FACULTE DE MEDECINE HENRI WAREMBOURG

Année 2012

THESE POUR L'OBTENTION DU DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

DE-IDENTIFICATION AUTOMATISEE DE COURRIERS MEDICAUX : PROPOSITION ET EVALUATION DE LA METHODE FASDIM

Présentée et soutenue publiquement le 17 Janvier 2012

Par Capucine MOURET-KUBIAK

Jury

Président :	**Monsieur le Professeur Beuscart**
Assesseurs :	**Monsieur le Professeur Frimat** **Monsieur le Professeur Duhamel** **Monsieur le Docteur Chazard**
Directeur de Thèse :	**Monsieur le Docteur Chazard**

Sommaire

Sigles utilisés

CCAM	*Classification Commune des Actes Médicaux*
CD	*Code de Déontologie*
CIM 10	*Classification Internationale des Maladies, 10ème révision*
CNIL	*Commission Nationale de l'Informatique et des Libertés*
CRF	*Conditional Random Field*
CSP	*Code de Santé Publique*
CT	*Code du Travail*
DINAMIT	*Dossier Informatisé pour les Autonomes et Médecines Interprofessionnelles du Travail*
DMST	*Dossier Médical Santé Travail*
FASDIM	*Fast and Simple De Identification Method*
FN	*Faux Négatif*
FP	*Faux Positif*
GISSET	*Groupement Inter Services Santé et Travail*
HAS	*Haute Autorité de Santé*
HIPAA	*Health Insurance Portability and Accountability Act*
IDS	*Integral Data Sante*
ISTNF	*Institut de Santé au Travail du Nord de la France*
MIRT	*Médecin Inspecteur Régional du Travail*
NAF	*Nomenclature d'Activités Française*
NER	*Name Entity Recognition*
NIH	*National Institute of Health*
NLM	*National Library of Medicine*
NLP	*Natural Language Processing*
OMS	*Organisation Mondiale de la Santé*
PHI	*Protected Health Information*
PHP	*Hypertext Preprocessor*
PSIP	*Patient Safety through Intelligent Procedures*
SIH	*Système d'Information Hospitalier*
SPP	*Suivi Post Professionnel*
SVM	*Support Vector Machine*
UMLS	*Unified Medical Language System*
VN	*Vrai Négatif*
VP	*Vrai Positif*
VPN	*Valeur Prédictive Négative*
VPP	*Valeur Prédictive Positive*

Introduction

I. Courriers médicaux : définition

A. Le dossier médical du patient hospitalisé

1. Définition

La prise en charge des patients hospitalisés impose une communication entre les différents membres de l'équipe soignante et une traçabilité assurant le suivi du patient. Les éléments relatifs à la santé et à la prise en charge de chaque patient sont consignés dans le dossier médical. Le dossier médical est défini en partie par la loi n°2002-303 du 4 mars 2002, relative « aux Droits des malades et à la qualité du système de santé », dite aussi « loi Kouchner » [CSP L.1111-7] comme étant l'« ensemble des informations concernant sa [du patient] santé détenues, à quelque titre que ce soit, par des professionnels et des établissements de santé, qui sont formalisées ou ont fait l'objet d'échanges écrits entre professionnels de santé […] ».

2. Contenu

Le contenu du dossier médical est précisé dans l'article R 710-2-2 du Code de Santé Publique [CSP R.710-2-2] et est présenté dans le Tableau 1.
Les seules informations communicables, selon l'article R710-2-2 du Code de Santé Publique [CSP R710-2-2], sont celles mentionnées dans les deux premières sections du Tableau 1.
Tout ou partie de ces éléments peut être informatisé au sein de l'établissement de santé.

Informations formalisées recueillies lors des consultations externes dispensées dans l'établissement, lors de l'accueil au service des urgences ou au moment de l'admission et au cours du séjour hospitalier
• La lettre du médecin qui est à l'origine de la consultation ou de l'admission • Les motifs d'hospitalisation • La recherche d'antécédents et de facteurs de risques • Les conclusions de l'évaluation clinique initiale • Le type de prise en charge prévu et les prescriptions effectuées à l'entrée • La nature des soins dispensés et les prescriptions établies lors de la consultation externe ou du passage aux urgences • Les informations relatives à la prise en charge en cours d'hospitalisation : état clinique, soins reçus, examens paracliniques, notamment d'imagerie • Les informations sur la démarche médicale, adoptée dans les conditions prévues à l'article L. 1111-4 [CSP L.1111-4] • Le dossier d'anesthésie • Le compte-rendu opératoire ou d'accouchement • Le consentement écrit du patient pour les situations où ce consentement est requis sous cette forme par voie légale ou réglementaire • La mention des actes transfusionnels pratiqués sur le patient et, le cas échéant, copie de la fiche d'incident transfusionnel mentionnée au deuxième alinéa de l'article R. 666-12-24 [CSP R.666-12-24] • Les éléments relatifs à la prescription médicale, à son exécution et aux examens complémentaires • Le dossier de soins infirmiers ou, à défaut, les informations relatives aux soins infirmiers • Les informations relatives aux soins dispensés par les autres professionnels de santé • Les correspondances échangées entre professionnels de santé
Informations formalisées établies à la fin du séjour
• Le compte-rendu d'hospitalisation et la lettre rédigés à l'occasion de la sortie • La prescription de sortie et les doubles d'ordonnance de sortie • Les modalités de sortie (domicile, autres structures) • La fiche de liaison infirmière
Informations mentionnant qu'elles ont été recueillies auprès de tiers n'intervenant pas dans la prise en charge thérapeutique ou concernant de tels tiers.

Tableau 1. Informations contenues dans le dossier médical

B. Les courriers médicaux

Parmi ces éléments, les courriers médicaux contiennent une certaine quantité d'information médicale. On appelle ici courrier médical toute lettre ou document établis par un médecin, ce qui comprend :

- les lettres de demande d'admission
- les comptes-rendus d'actes (opératoires, d'accouchement, d'explorations fonctionnelles ou d'examens complémentaires)
- les lettres de sortie d'hospitalisation, qui sont des comptes-rendus d'hospitalisation essentiellement destinés à des médecins extérieurs à l'établissement

Les courriers médicaux ont une place essentielle dans les dossiers médicaux, puisqu'ils synthétisent les différents événements survenus, en reprenant les antécédents, la démarche diagnostique, les résultats d'examens complémentaires, la

description de la prise en charge effectuée par le médecin rédigeant le courrier et la conclusion médicale ainsi que les consignes permettant de poursuivre les soins.

II. L'intérêt d'anonymiser les courriers

Les courriers médicaux informatisés constituent une source importante de données exploitables dans différents domaines : essentiellement pour la recherche médicale mais également en santé publique, et peut-être bientôt en santé travail.
Les courriers médicaux anonymisés forment une base de données pouvant contenir un nombre quasiment illimité de séjours hospitaliers, exploitables dans le cadre de lecture experte ou de la fouille de données automatisée, détaillées ci-dessous.
Ces courriers pourraient être exploités dans le milieu assurantiel : l'Assurance Maladie pourrait automatiser un précontrôle à distance des dossiers qui seraient dé-identifiés. A l'issue de ce pré-contrôle à distance, sélectionnant les dossiers comprenant les critères choisis selon l'enquête, un contrôle à dire d'experts par des médecins inspecteurs pourrait être effectué sur les dossiers correspondants (dossiers non anonymes dans ce cas) dans l'établissement.

A. Contexte législatif

Afin de manipuler les données contenues dans les dossiers médicaux, une condition absolue est requise : la vie privée du patient doit être respectée selon l'article L 1110-4 du Code de la Santé Publique [CSP L.1110-4] (modifié par l'article 2 de la loi du 13 août 2004 du Code de la Sécurité Sociale), l'article 29 de la loi CNIL n°78-17 du 6 janvier 1978 [CNIL 2011 (1)] ainsi que les articles 4,72 et 73 du Code de Déontologie [CD 4 ; CD 72 ; CD 73]. Des sanctions sont prévues en cas de manquements à cette obligation [CP 226-13, CP 226-14] : « La révélation d'une information à caractère secret par une personne qui en est dépositaire soit par état ou par profession, soit en raison d'une fonction ou d'une mission temporaire, est punie d'un an d'emprisonnement et de 15 000 euros d'amende ».

B. Protection de la vie privée du patient

Il existe deux manières de procéder pour remplir cette condition : soit le consentement libre et éclairé du patient est recueilli afin d'utiliser les informations le concernant, soit les documents sont traités dans le but d'enlever toutes les informations susceptibles d'entraîner l'identification du patient, ce qui est appelé « dé-identification ». Chaque approche exige une charge de travail et de temps importante. Certaines exploitations des documents ne nécessitent pas de retrouver l'identité du patient. Par exemple, des travaux d'épidémiologie hors registres ou une recherche d'association entre des expositions et un effet, dont le but n'est pas d'indemniser les personnes, mais d'accroître la connaissance scientifique ne demandent pas de suivi des patients dans le temps.

Il faut toutefois noter que la dé-identification ne dispense pas les personnes détentrices de l'information médicale d'informer les patients (ou les salariés dans le cas de la Santé au Travail) que les informations médicales les concernant peuvent être utilisées à des fins de recherche, dans l'anonymat complet, et qu'ils ont un droit d'opposition à cette exploitation.

C. Utilisation de courriers dé-identifiés

L'utilité des courriers médicaux dans ce contexte peut revêtir deux aspects : soit ces courriers sont utiles pour effectuer une lecture à dire d'experts de cas après une requête, soit le contenu de ces courriers est exploité de manière automatisée afin d'en extraire des informations médicales.

1. Lecture experte

Des bases de données informatisées contenant des informations médicales peuvent être interrogées lors de projets de recherche ou d'études épidémiologiques, pour en retirer des données et dégager des tendances : il s'agit de requête informatique.
Toutefois, les résultats obtenus nécessitent le plus souvent une lecture des courriers médicaux par des experts afin d'être validés. Cette lecture s'appuie en partie sur des courriers en texte libre contenant à la fois des données médicales et des données personnelles.

2. Exploitation automatisée

Un travail de recherche au sein des hôpitaux pourrait être mené avec l'objectif de trouver des associations entre des professions, des expositions professionnelles et des pathologies décrites dans les textes libres contenus dans les dossiers médicaux, ajoutées à l'analyse des diagnostics ou des actes d'hospitalisation (ou des actes ambulatoires) disponibles dans le Système d'Information Hospitalier (SIH). Cette recherche d'associations pourrait se faire en combinant des techniques de *Data Mining* et de *Semantic Mining*.
Le *Data Mining*, ou fouille de données en français, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques [Wikipédia 2011 (1)]. Les méthodes de fouille de données permettent donc de dégager du sens d'un ensemble de données d'allure a priori disparate et créent de la sémantique. La sémantique dégagée prend généralement trois formes (traduction par des signifiants formels) issues de l'intelligence artificielle : la tableau, le graphe ou l'arbre (cas particulier de graphe nécessitant une théorie et une exploitation spécifique). Ce sont des signifiants, dans le sens où ils représentent des connaissances. De telles structures sont ensuite annotées dans les données de départ, chaque donnée portant alors la marque de son appartenance à une branche de l'arbre, une case du tableau, etc. L'analyse prend alors un niveau de compréhension plus complexe.
Dans ces cas, il est moins fastidieux de recourir à la dé-identification plutôt qu'au recueil du consentement libre et éclairé du patient puisque ces démarches s'appliquent à de très grandes quantités de documents (plusieurs dizaines de milliers de courriers). En effet, dès son admission, le patient est informé que les informations concernant sa santé peuvent être utilisées à des fins de recherche, il reçoit

également l'information qu'il est en droit de s'opposer à l'exploitation des données le concernant. Sa non-opposition fait donc office d'accord global quant à l'utilisation des données de santé non nominatives le concernant.
D'autre part, la dé-identification est la solution la plus sécurisée lorsqu'elle est possible en vue d'assurer le respect de la vie privée de la personne (recommandations de la Commission Nationale de l'Informatique et des Libertés - CNIL) [CNIL 2011 (2)].
De plus, les contraintes de protection de la vie privée du patient sont tellement strictes qu'elles peuvent générer des blocages dans la transmission d'information entre différents services. Ce frein à l'acquisition de données au sein d'un établissement est résolu en partie par la dé-identification des documents.

