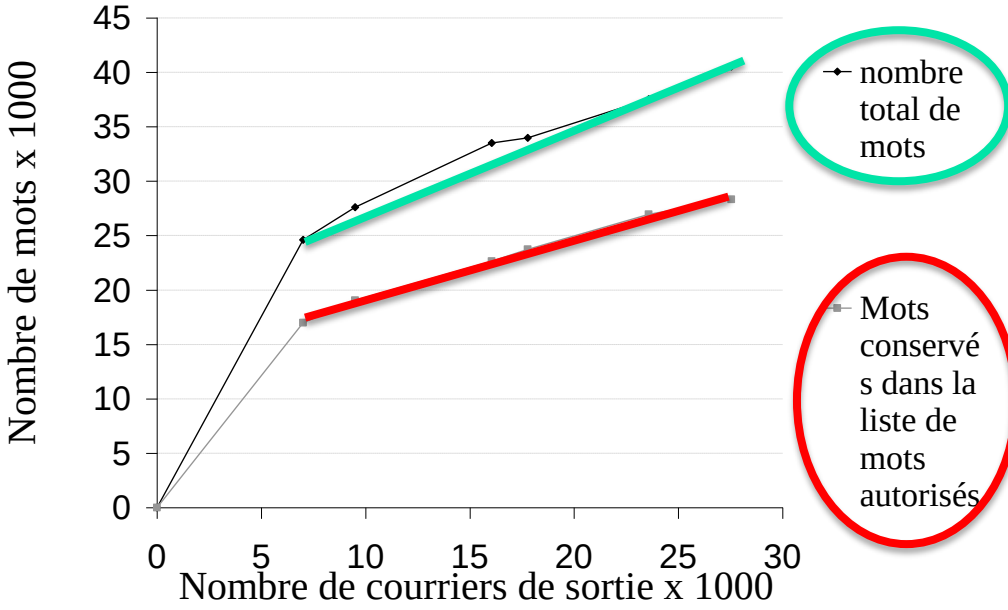
Nombre d'heures de travail en fonction du nombre de courriers de sortie à dé-identifier

Nombre de mots à trier en fonction du nombre de courriers de sortie à dé-identifier

Résultats
Évaluation de la charge de travail



Nombre d'heures de travail en fonction du nombre de courriers de sortie à dé-identifier



Nombre de mots à trier en fonction du nombre de courriers de sortie à dé-identifier

Conclusion

- FASDIM :
  - Excellents résultats, comparables à ceux obtenus en langue anglaise
  - Facilement reproductible sans matériel par les établissements de santé
  - Opérationnel très rapidement
- Publications :
  - Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart JB, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform. 2014 Apr;83(4):303-12. doi: 10.1016/j.ijmedinf.2013.11.005. Epub 2013 Dec 7.
  - Chazard E, Mouret C, Ficheur G, Beuscart R. Déidentification automatisée de courriers médicaux : la méthode FASDIM. Revue d'Épidémiologie et de Santé Publique, Volume 60, Supplement 1, March 2012, Page S18
- Code « open source » téléchargeable sur http://www.fasdim.com