III. État de l'art de la dé-identification des courriers

Ce travail s'intéressera au cas plus particulier des courriers médicaux, ce terme englobant ici :

- les comptes-rendus d'actes diagnostiques (certaines explorations endoscopiques ou fonctionnelles)
- les comptes-rendus d'actes thérapeutiques (accouchements, comptes-rendus opératoires)
- les courriers de sortie

A. Définitions

Il est important de noter que deux termes sont fréquemment employés en tant que synonymes en langue anglaise : il s'agit des termes « anonymization », que l'on pourrait traduire en français par « anonymisation », et « de-identification » dont l'équivalent français serait « dé-identification ». Néanmoins selon certains auteurs, ces deux notions sont différentes.

Anonymisation : est le fait de rendre anonyme (*the anonymization* en anglais), consiste à enlever les noms et prénoms d'une personne dont on veut conserver l'anonymat.

Dé-identification : est le fait de rendre non identifiable (*the de-identification* en anglais) un document, ce qui s'avère plus complexe puisque toutes les données susceptibles d'identifier une personne doivent être retirées. Cela comprend le nom, mais aussi la date de naissance, l'adresse postale, toutes les données identifiantes telles qu'elles sont définies ci-dessous sous le terme « PHI ». La dé-identification est donc un concept plus large que l'anonymisation.
Notre objectif est donc bien de rendre des documents « non identifiables ». Toutefois, la CNIL et la plupart des auteurs francophones emploient le terme « anonymisation ». Par souci de rigueur et de cohérence scientifique, nous utiliserons le terme plus exact de « dé-identification » tout au long de ce travail.

Données identifiantes « PHI » : Les données dites « identifiantes » (*Protected Health Information* en anglais, « PHI ») sont décrites dans la section « Matériel » (voir Matériel, II, page 28).

Pseudonymisation : La pseudonymisation est le fait de remplacer des données identifiantes par des données fictives réalistes de même catégorie d'information (un nom de famille est remplacé par un autre nom de famille, une date par une autre date, etc.).
La pseudonymisation peut être réversible ou irréversible, selon qu'un lien est maintenu entre les données d'origine et les pseudonymes. Dans le cas de la pseudonymisation réversible, le retour aux données originales identifiantes est possible grâce à des clés secrètes connues des seuls utilisateurs habilités à avoir accès aux informations.
L'intérêt de la pseudonymisation est le maintien du confort de lecture en créant des documents très réalistes. La sécurité est également renforcée quant au respect de la confidentialité. En effet, s'il demeure un nom de patient dans le document pseudonymisé, rien ne permet de deviner qu'il s'agit du véritable nom d'origine.

B. Métriques d'évaluation des procédés de dé-identification

L'évaluation des méthodes de dé-identification retrouvée dans la littérature montre l'usage de trois métriques fréquemment utilisées. Ce point est détaillé dans la section Méthode (voir Méthode, II.A.2, page 45).

1. Rappel

Le rappel est la proportion de données identifiantes retirées de manière pertinente du texte parmi celles qui devraient être retirées. Le rappel reflète l'efficacité de la méthode, il s'agit de la sensibilité. La formule est présentée dans la Formule 1.

$$Rappel = R = \frac{VP}{\#identifiants} = \frac{VP}{VP + FN}$$

Formule 1. Rappel

2. Précision

La précision est la proportion de données identifiantes enlevées sur l'ensemble des données supprimées par la méthode de dé-identification dans le document. La précision correspond à la Valeur Prédictive Positive (VPP). La formule est présentée dans la Formule 2.

$$Précision = P = \frac{VP}{\#mots\ supprimés} = \frac{VP}{VP + FP}$$

Formule 2. Précision

3. F-measure

La *F-measure* est une moyenne harmonique des deux scores précédents (Rappel et Précision). Cette mesure permet de comparer les différentes méthodes avec un seul chiffre. La formule est présentée dans la Formule 3.

$$F\_measure = F = \frac{1}{\frac{1}{2} * (\frac{1}{P} + \frac{1}{R})}$$

Formule 3. F-measure

C. Les différents procédés existants

Rendre les documents non identifiables peut être obtenu soit manuellement en analysant et en supprimant les informations sensibles, soit de manière automatisée via un programme informatique. La première méthode, si elle semble plus sûre en termes de confidentialité, présente néanmoins l'inconvénient d'être fastidieuse et coûteuse en termes de temps de travail [Dorr 2006] et de personnel requis. La deuxième méthode permet de traiter un nombre quasiment illimité de documents : il existe de nombreux outils en anglais. Une revue récente [Meystre 2010] fait un état des lieux de ces méthodes et de leur évaluation. Ces méthodes recourent à deux types de procédés :

- d'une part le *pattern matching* [Beckwith 2006, Berman 2003, Fielstein 2004, Friedlin 2008, Gupta 2004, Morrison 2009, Neamatullah 2008, Ruch 2000 Sweeney 1996]
- d'autre part le *machine learning* [Aramaki 2006, Gardner 2008, Hara 2006, Szarvas 2007, Uzuner 2008, Wellner 2007]

Certaines méthodes combinent les deux approches.

1. Le pattern matching

Le *pattern matching* consiste à élaborer des règles permettant de garder ou de rejeter des mots contenus dans les documents selon qu'ils appartiennent ou non à des listes (appelées également dictionnaires mais il s'agit plutôt de lexiques) pré-établies par des experts ou récupérées auprès d'organisations officielles.

> *Exemples :*
> - *liste de noms propres*
> - *liste de noms de communes*
> - *liste de termes médicaux – Unified Medical Language System [UMLS 2009]*
> - *liste de noms d'hôpitaux*

L'inconvénient de cette méthode est qu'elle requiert la construction d'une base de données importante, élaborée manuellement. Cette méthode est alors très

performante sur le secteur ciblé : par exemple, pour anonymiser des courriers provenant des États-Unis, les listes établies comprennent les noms des citoyens américains, les noms des villes des différents états, etc. et permettent d'extraire des documents les noms contenus dans les listes. La conséquence est que l'exportation d'une telle méthode perd énormément en performance dès lors qu'elle est transposée dans un autre pays : les listes ne sont plus exhaustives.

Sweeney *et al.* ont développé plusieurs méthodes : *Scrub system* puis *Datafly* [Sweeney 1996]. Ces méthodes utilisent de multiples algorithmes de détection des données personnelles, regroupées sous le terme de « PHI » pour *Protected Health Information*, associés à une catégorisation de ces PHI dans les documents en texte libre. Leurs résultats sur 275 dossiers pédiatriques (incluant 3 198 lettres) montrent une valeur prédictive positive de 99 %. Le taux de faux positifs et le détail du corpus d'évaluation ne sont pas fournis.

Le *De-identifier* de Berman *et al.* tire partie du métathésaurus de l'UMLS, une collection de vocabulaire médical sponsorisée par le *National Institute of Health* (NIH) et développé par la *National Library of Medicine* (NLM) [Berman 2003]. Le principe est que les mots qui n'appartiennent ni à cette nomenclature ni à une liste de *stop words* sont des PHI et sont donc retirés du texte.

Neamatullah *et al.* ont développé DE-ID *de-identifier*, qui repose sur des dictionnaires construits manuellement et des expressions rationnelles [Neamatullah 2008]. Leur système obtient un rappel de 96,7% et une précision de 74,9% sur un corpus de 2 434 notes infirmières.

DE-ID *de-identifier* est utilisé par Velupillai *et al.* [Velupillai 2009] et Grouin *et al.* [Grouin 2009] pour être traduit respectivement en suédois et en français ; de même, ce système a été repris par Tu *et al.* [Tu 2010] dans l'objectif de l'adapter pour la région de l'Ontario au Canada.

Gupta *et al.* ont conçu une méthode de dé-identification, évaluée sur des rapports d'anatomopathologie [Gupta 2004]. Leur approche localise les portions de phrases ne contenant pas de données personnelles et sépare le reste du texte en ajoutant des tags de dé-identification. Pour la dé-identification de phrases contenant des informations médicales pertinentes, l'algorithme utilise le métathésaurus de l'UMLS. L'évaluation est réalisée sur des rapports chirurgicaux et consiste à reporter le taux de faux positifs et de faux négatifs. Le rappel et la précision n'ont pas été calculés.

Beckwith *et al.* identifient les PHI qui figurent dans les en-têtes (par exemple, les numéros de dossiers médicaux et les noms de patients) puis suppriment ces PHI des documents [Beckwith 2006]. Afin de repérer les dates, les numéros identifiants et les adresses, des motifs sont utilisés ; de même, les marqueurs de civilité ou de titres (comme Monsieur, Madame, Docteur, etc.) sont employés pour localiser ces concepts dans le texte et ensuite les supprimer. Le traitement du texte se termine par une comparaison du texte d'origine avec une base de données contenant des noms propres. Beckwith *et al.* rapportent qu'ils suppriment 98,3 % d'identifiants dans les rapports d'anatomopathologie provenant de 3 institutions. Ils précisent également que 2,6 mots non PHI par document sont enlevés.

Ruch *et al.* emploient des méthodes de *Natural Language Processing* (NLP) afin de désambiguïser les mots et les reconnaître [Ruch 2000]. Ils utilisent également des règles contextuelles pour identifier les PHI. Le système trouve 98-99 % des informations identifiantes personnelles sur le corpus de test.

2. *Le machine learning*

Le *machine learning* regroupe plusieurs méthodes, plus récentes dans le domaine informatique. Le principe est le suivant : un programme informatique apprend à reconnaître ce qu'est une donnée identifiante à partir d'un corpus de textes annotés par des experts, puis est ensuite capable de les reconnaître et de les enlever dans des textes inconnus. Ces méthodes sont très performantes, et à l'instar des méthodes basées sur des règles et des dictionnaires, peuvent être étendues dans différents domaines. L'inconvénient de cette approche réside dans la constitution du corpus d'entraînement : celui-ci doit être le plus complet possible, c'est-à-dire présenter un maximum de cas de figures afin d'assurer la performance de la méthode, ce qui nécessite un temps de travail conséquent par des experts. Cet inconvénient est en partie contourné par le partage de ces corpus d'entraînement.
En 2006, i2b2 a organisé un concours de dé-identification [Uzuner 2007], les corpus d'entraînement étaient donc partagés et les résultats ont pu être comparés entre les participants. Les données d'entraînement et de test sont composées de 889 courriers de sortie contenant respectivement 470 000 mots pour le set d'entraînement et 140 000 mots pour le set de test. La méthode avec la meilleure performance atteint des scores proches de 98 % pour la F-measure pour toutes les catégories.
Uzuner *et al.* remarquent que les méthodes d'apprentissage statistique obtiennent les meilleures performances, suivies par les méthodes hybrides combinant les règles et du *machine learning*, les systèmes de *machine learning* purs sans règles ni modèle et enfin les systèmes basés uniquement sur des règles [Uzuner 2007].
Szarvas *et al.* utilisent une approche basée sur la reconnaissance d'entités nominales (*Named Entity Recognition* - NER) par des systèmes statistiques destinée à être appliquée sur des documents semi structurés [Szarvas 2007].
Aramaki *et al.* ont élaboré une méthode de dé-identification basée sur une approche de *Conditional Random Field* (CRF) combinant des caractéristiques non-locales avec des caractéristiques locales pour identifier les PHI [Aramaki 2006].
Hara *et al.* ont adopté une méthode utilisant le *Support Vector Machine* (SVM) et des règles [Hara 2006].
Wellner *et al.* ont adapté deux outils de NER, Carafe et Lingpipe pour développer une méthode de dé-identification [Wellner 2007].
Uzuner *et al.* utilisent le SVM pour identifier le caractère PHI ou non de mots individuels [Uzuner 2007]. La méthode est basée sur l'orthographe, le contexte local et les caractéristiques des mots. La structure définie se sert de cette représentation du contexte local pour repérer des PHI dans les textes cliniques.
Comme le rapportent Aberdeen *et al.* et Meystre *et al.*, les systèmes de *machine learning* ne requièrent que peu de développements supplémentaires des programmes pour conserver de bons résultats lorsqu'ils sont appliqués sur des documents de même langage [Aberdeen 2010, Meystre 2010]. Ainsi, certaines équipes ayant participé au concours de dé-identification organisé par i2b2 ont pu démontrer que les résultats obtenus sur le corpus de test imposé lors de la compétition sont similaires à ceux obtenus sur leur propre corpus d'évaluation. L'adaptation est rapide : quatre heures par exemple pour Aberdeen *et al.* [Aberdeen 2010].

3. Exportabilité des méthodes de dé-identification

La plupart des méthodes rapportées dans la littérature sont élaborées en langue anglaise et sont difficilement exportables dans d'autres langages, c'est-à-dire que la simple traduction des listes par exemple, ne permet pas d'obtenir des documents anonymisés dans la langue choisie ; en effet les règles établies sont propres à la langue traitée. Il en va de même pour les méthodes de *machine learning*, bien que les articles concernant cette exportabilité partent des méthodes basées sur des listes associées à des règles. Ainsi, [Velupillai 2009] et [Grouin 2009] ont tenté d'adapter la méthode de [Neamatullah 2008] (car libre de droits) devant des besoins similaires de dé-identification automatisée sur des documents en langue danoise. Pantazos *et al.* ont développé leur propre méthode et ont obtenu un rappel de 99,5 %, une précision de 92,3 % et une F-measure de 95,7 % [Pantazos 2011].

4. Les méthodes de dé-identification en langue française

Cependant aucune méthode en langue française n'est actuellement disponible de manière libre de droits ; d'autre part, l'adaptation en français de méthodes anglaises a entraîné un très grand nombre de suppressions inappropriées de mots dans les documents [Grouin 2009], avec un rappel à 65 % pour une précision de 23 %. Cette même équipe a donc développé sa propre méthode, évaluée sur un faible échantillon, fournissant de meilleurs résultats (rappel à 83 % et précision à 92 %) mais encore insuffisants pour garantir la confidentialité du patient.

5. Synthèse des performances

Le Tableau 2 ci-dessous récapitulatif des différentes méthodes présentées dans la littérature compare les résultats obtenus entre les différents auteurs cités précédemment.

Référence	Méthode	Langue	Précision	Rappel	F-measure
Aberdeen 2010	Machine learning	EN	94,3 %	97,8 %	96 %
Tu 2010	Pattern matching		91,3 %	88,3 %	90 %
Uzuner 2007	Machine learning	EN	99 %	97 %	98 %
Friedlin 2008	Pattern matching	EN	-	99,47 %	-
Thomas 2002	Pattern matching	EN	-	98,7 %	-
Dorr 2006	Manuelle	EN	95,9 %	99,6 %	-
Neamatullah 2008	Pattern matching	EN	75 %	94 %	-
Grouin 2009	Pattern matching	FR EN	92 % 23 %	83 % 65 %	- -
Velupillai 2009	Pattern matching	SW	3 à 9 %	56 à 76 %	4 à 16 %
Beckwith 2006	Pattern matching	EN	98,3 %	-	-
Taira 2002	Pattern matching et Machine learning	EN	99 %	94 %	-
Wellner 2007	Machine learning	EN	98 %	96 %	96 %
Sweeney 1996	Pattern matching	EN	-	-	-
Aramaki	Machine learning	EN	-	-	-
Szarvas 2007	Machine learning	EN	-	-	96 %
Ruch 2000	Pattern matching	FR	-	99 %	-

Tableau 2. Récapitulatif des résultats obtenus par les différents auteurs cités

IV. Objectifs de ce travail

A. Objectifs généraux

Les objectifs sont les suivants : construire avec un minimum de matériel une méthode de dé-identification des courriers médicaux, cette méthode devant être performante, c'est-à-dire assurer la confidentialité des données, et devant générer des documents exploitables, c'est-à-dire sans perte d'information médicale. Un hôpital doit pouvoir facilement et rapidement mettre en place la méthode sans matériel autre que les documents à dé-identifier.
L'objectif principal de ce travail est d'élaborer FASDIM, *a Fast and Simple De-Identification Method*, permettant de rendre des documents en texte libre exempts de données personnelles identifiantes.

B. Objectifs opérationnels

1. Définir la méthode

Le premier objectif est de définir le concept de la méthode de dé-identification : définir les différentes étapes nécessaires en fonction du type de données identifiantes à supprimer dans le texte.

2. Implémenter la méthode

Le deuxième objectif est d'implémenter la méthode en PHP, acronyme récursif pour *PHP : Hypertext Preprocessor*, qui est un langage de programmation orienté objet généraliste et Open Source, initialement conçu pour le développement d'applications web et également utilisable en ligne de commande ou en *Graphic User Interface* [PHP 2011].

3. Évaluer la méthode

A l'issue de cette implémentation, la méthode doit être évaluée sur trois points :

- La première évaluation porte sur l'efficacité de la méthode, c'est-à-dire sa capacité à assurer la confidentialité des données tout en protégeant le contenu des documents.
- Ce deuxième point est évalué plus en détail : l'évaluation de la conservation de l'information médicale permet de mesurer la capacité de la méthode à générer des documents exploitables sur le plan médical. Effectivement, il est possible qu'une méthode supprime à tort certains mots d'un document, sans pour autant que cela nuise à l'interprétabilité médicale de ce document. A notre connaissance ce point n'a jamais été investigué.
- Enfin, la charge de travail relative à l'élaboration du programme ainsi qu'à sa mise à jour est également quantifiée et exprimée en nombre d'heures en fonction du nombre de courriers à dé-identifier.

4. Évaluer l'utilisation de la méthode en Santé Travail

Le dernier objectif opérationnel est de montrer le potentiel d'application d'une telle méthode de dé-identification en santé travail.

C. Objectifs annexes

1. Exportation de la méthode

Un établissement de santé souhaitant dé-identifier une base de données informatisée contenant des documents en texte libre, doit pouvoir implémenter facilement et rapidement la méthode sans matériel préalable, à partir de la seule lecture de ce travail. Grâce à la présentation de l'évaluation de la charge de travail, l'établissement voulant mettre en place la méthode saura exactement le temps de travail requis, et sera complètement indépendant dans sa démarche : aucune autre aide ne sera requise.

2. Indépendance de la langue

Le concept de la méthode doit permettre de la reproduire dans d'autres langues avec la même efficacité. Autant que possible, la procédure à suivre ne doit pas prendre en compte la grammaire, la syntaxe. Les mots sont traités indépendamment, aucun lien grammatical n'est établi entre eux.

> *Exemples : sont considérés comme deux mots différents :*
> - *« mange » et « mangeait »*
> - *« lien » et « liens »*

Cet objectif ne pourra être atteint naturellement que pour les langues dont l'écriture repose sur un alphabet, latin de préférence.

Matériel

Le matériel consiste essentiellement en un jeu de courriers à dé-identifier associé aux noms des patients correspondants. Ce dernier point est cependant facultatif pour l'élaboration de la méthode. Les données identifiantes à supprimer sont également listées et détaillées.

I. Données initiales

A. Base de données

Ce travail de dé-identification a été mené dans le cadre du projet européen PSIP [PSIP 2011] pour *Patient Safety through Intelligent Procedures in medication*. Ce projet a pour vocation de prévenir les effets indésirables des médicaments dans les services cliniques des hôpitaux [Chazard 2011]. Un programme informatique analyse le contenu de séjours hospitaliers d'un hôpital périphérique du CHU de Lille et génère des règles de survenue d'effets indésirables médicamenteux. Ces règles sont des associations statistiques entre des événements survenus lors d'une hospitalisation. Une revue de ces règles par des experts est donc nécessaire afin de les valider. Pour réaliser cette revue de cas, une base de données contenant environ 50 000 documents médicaux informatisés correspondant aux séjours analysés est mise à disposition des experts. Ces documents proviennent des dossiers médicaux et consistent essentiellement en : courriers de sortie, courriers de transfert, comptes-rendus opératoires, comptes-rendus d'examens d'imagerie, fiches de liaison.
Ces documents sont fournis sous format Word dans un fichier compressé (.zip). A chaque document est attribué un numéro arbitraire de séjour permettant de rapporter les documents correspondant au même séjour. Ces numéros arbitraires de séjour sont répertoriés dans une table.
La base de données est composée des documents en texte libre informatisés des séjours d'un hôpital périphérique des années 2009, 2010 et 2011. Ces documents proviennent de divers services de médecine : services de cardiologie, de gériatrie, de pneumologie, de gastro-entérologie.
La Figure 1 montre un exemple fictif de courrier de sortie d'hospitalisation. Les données personnelles identifiantes sont principalement contenues dans l'en-tête et en fin de document, mais cependant le texte n'est pas structuré car ces informations peuvent être retrouvées dans le corps du texte. Par exemple, le nom de la patiente est repris en fin de paragraphe résumant les traitements qu'elle a reçus durant son séjour. L'exemple cité montre également un cas très fréquemment retrouvé dans les

documents médicaux en texte libre : il s'agit du nom de famille mal orthographié. En effet, alors que le nom est « Dupont » dans la phrase introductive du courrier, il est noté « Dupond » dans le texte. La Figure 1 attire également l'attention sur les différences de format de date qui peuvent être retrouvés dans les courriers.

> *Exemples de format de date retrouvés dans les courriers :*
> - *16.01.2000*
> - *08/02/74*
> - *13 au 16 janvier 2000*

Cet exemple de courrier met en évidence les problèmes posés par l'orthographe des noms propres. L'exemple de ce courrier permet également de citer le contenu d'un courrier de sortie d'hospitalisation de manière générale. Il apparaît notamment :

- des diagnostics (« diabète, asthme »)
- des actes (« bilan gazométrique »)
- des résultats de biologie (« hémoglobine à 15,3 »)
- des noms de médicaments ainsi que leur posologie (« Lévothyrox 125 : 1 par jour »)

16.01.2000

Dr Roger Pierre
3 rue de la gare
99280 Ville-sur-mer

Dr Caroline Jean
Chirurgie endocrinologie
Hôpital Ambroise Paré
99280 Ville-sur-mer

er/nm
nda : 895319413

Chers confrères,

Votre patiente madame Dupont Cunégonde, née le 08/02/74, demeurant 179 avenue Jules Verne, 99280 Ville-sur-mer, a été hospitalisée dans le service de pneumologie du 13 au 16 janvier 2000 pour un nouvel épisode de décompensation respiratoire.

Ses antécédents sont marqués par un asthme, du diabète, un goitre thyroïdien opéré récemment.
La patiente nous a été adressée en raison d'une décompensation respiratoire avec dyspnée et toux sans expectoration.
A l'examen clinique à l'entrée on retrouve à l'auscultation pulmonaire des ronchi aux deux bases. L'abdomen était souple, l'auscultation cardiaque était sans particularité.
Le bilan biologique à l'entrée montre une hémoglobine à 15.3, des leucocytes à 5 310, une CRP élevée à 60 témoignant d'un probable syndrome inflammatoire. L'examen cytobactériologique des crachats s'est révélé négatif.

La radiographie pulmonaire ne retrouve pas de foyer visible de pneumopathie.
Le bilan gazométrique initial montre une pO_2 à 61, une pCO_2 à 34 et un pH à 7.43.
La patiente a bénéficié dans notre service d'un traitement à visée respiratoire associant des aérosols de broncho-dilatateurs, une kinésithérapie active, une corticothérapie par voie générale et une antibiothérapie par Augmentin permettant une amélioration rapide.
Madame Dupond souhaite regagner rapidement son domicile.

Au total : surinfection bronchique chez une patiente porteuse d'un asthme.

Le traitement de sortie est le suivant :
- Lévothyrox 125 : 1 par jour,
- Solupred 20 : 2 cps par jour pendant 5 jours puis arrêt,
- Augmentin : 1 g x 3 par jour pendant 8 jours.
- Sérétide 500 : 1 bouffée x 2 par jour,
- Ventoline a la demande si besoin 4 bouffées par jour,
- Azantac 150 : 1 cp par jour.

Madame Dupont sera revue par le Docteur Dubois d'ici 15 jours.
Bien cordialement.

Dr Joyeux Dr Lefebvre Dr Martin

Figure 1. Exemple *fictif* de courrier de sortie d'hospitalisation

B. Système d'Information Hospitalier (SIH)

Grâce au Système d'Information Hospitalier (SIH), les noms et prénoms des patients dont nous possédons les courriers sont disponibles, sous forme d'un tableau reliant les noms et prénoms au numéro arbitraire de séjour. Le tableau fourni par le SIH comporte 3 876 lignes, un exemple fictif de présentation de ce tableau est donné dans le Tableau 3.

Identifiant Séjour	Prénom	Nom	Nom Usuel
159786452	Gustave	Flaubert	
945678584	Élisabeth	Bennett	Darcy
324548677	Catherine	Tully	

Tableau 3. Exemple *fictif* de présentation de tableau fourni par le SIH

II. Données identifiantes à supprimer

En France, la CNIL souligne l'importance du respect de la confidentialité des patients dans le milieu médical [CNIL 2011 (2)] et met en avant la nature des données à protéger, ces données étant dites « sensibles » dès lors qu'elles sont susceptibles de révéler l'identité d'un patient. Il peut s'agir des noms, prénoms des patients, de la date de naissance, d'un matricule de Sécurité Sociale, etc. Toutefois la liste fournie n'est pas exhaustive et non officielle. De plus, la CNIL préconise une anonymisation par hachage [CNIL 2011 (3)], qui ne nous concerne pas puisque la dé-identification que nous recherchons est irréversible, dans le sens où aucun moyen ne permet de retrouver les données identifiantes originales (qui sont perdues) et aucune notion de suivi n'est nécessaire.
Aux Etats-Unis, l'*Health Insurance Portability and Accountability Act* (HIPAA) donne des recommandations pour protéger la vie privée des patients. Ces recommandations contiennent une liste d'items qui doivent être retirés de tout document nécessitant d'être « dé-identifié ». Ces données sont nommées *Protected*

Health Information (PHI) et regroupées en 18 catégories décrites dans le Tableau 4 (voir Tableau 4).

Catégories de données identifiantes « PHI »
1. Noms
2. Indications géographiques
3. Adresses email
4. Numéro de Sécurité Sociale
5. Numéros de licence ou de certificat
6. Numéros d'immatriculation d'un véhicule
7. Identifiants biométriques
8. Photographies de face
9. Tous les éléments de date
10. Numéros de téléphone
11. Numéros de fax
12. Numéros de plan de santé
13. Numéros de compte
14. URL
15. Adresses IP
16. Numéros de dossiers
17. Numéros d'identification ou de série
18. Tout autre numéro ou identifiant ou code
19. + Noms de soignants ou de structure de soins
20. + Adresses de structures de soins
NB : le signe + montre qu'il s'agit d'items ajoutés usuellement à la liste fournie par l'HIPAA

Tableau 4. Catégories de PHI

Après une revue de la littérature concernant la dé-identification des documents médicaux, il apparaît que les noms de soignants ainsi que les noms et adresses de structures de soins sont considérés comme identifiants chez un certain nombre d'auteurs [Beckwith 2006, Friedlin 2008, Neamatullah 2008, Ruch 2000, Szarvas 2007, Tu 2010, Uzuner 2008, Wellner 2007]. Nous avons donc décidé de retirer ces données également, en plus des PHI décrits par l'HIPAA.

Méthode

La méthode est définie en trois étapes : tout d'abord élaborer FASDIM, évaluer ensuite d'une part la performance de la méthode et d'autre part quantifier la perte d'information engendrée par le traitement des documents, et enfin évaluer la charge de travail et le temps requis afin de mettre à jour FASDIM.

I. L'algorithme FASDIM

Le principe de la méthode repose sur la construction d'une liste de mots autorisés à apparaître dans les courriers dé-identifiés. Nous considérons que tous les autres mots, y compris les chiffres, doivent disparaître. La méthode a comme point de départ des courriers médicaux contenant les données personnelles, ces courriers subissant une série de simplifications dont l'objectif est de limiter le nombre total de mots différents contenus dans les courriers. En effet, limiter le nombre total de mots différents permet de simplifier la création de la liste de mots autorisés réalisée manuellement à partir des différents mots retrouvés dans les courriers. En parallèle, une méthode d'identification des motifs comprenant des données numériques à conserver est appliquée. Cette méthode traite chaque courrier indépendamment, aucun lien n'est établi entre les courriers d'un même patient, ni d'un même séjour. L'ensemble de ces étapes est schématisé sur la Figure 2.
Tout au long de cette section, nous suivrons ce schéma en mettant en gras la partie faisant l'objet d'une description détaillée.

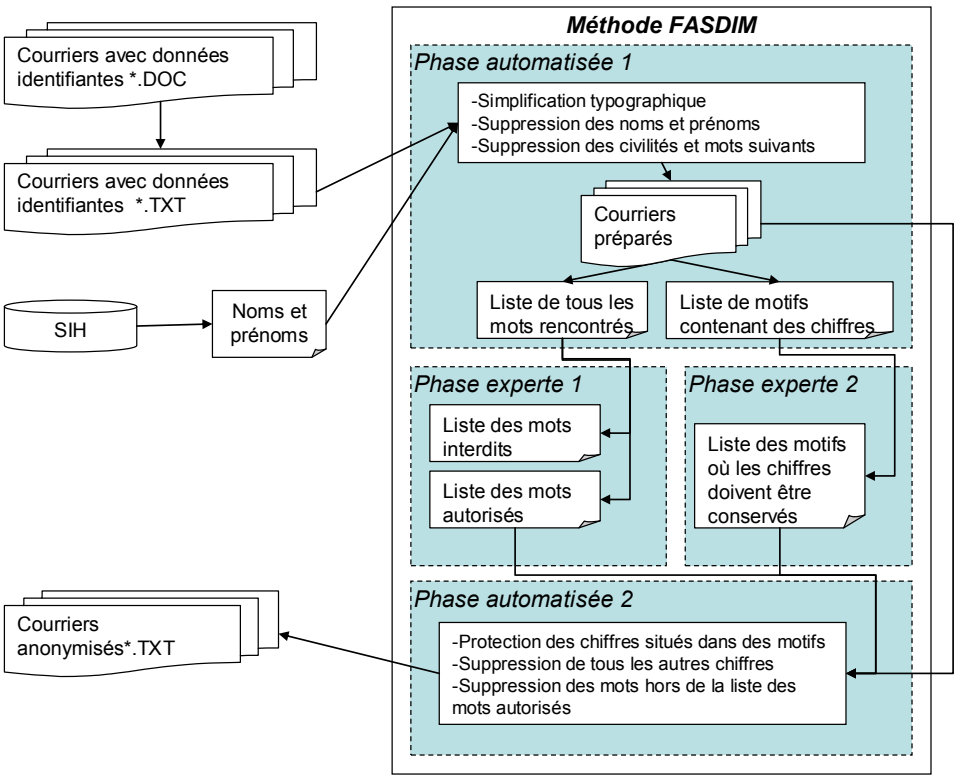

Figure 2. Étapes de FASDIM

A. Transformation des courriers

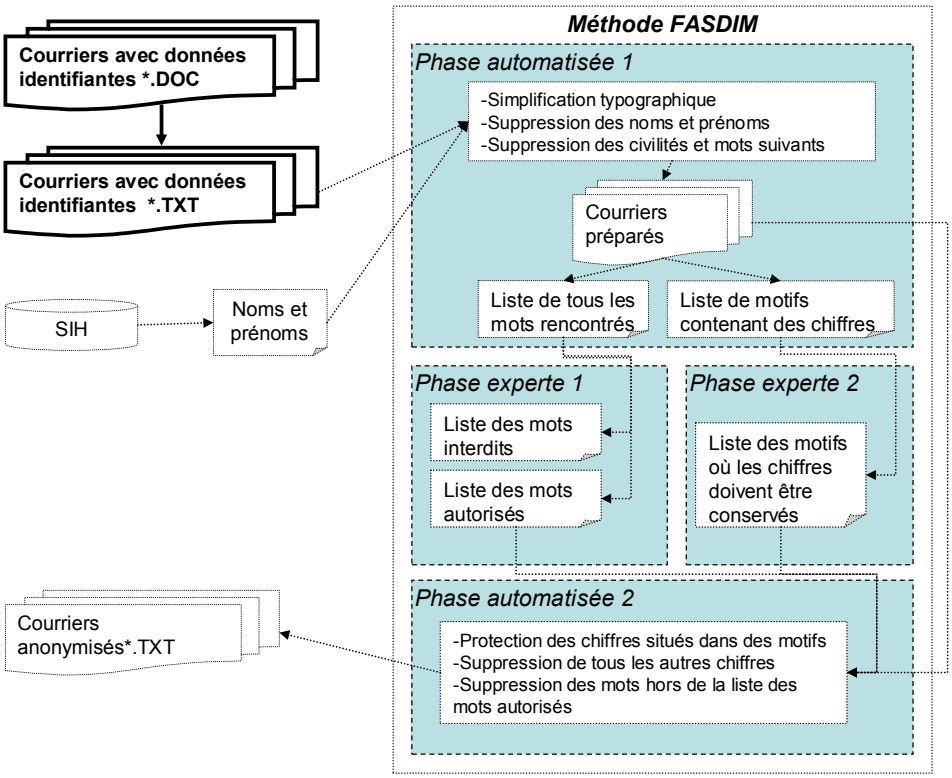

Figure 3. Étape de transformation des courriers

La toute première étape consiste à changer le format des courriers médicaux ; ces courriers sont fournis initialement au format fichier binaire Microsoft Word « .doc » dans une archive compressée. Un programme, écrit en PHP, permet de décompresser le fichier puis de convertir les courriers contenus dans ce fichier au format « .txt », en invoquant la librairie COM pour PHP de Microsoft®. Un fichier TXT par courrier est créé et rangé dans un nouveau dossier.
La transformation en format « .txt » permet de conserver le texte brut, les fichiers au format .doc sont beaucoup plus difficiles à traiter en PHP puisqu'ils contiennent à la fois le texte et le code de mise en page associé.

B. Simplification typographique

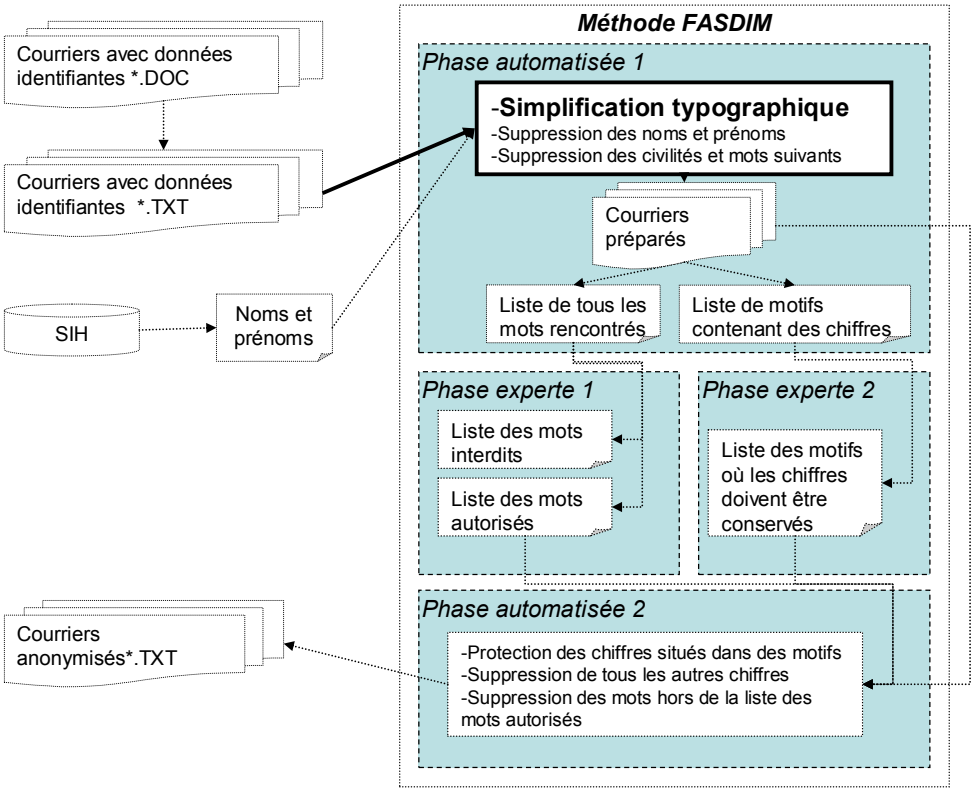

Figure 4. Étape de simplification typographique

Les documents médicaux d'origine sont en format texte brut et subissent une première modification : tous les caractères sont mis en minuscules ; les accents, les trémas et les caractères spéciaux sont supprimés et remplacés par les caractères simples associés. Ainsi :

> *Exemples*
> - *« événement » est remplacé par « evenement »*
> - *« cœur » est remplacé par « coeur »*
> - *« façon » est remplacé par « facon »*

La liste des caractères spéciaux associés à leur transformation est donnée dans le Tableau 5 (voir Tableau 5).

Caractères initiaux	Caractères de remplacement
à, â, ä	a
é, è, ê, ë	e
î, ï	i
ô, ö	o
ù, û, ü	u
ÿ	y
œ	oe
æ	ae
ç	c

Tableau 5. Caractères spéciaux et leur modification

Cette première étape a pour objectif de supprimer les fautes et polymorphismes dus à l'accentuation, et de faciliter la dernière étape (qui est la création d'une liste de mots autorisés) en diminuant le nombre total de mots différents présents dans les courriers. Cette étape facilite la résolution des problèmes posés par l'encodage des caractères.

C. Suppression des noms et prénoms des patients à partir de la base de données du SIH

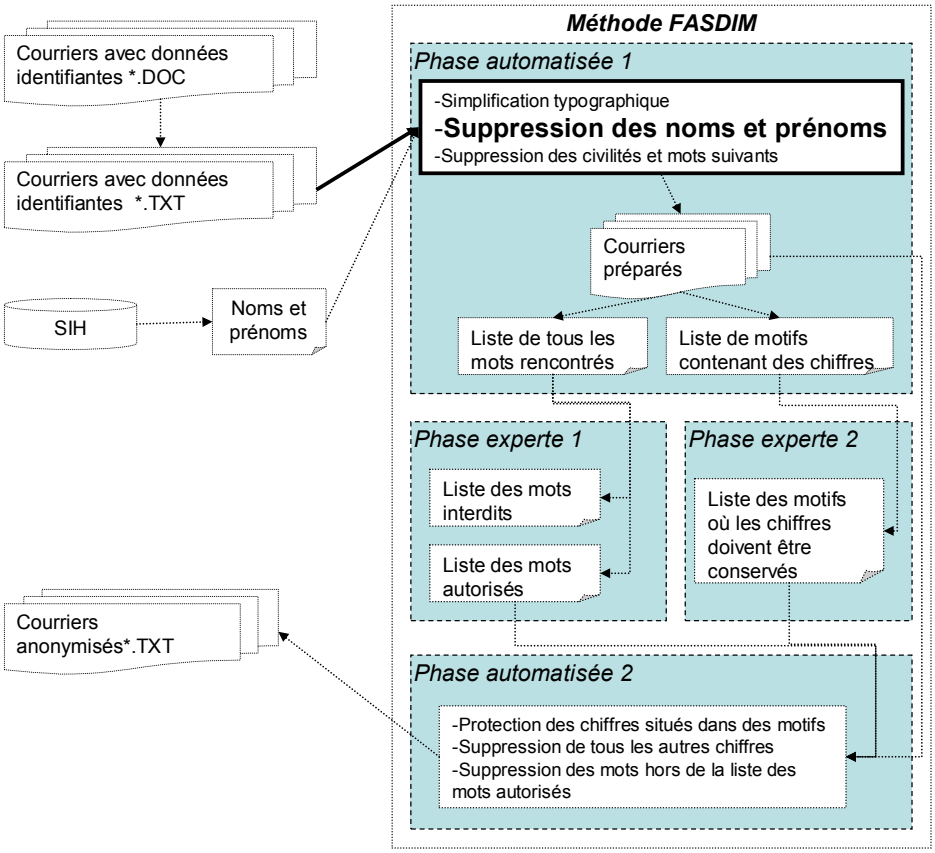

Figure 5. Étape de suppression des noms et prénoms

A partir de la table de correspondance comprenant les noms et prénoms des patients reliés à leur numéro de séjour arbitraire, ces données sont repérées dans le document puis remplacées par une arobase. Cette étape est facultative et dépend des données disponibles. Les noms et prénoms fournis subissent la même simplification typographique que les documents devant être dé-identifiés, pour les mêmes raisons (limiter les polymorphismes dus aux accents essentiellement). Les prénoms composés voient les deux prénoms traités séparément.

> *Exemple : Jean-François est interprété en 2 temps : le programme supprime d'abord tous les « jean » puis tous les « francois » présents dans le courrier.*

D. Suppression des motifs incluant une civilité

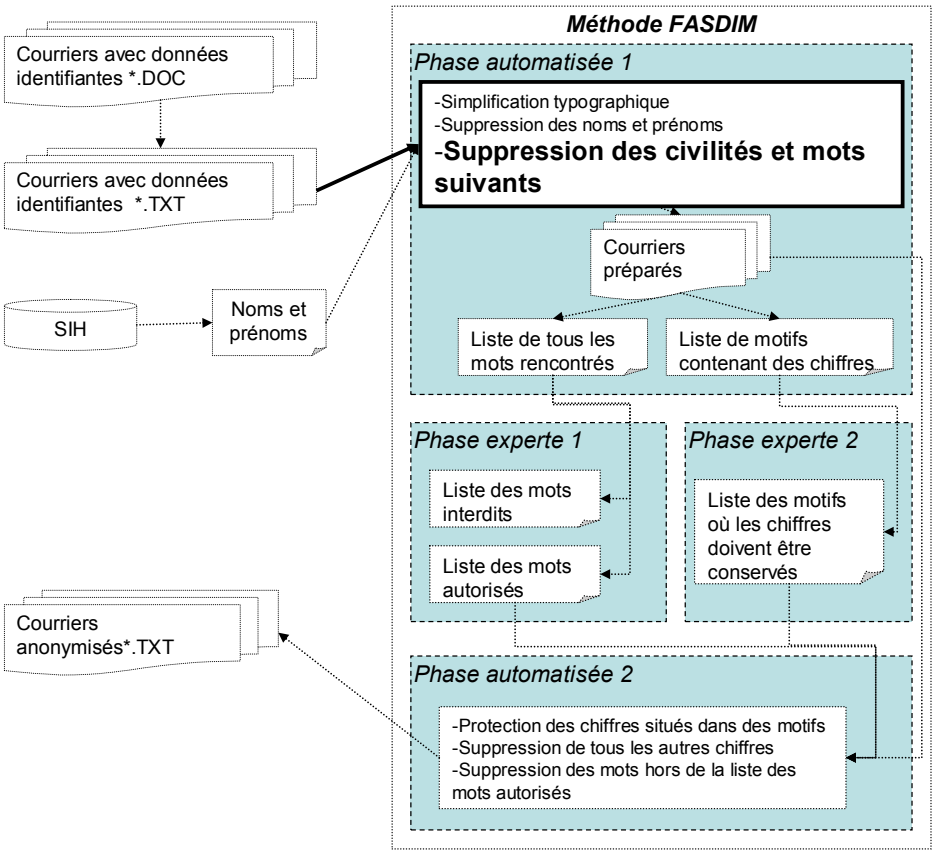

Figure 6. Étape de suppression des titres et civilités

Des motifs reprenant des civilités et des titres ont été définis : il s'agit des civilités comme « Monsieur », « M. », « Mr. », « Madame » et des titres comme « docteur » ou « Dr. » puis du premier ou des deux premiers mots suivants selon le cas de figure (présence ou non d'un point ou d'une virgule). 48 motifs ont ainsi été pris en compte et sont répertoriés dans le Tableau 6. Ces motifs, une fois repérés dans le document, sont remplacés par des arobases.

Objectif : supprimer civilité ou titre suivi du nom et prénom
"monsieur"[espace][mot][espace][mot]
"m."[espace][mot][espace][mot]
"mr."[espace][mot][espace][mot]
"mr"[espace][mot][espace][mot]
"madame"[espace][mot][espace][mot]
"mme."[espace][mot][espace][mot]
"mme"[espace][mot][espace][mot]
"mademoiselle"[espace][mot][espace][mot]
"melle."[espace][mot][espace][mot]
"melle"[espace][mot][espace][mot]
"docteur"[espace][mot][espace][mot]
"dr. "[espace][mot][espace][mot]
"dr"[espace][mot][espace][mot]
"professeur"[espace][mot][espace][mot]
"pr."[espace][mot][espace][mot]
"pr"[espace][mot][espace][mot]
Objectif : supprimer civilité ou titre suivi de l'initiale du prénom, d'un point et du nom
"monsieur"[espace][lettre] "." [+/-espace][mot]
"m."[espace][lettre] "." [+/-espace][mot]
"mr."[espace][lettre] "." [+/-espace][mot]
"mr"[espace][lettre] "." [+/-espace][mot]
"madame"[espace][lettre] "." [+/-espace][mot]
"mme."[espace][lettre] "." [+/-espace][mot]
"mme"[espace][lettre] "." [+/-espace][mot]
"mademoiselle"[espace][lettre] ". " [+/-espace][mot]
"melle."[espace][lettre] "." [+/-espace][mot]
"melle"[espace][lettre] "." [+/-espace][mot]
"docteur"[espace][lettre] "." [+/-espace][mot]
"dr. "[espace][lettre] "." [+/-espace][mot]
"dr"[espace][lettre] "." [+/-espace][mot]
"professeur"[espace][lettre] "." [+/-espace][mot]
"pr."[espace][lettre] "." [+/-espace][mot]
"pr"[espace][lettre] "." [+/-espace][mot]
Objectif : supprimer civilité ou titre suivi uniquement d'un nom
"monsieur"[espace][mot]
"m."[espace][mot]
"mr. "[espace][mot]
"mr"[espace][mot]
"madame"[espace][mot]
"mme."[espace][mot]
"mme"[espace][mot]
"mademoiselle"[espace][mot]
"melle."[espace][mot]
"melle"[espace][mot]
"docteur"[espace][mot]
"dr."[espace][mot]
"dr"[espace][mot]
"professeur"[espace][mot]
"pr."[espace][mot]
"pr"[espace][mot]

Tableau 6. Liste des motifs comprenant une civilité ou un titre

Le choix d'enlever deux mots suivant un titre ou une civilité est un moyen d'enlever le nom et le prénom quand ce dernier figure. Cependant le prénom n'étant pas repris systématiquement dans le texte, le risque est d'avoir une suppression inopportune d'un mot qui aurait dû être laissé en place. Néanmoins, dans la très grande majorité

des cas, ce mot enlevé est un verbe ou un auxiliaire et son absence n'affecte ni la lisibilité ni la compréhension du texte.

E. Création d'une liste de mots autorisés

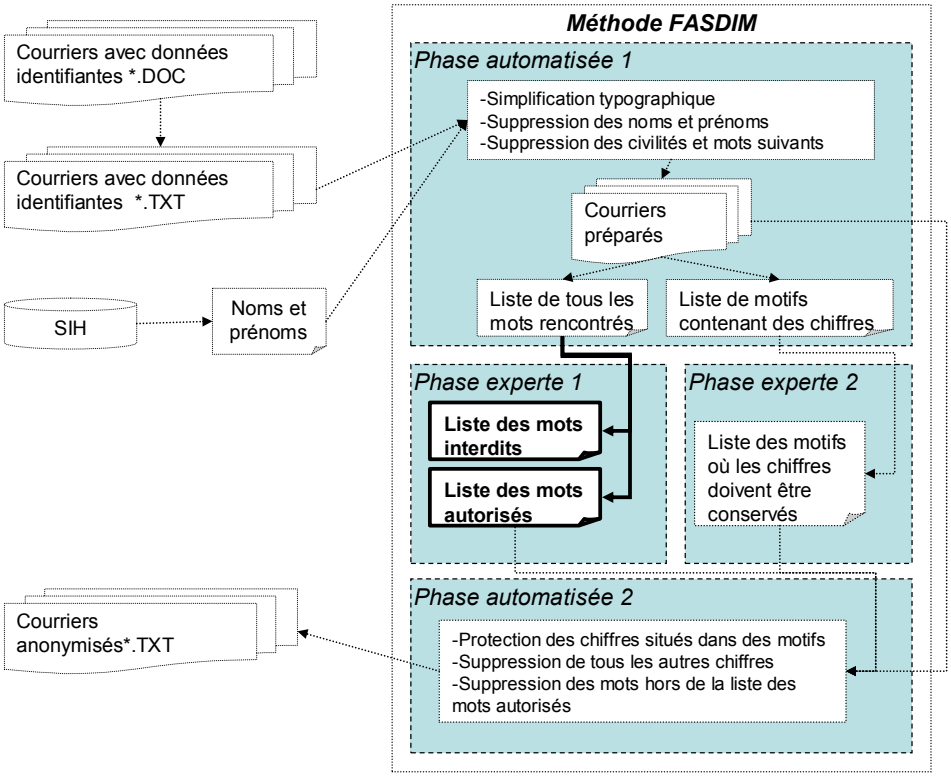

Figure 7. Étape de création d'une liste de mots autorisés

La création d'une liste de mots autorisés résulte de la nécessité de supprimer les portions d'adresse autres que numériques ainsi que les noms de soignants et les noms de patients mal orthographiés. Ainsi, tous les mots différents, y compris les différentes conjugaisons d'un même verbe, les accords de noms communs ou d'adjectifs, sont extraits avec leur nombre d'occurrences. La liste ainsi obtenue est ensuite revue par des experts afin de constituer les deux listes suivantes : une première liste de « mots autorisés » et une seconde de « mots interdits ». La liste de mots interdits comprend tous les mots qui sont :

- soit des mots communs très probablement utilisés dans des adresses

> *Exemples :*
> - *tilleuls*
> - *acacias*

- soit des noms propres qui ne sont pas des éponymes médicaux

Cependant, lorsqu'un éponyme médical est un nom propre attribué fréquemment, il est porté sur la liste des mots interdits, ceci afin de garantir la confidentialité du patient. Dans tous les cas douteux, c'est la confidentialité qui est privilégiée : on supprimera donc parfois un mot à tort dans un certain contexte, alors que dans un autre contexte la confidentialité du patient sera protégée.

> *Exemples :*
> - *« mordu par un renard »*
> - *« Dr Renard »*
>
> - *« nous avons pris la liberté de… »*
> - *« 35 Bd de la liberté »*

Les mots pertinents du point de vue médical mais mal orthographiés sont conservés dans la liste de mots autorisés.

> *Exemple :*
> *« férosémide » à la place de « furosémide »*

Cette approche sera menée avec prudence dans la mesure où certains mots peuvent être interprétés comme étant des mots mal orthographiés alors qu'ils sont en réalité des noms de ville ou de personne.

> *Exemple :*
> *« hergnie » (nom de ville) considéré comme « hernie » mal orthographié*

F. Création d'une liste de motifs autorisant les chiffres

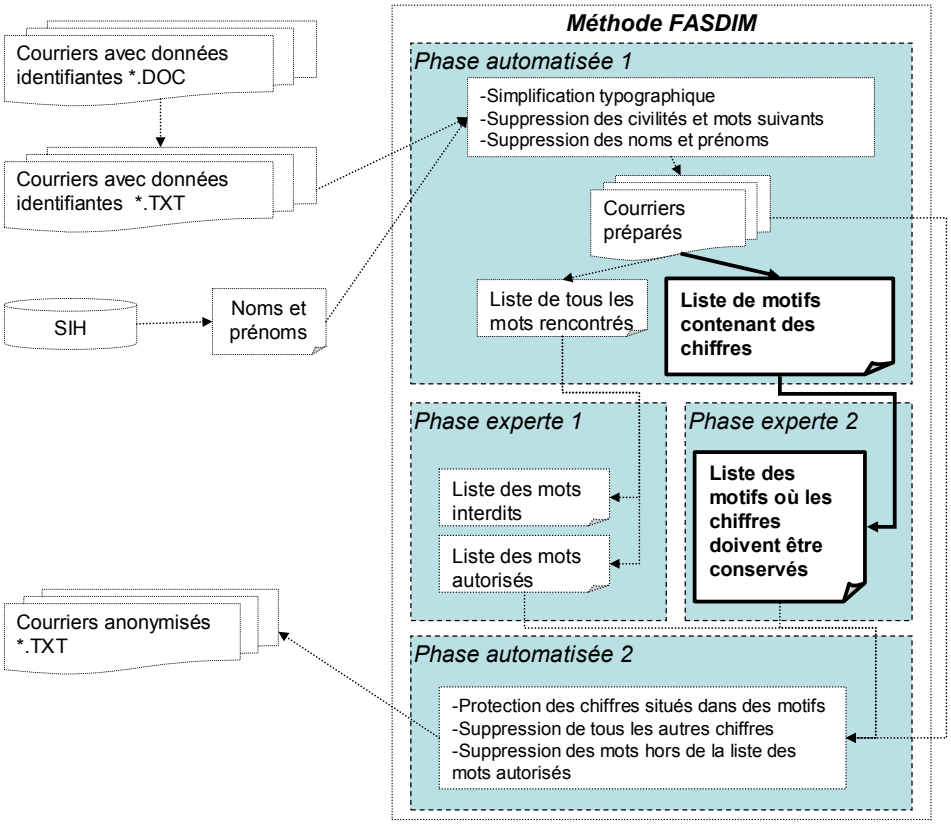

Figure 8. Étape de création d'une liste de motifs contenant des chiffres

Certaines données numériques, comme les dates de naissance, les dates de séjour, les indices biométriques et les portions d'adresse (numéro de rue, codes postaux…) nécessitent d'être enlevées, alors que d'autres données numériques ont besoin d'être conservées car elles constituent une information médicale importante. Il s'agit essentiellement de :

- données de biologie
- posologies médicamenteuses
- mesures physiques

> *Exemples :*
> - *« glycémie à 2,5 g/l »*
> - *« kardégic 160 mg »*
> - *« tension artérielle à 16/10 »*

Nous avons identifié à l'aide d'expressions rationnelles les motifs comprenant des chiffres et les deux mots les suivant ou les précédant. Les expressions rationnelles sont des chaînes de caractères aussi appelées motifs, ces motifs étant fréquemment retrouvés dans l'ensemble étudié. Des phrases contenant ces motifs ont ensuite été revues manuellement par des experts afin de déterminer si elles pouvaient correspondre d'une manière fiable à une donnée médicale, non identifiante.

1. La préposition « à »

Nous avons constaté après vérification, sur plus de 300 exemples qu'après la préposition « a » (l'accent étant supprimé à la suite de l'étape de simplification typographique, voir Méthode, I.B, page 33, aucune discrimination n'est possible entre la préposition et l'auxiliaire), les chiffres ne concernaient pas des données identifiantes. Nous avons donc choisi de garder tous les chiffres se situant après « a » dans le texte : ceci permet de garder dans le texte remanié beaucoup de données biologiques, comme toutes les formulations qui ne sont pas suivies d'unités de mesure et donc par conséquent, non repérées par une expression rationnelle.

> *Exemples :*
> - *« la créatinine est à 10 »*
> - *« urée à 0,5 »*

2. Les unités de mesure

Les unités de mesures présentes dans les courriers ont été listées à l'issue de l'étude des mots suivant des chiffres.

> *Exemples :*
> - *« g » pour grammes*
> - *« l » pour litres*
> - *« ui » pour unités internationales*
> - *« torr »*

3. Les formes galéniques

Les différentes formes galéniques ainsi que les abréviations correspondantes retrouvées dans les courriers sont listées.

> *Exemple :*
> *« cp », « cps » ou « comprimes »*

4. Les noms de médicaments

Environ 400 noms de médicaments fréquemment suivis de chiffres ont été répertoriés.

> *Exemples :*
> - *« acebutolol »*
> - *« renitec »*

5. Autres catégories

Certains termes médicaux sont également apparus dans les motifs.

Exemples :
- *« bav »*
- *« diamètre »*
- *« aiguille »*

Certains termes comme « aiguille » ont une pertinence médicale discutable, néanmoins toutes les données non personnelles ont besoin d'être conservées, afin de conserver la lisibilité du texte dé-identifié.

G. Suppression des chiffres non protégés

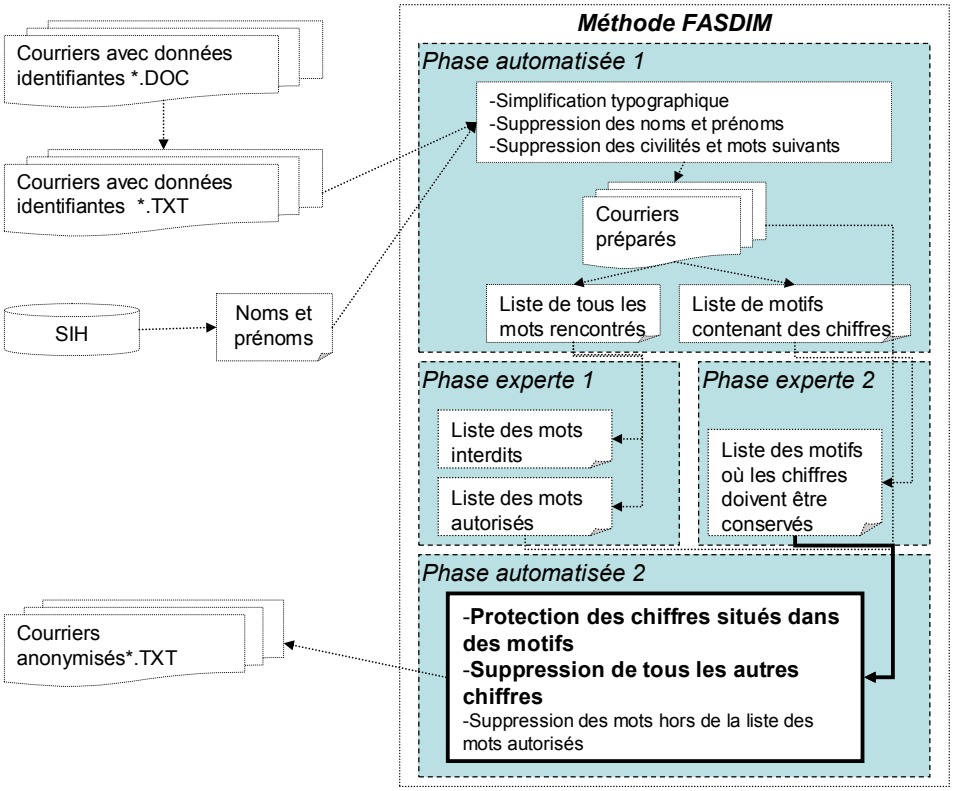

Figure 9. Étape de suppression des chiffres non protégés

Les motifs comprenant des chiffres que nous souhaitons préserver sont protégés, tandis que tous les autres chiffres présents dans le courrier sont supprimés.
Un exemple de cette étape de protection des chiffres contenus dans certains motifs définis ci-dessus, puis de la suppression de tous les autres chiffres contenus dans le courrier est donné Figure 10.

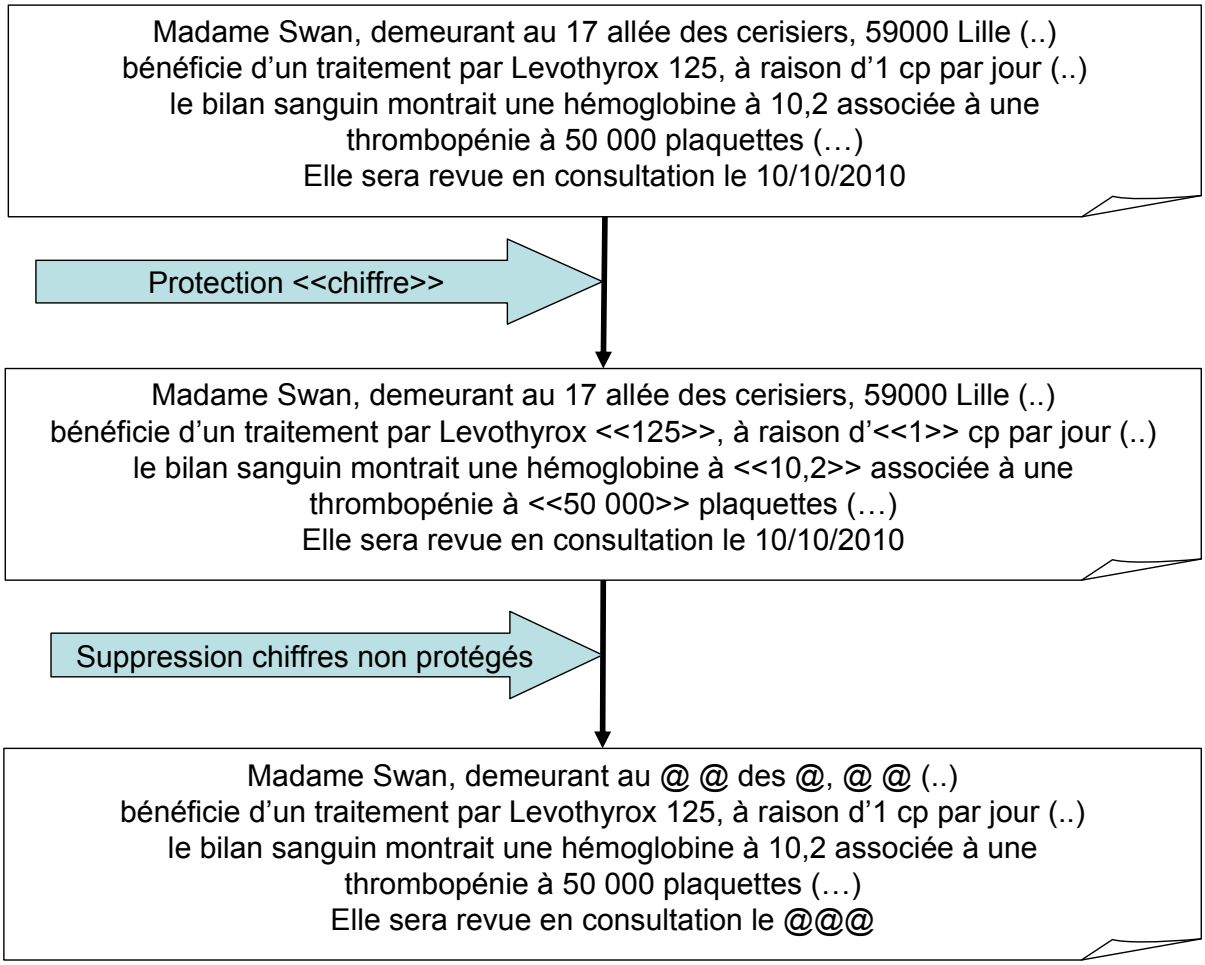

Figure 10. Exemple *fictif* de protection puis suppression de chiffres dans les courriers

H. Suppression des mots non autorisés

Enfin, la dernière étape consiste à supprimer des courriers prétraités tous les mots qui n'appartiennent pas à la liste de mots autorisés établie précédemment (voir Figure 11). Ces mots sont remplacés par des arobases, puis les courriers ainsi dé-identifiés sont enregistrés dans un dossier dédié.

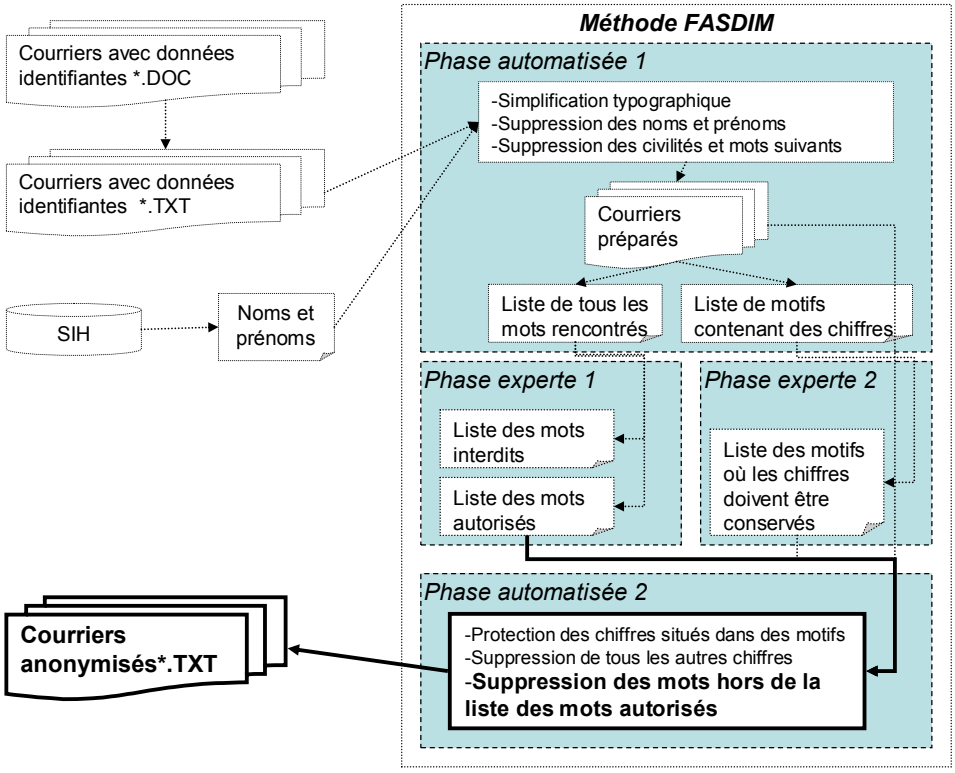

Figure 11. Étape de suppression des mots n'appartenant pas à la liste de mots autorisés

I. Les itérations suivantes de la liste de mots autorisés

Les deux listes de mots (autorisés et interdits) peuvent être mises à jour à chaque fois qu'un set de nouveaux courriers a besoin d'être dé-identifié, c'est pourquoi nous conservons également la liste de mots interdits : ainsi nous extrayons des nouveaux courriers les « nouveaux mots », c'est-à-dire les mots qui ne sont jamais apparus auparavant et qui par conséquent n'ont jamais été placés dans une des deux listes. Cette mise à jour est optionnelle du point de vue de la confidentialité du patient : en effet, elle permet certes une meilleure lisibilité et diminue la perte d'information, mais on pourrait très bien envisager de ne pas la réaliser. Les nouvelles données identifiantes susceptibles d'être contenues dans les nouveaux courriers ne seraient pas autorisées et donc n'apparaîtraient pas sur les courriers dé-identifiés : les mots nouveaux non listés sont remplacés par des arobases. Il apparaît qu'au terme d'un certain nombre de courriers traités, le nombre de nouveaux mots diminue et que la proportion de mots mal orthographiés augmente. De plus, la plupart du temps, une faute d'orthographe n'apparaissant qu'après le traitement de 20 000 courriers ne concerne qu'un seul courrier. La rentabilité de la revue des listes est donc faible.

II. Évaluation de la méthode

A. Évaluation de l'efficacité

1. Matériel d'évaluation

Un corpus de 508 courriers de sortie est constitué par tirage au sort parmi 17 776 courriers disponibles afin d'évaluer l'efficacité de FASDIM. L'évaluation porte sur les courriers de sortie car ce sont les documents qui contiennent le plus d'informations, que ce soit des informations personnelles ou médicales, comparées aux comptes-rendus d'imagerie ou opératoires.

2. Les outils de mesure de l'efficacité de FASDIM

Deux experts ont annoté les données identifiantes sur les courriers originaux (non traités par le programme) puis ont comparé ces données avec les courriers dé-identifiés par FASDIM. Ont ainsi été dénombrés :

- Le nombre de Vrais Positifs (VP) : le nombre de PHI enlevés par FASDIM
- Le nombre de Faux Négatifs (FN) : le nombre de PHI laissés en place par FASDIM alors qu'ils auraient dû être supprimés
- Le nombre de Faux Positifs (FP) : le nombre de mots supprimés par FASDIM alors qu'ils auraient dû rester en place dans le courrier dé-identifié.

Une fois que tous les courriers sont revus, les statistiques suivantes sont calculées :

- Le rappel, qui représente la proportion de PHI retrouvés et supprimés par FASDIM dans les courriers, parmi tous les PHI présents dans les courriers. Son mode de calcul est présenté dans l'Formule 4. On parle ici de rappel plutôt que de sensibilité, l'évaluation portant sur un programme informatique [Wikipedia 2011 (2)]
- La précision, qui représente la proportion de mots enlevés de manière appropriée parmi tous les mots supprimés par FASDIM. Son mode de calcul est présenté dans l'Formule 5. Le terme statistique correspondant est la Valeur Prédictive Positive (VPP)
- La F-measure, qui représente la moyenne harmonique entre le rappel et la précision et permet d'évaluer la qualité globale du procédé. Son mode de calcul est présenté dans l'Formule 6

$$Rappel = R = \frac{VP}{\#identifiants} = \frac{VP}{VP + FN}$$

Formule 4. Rappel

$$Précision = P = \frac{VP}{\#mots\ supprimés} = \frac{VP}{VP + FP}$$

Formule 5. Précision

$$F\_measure = F = \frac{1}{\frac{1}{2} * (\frac{1}{P} + \frac{1}{R})}$$

Formule 6. F-measure

Ces mesures sont couramment utilisées et admises dans le domaine de la dé-identification [Aberdeen 2010, Dalianis 2010, Hirpcsak 2005, Szarvas 2007, Tu 2010, Uzuner 2008, Velupillai 2009, Wellner 2007, Yeniterzi 2010] et d'une manière générale dans le domaine du *Semantic Mining*.

3. L'annotation et le classement par catégories des données identifiantes

L'évaluation doit prendre en compte un phénomène non négligeable dans le domaine de la dé-identification : les données identifiantes n'ont pas la même valeur. En effet, nous pouvons considérer qu'il est plus préjudiciable pour le patient que son nom de famille soit resté sur le document final plutôt que le nom d'une ville ou d'un médecin. C'est pourquoi il faut pouvoir détailler les faux négatifs pour montrer l'impact du manque d'efficacité. Nous avons donc regroupé en 7 catégories plus génériques les 18 items désignés par l'HIPAA ainsi que les catégories supplémentaires concernant l'identification de soignants et de structures de soins. Ces catégories sont présentées dans le Tableau 7.

Catégories de PHI retenues pour l'évaluation	Items de l'HIPAA correspondants
Noms et prénoms de patients	noms
Dates	dates
Indices géographiques	adresse postale adresse email url
Indices biométriques	taille poids
Données numériques identifiantes	numéro de Sécurité Sociale numéro de dossier numéro de téléphone numéro de Fax numéro de série de matériel prothétique ou implantable numéro de licence ou de certificat numéro d'immatriculation d'un véhicule numéro de compte numéro de plan santé adresses IP tout autre numéro identifiant ou code
Noms de soignants	-
Noms de structure de soins ou adresses de médecin	-

Tableau 7. Catégories de PHI évaluées et leur correspondance aux items de l'HIPAA

Les experts disposent d'un tableau Excel, avec une ligne par courrier (numéroté), puis une première colonne pour y consigner le nombre de PHI présents dans le courrier original, suivie d'une deuxième colonne recevant le nombre de faux positifs, et enfin une colonne par catégorie susmentionnée afin d'y inscrire le nombre de faux négatifs correspondants, retrouvés dans le courrier dé-identifié. Une dernière colonne « commentaire » permet de noter le détail soit des faux négatifs soit des faux positifs, dans l'objectif d'améliorer la méthode après évaluation.
Le dénombrement se fait non pas par mot, mais par concept : ainsi, le nom et le prénom (qu'il s'agisse d'un patient ou d'un médecin) sont comptés comme 1 élément. Pour les adresses, nous avons procédé ainsi : une adresse est globalement divisée en deux parties. La première partie comprend le numéro de voie, le nom de cette voie et éventuellement une précision supplémentaire (comme une lettre de bâtiment ou un nom de résidence) ; la seconde partie comprend le code postal et le nom de la commune. Cette démarche est motivée par le fait que la révélation d'un code postal est équivalente à la révélation du nom de la commune : il est en effet très aisé de retrouver l'un à partir de l'autre et réciproquement. Concernant la première partie d'adresse, ce principe ne se vérifie pas mais s'en rapprochait suffisamment pour que nous le considérions du même point de vue global.

> *Exemple :*
> - *adresse initiale « [21 boulevard de la liberté][99999 Ville] »*
>
> *⇨ 2 PHI comptés dans « courrier original »*
> - *s'il demeure « [@ @ de la liberté][@ @] »*
>
> *⇨ 1 PHI restant dans « courrier dé-identifié ».*

Une colonne supplémentaire dans le tableau Excel de recueil des données permet de noter le nombre d'auxiliaires verbaux enlevés par la suppression des titres et

civilités et comptés initialement comme des faux positifs. En effet, nous souhaitons quantifier dans quelle mesure ces faux positifs, qui altèrent peu la lisibilité, pénalisent la précision.

B. Évaluation de la perte d'information

Cette évaluation est réalisée sur le même corpus de courriers. L'objectif est de quantifier la perte de l'information médicale contenue dans ces documents. Les courriers précédemment évalués sont donc repris, les experts relèvent le nombre de codes selon la CIM-10, le nombre de médicaments et le nombre d'actes selon la CCAM présents d'abord dans les courriers originaux puis dans les courriers dé-identifiés. Les codes de la CIM-10 sont divisés en 3 catégories : biologie, actes et autres (pour diagnostics essentiellement). Pour cela, les experts disposent d'un tableau Excel, avec une ligne par courrier, des colonnes reprenant les catégories :

- médicaments (ex : paracétamol)
- biologie CIM-10 (ex : R739, hyperglycémie ; R71, anomalie des globules rouges)
- actes CIM-10 (ex : Z491, dialyse extra-corporelle)
- autres CIM-10, c'est-à-dire l'ensemble des autres codes contenus dans la CIM-10 (ex : I10, HTA)
- actes CCAM (ex : HHFA001, appendicectomie)
- « commentaires »

Cette dernière colonne a pour but de noter la conséquence de la perte d'un mot : aucun changement de code, changement de code ou disparition du code. Le décompte des médicaments ne comprend que le nom : il ne fallait compter ni la posologie, ni la durée de traitement ni la voie d'administration. Les maladies et symptômes sont comptés par concepts : ainsi « adénocarcinome du sein gauche » compte pour 1, mais « adénocarcinome du sein gauche opéré et compliqué d'un lymphœdème » compte pour 3 (2 pour la catégorie « autres CIM-10 » et 1 pour la catégorie « actes CCAM »). Les antécédents sont également comptés, de même que les redondances dans le texte. Cet exemple de concepts dénombrés dans un extrait de courrier fictif dé-identifié est illustré dans la Figure 12. Les concepts sont entourés en pointillés ; le nombre retenu pour le décompte de l'évaluation figure en gras, associé à la catégorie d'information auquel il correspond.

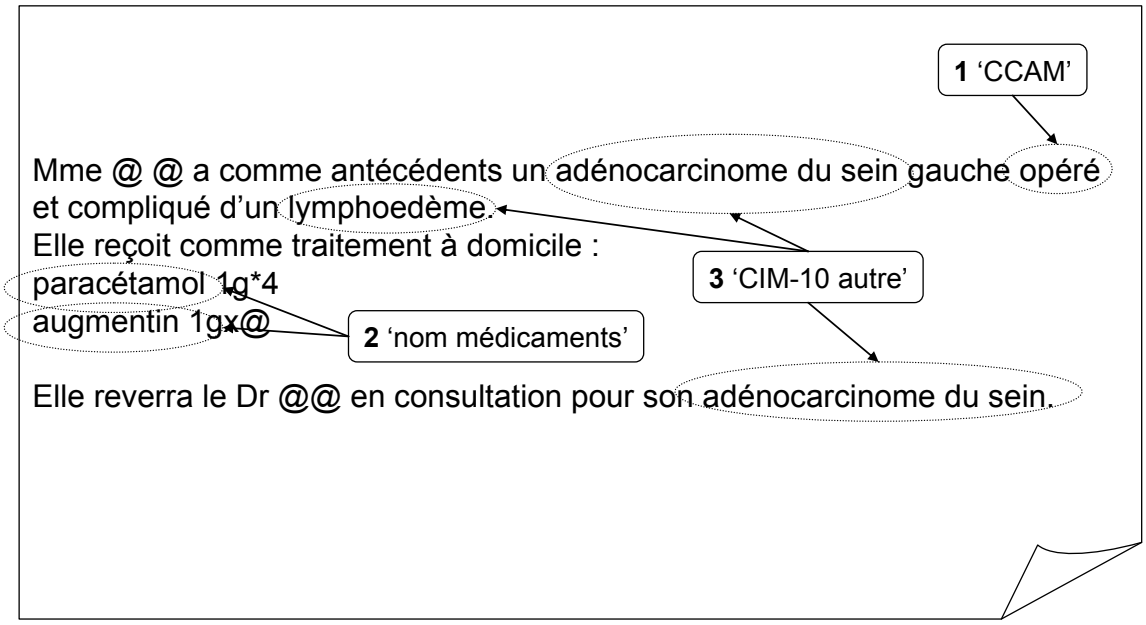

Figure 12. Exemple *fictif* de dénombrement de concepts médicaux dans un extrait de courrier médical dé-identifié

Les experts ont également à disposition le logiciel Medor [Miro 2011] afin de rechercher les codes (CIM-10 et CCAM) existants.
A l'issue de ce recueil, le pourcentage d'information conservée est calculé par catégorie : nombre de codes dans le courrier dé-identifié divisé par le nombre de codes dans le courrier original. Les résultats propres à chaque expert sont également différenciés dans le but de les comparer. En effet, il est reconnu qu'il existe une variabilité entre les experts dans le codage manuel.

C. Évaluation du temps de travail

Lors de chaque mise à jour des listes à réception des nouveaux jeux de courriers, le temps de travail requis pour implémenter la méthode est consigné dans un tableau Excel, de même que le nombre total de nouveaux mots et le nombre de nouveaux mots ajoutés à la liste de mots autorisés.

Résultats

I. Une application qui fonctionne

FASDIM est appliquée en routine dans le cadre du projet PSIP et a permis de dé-identifier 27 540 documents contenus dans les dossiers médicaux. L'utilisation faite par les experts montre que les documents obtenus sont exploitables, tout en garantissant le respect de la vie privée du patient. A titre d'illustration, les Figures 13 et 14 donnent un exemple de courrier de sortie d'hospitalisation non anonymisé puis dé-identifié par FASDIM. Ces exemples sont fictifs (toutes les données identifiantes ont été remplacées par des pseudonymes et des dates différentes) mais réalistes puisqu'ils reprennent la plupart des cas de figure pouvant générer des difficultés pour les méthodes de dé-identification. Ainsi, le nom d'un praticien étant également un nom commun (« docteur coeur » dans l'exemple de courrier) disparaît dans le courrier dé-identifié par FASDIM grâce à la suppression des mots suivant les motifs de titres et de civilités. Cette précaution permet de laisser des mots tels que « cœur » dans la liste de mots autorisés et ainsi de préserver l'information pertinente, autre qu'identifiante.

Le 25.04.2009
Monsieur le Docteur Sarrasin

À

Madame le Docteur Dubois
108 avenue de la République
99280 Ville

sa/nm

Cher confrère,

Votre patient, Monsieur Jean Fontaine, né le 08/07/1929, demeurant 35 rue des tilleuls 99875 Village, a été brièvement hospitalisé dans le service du 25/03/2009 au 22/04/2009, à la suite de la coronarographie.

Ce patient, comme vous le savez, est porteur d'antécédents :
- d'infarctus du myocarde en 2006 pour lequel il a été stenté,
- il est aussi porteur d'une BPCO,
- d'une insuffisance cardiaque,
- et d'un DNID.

À l'entrée dans le service, le malade était stable. À l'examen clinique, on retrouve à l'auscultation pulmonaire quelques sibilants modérés. L'abdomen est souple. Le reste de l'examen clinique est sans particularité.

La coronarographie avait conclu à l'occlusion de l'IVA avec occlusion de la coronaire droite avec un réseau de collatéralités assez développées. L'importance des calcifications et l'état global des artères contre indiquaient l'angioplastie. Suite à cette coronarographie, nous nous sommes orientés vers un traitement médical de cette cardiopathie ischémique. Un traitement par anticoagulants efficaces avec un relais par Previscan a été instauré par le Docteur Martin. Le Plavix a été poursuivi, des bêta bloquants de type Temerit 5 mg ont été introduits.

La radiographie pulmonaire a montré des signes en faveur d'un œdème pulmonaire avec œdème interstitiel bilatéral et un épanchement pleural bilatéral plus marqué à droite.

Malgré le traitement déjà instauré suite à la précédente hospitalisation, on assistait à la persistance de l'hypoxie avec une pO_2 à 54, pCO_2 à 48 et un pH à 7,43, nous faisant poser la question d'une oxygénothérapie au long cours.

La réalisation de plusieurs oxymétries a conclu à une bonne saturation nocturne sous O_2 à 2 l/mn. L'EFR montre un trouble ventilatoire mixte à prédominance obstructive avec un VEMS à 40%.

Au total : décompensation de BPCO sur surinfection bronchique suivie de modifications électrocardiographiques dont la réalisation de la coronarographie avait montré une occlusion de l'IVA et une occlusion de la coronaire droite, un traitement médical a été instauré.

Il sera bien sur utile que le patient soit revu par son cardiologue habituel (Docteur Cœur). Le traitement de sortie est le suivant :
- Ikorel 20 mg : 2/j,
- Elisor 20 mg : 1/j,
- Temerit 5 mg : 1 x 2/j,
- old : 2 l/mn, 18h/24,
- Plavix 75 : 1/j,
- Lasilix 40 : 1/j,
- Xyzall 5 : 1/j,
- Omacor : 1/j,
- Previscan : 1/4 /j,
- Symbicort 400 : 2 x 2/j,
- Spiriva : 1 gel le matin,
- Triatec 5 mg : 1/j.

Bien cordialement.

Docteur Leblanc Marie Bernard, interne.

Figure 13. Exemple *fictif* de courrier de sortie non dé-identifié

le @@@

A

@@@
@@ de la @
@@

sa/nm

cher confrere,

votre patient, @@@, ne le @@@, @@@ des @@@, a ete brievement hospitalise dans le service du @@@ au @@@, à la suite de la coronarographie.

ce patient, comme vous le savez, est porteur d'antecedents :

- d'infarctus du myocarde en @ pour lequel il a ete stente,
- il est aussi porteur d'une bpco,
- d'une insuffisance cardiaque,
- et d'un dnid.

a l'entree dans le service, le malade etait stable. a l'examen clinique, on retrouve a l'auscultation pulmonaire quelques sibilants moderes. l'abdomen est souple. le reste de l'examen clinique est sans particularite.

la coronarographie avait conclu a l'occlusion de l'iva avec occlusion de la coronaire droite avec un reseau de collateralites assez developpees. l'importance des calcifications et l'etat global des arteres contre indiquaient l'angioplastie. suite a cette coronarographie, nous nous sommes orientes vers un traitement medical de cette cardiopathie ischemique. un traitement par anti-coagulants efficaces avec un relais par previscan a ete instaure par le @@. le plavix a ete poursuivi, des beta bloquants de type temerit 5 mg ont ete introduits.

la radiographie pulmonaire a montre des signes en faveur d'un oedeme pulmonaire avec oedeme interstitiel bilateral et un epanchement pleural bilateral plus marque a droite.

malgre le traitement deja instaure suite a la precedente hospitalisation, on assistait a la persistance de l'hypoxie avec une po2 a 54, pco2 a 48 et un ph a 7,43, nous faisant poser la question d'une oxygenotherapie au long cours.

la realisation de plusieurs oxymetries a conclu a une bonne saturation nocturne sous o2 a 2 l/mn. l'efr montre un trouble ventilatoire mixte a predominance obstructive avec un vems a 40%.

au total : decompensation de bpco sur surinfection bronchique suivie de modifications electrocardiographiques dont la realisation de la coronarographie avait montre une occlusion de l'iva et une occlusion de la coronaire droite, un traitement medical a ete instaure.

il sera bien sur utile que le patient soit revu par son cardiologue habituel (@@). le traitement de sortie est le suivant :

- ikorel 20 mg : 2/j,
- elisor 20 mg : 1/j,
- temerit 5 mg : 1 x 2/j,
- old : 2 l/mn, 18h/24,
- plavix 75 : 1/j,
- lasilix 40 : 1/j,
- xyzall 5 : 1/j,
- omacor : 1/j,
- previscan : 1/4 /j,
- symbicort 400 : 2 x 2/j,
- spiriva : 1 gel le matin,
- triatec 5 mg : 1/j.

bien cordialement.

@@@ @@

Figure 14. Exemple *fictif* de courrier de sortie dé-identifié

II. Évaluation de l'efficacité de FASDIM

Les 508 courriers de sortie sont évalués. Le rappel est de 98,1 %, la précision est de 79,6 % et la F-measure de 87,9 % (Tableau 8).

Mesures	Valeurs
Nombre de courriers de sortie	508
Nombre total de PHI	9 914
Nombre moyen de mots par courrier	510
Nombre moyen de PHI par courrier	20
Faux Positifs (FP)	2 537
Faux Négatifs (FN)	183
Nombre FN par courrier	0,38
Rappel (R)	98,1%
Précision (P)	79,6%
F-measure (F)	87,9%

Tableau 8. Résultats de l'évaluation de l'efficacité de FASDIM

Les courriers contiennent en moyenne 510 mots chacun.
Ces courriers contiennent 9 914 PHI soit en moyenne 20 PHI par courrier.
183 faux négatifs sont retrouvés parmi les 508 courriers de sortie : 183 PHI ont été ignorés par FASDIM, soit 0,38 PHI par courrier. La nature de ces PHI est détaillée dans le Tableau 9 selon les catégories définies dans le Tableau 7 (voir Méthode, II.A.3, page 46).

Catégorie de PHI	Proportion
Information partielle sur un lieu	63,7 %
Noms de soignants	23 %
Données biométriques (poids uniquement)	5,5 %
Portion de dates ou âge	4,4 %
Noms de structures de soins	3,3 %
Noms de patient	0%
Autres nombres	0%

Tableau 9. Détails des faux négatifs

Au regard de ce détail, nous pouvons considérer que la confidentialité du patient est préservée. En effet, parmi les PHI restant dans les courriers dé-identifiés, aucun n'appartient à la catégorie des noms-prénoms de patient et aucune date de naissance n'est retrouvée. Près de 63 % des PHI restants sont des portions d'adresse, dont aucune ne permet d'identifier précisément une adresse. La deuxième catégorie de PHI ignorés est représentée par les noms de soignants : 23 %. Ces noms ne seront pas nécessairement identifiés comme étant des noms de soignants par un lecteur ayant en première lecture le courrier dé-identifié : le mot laissé en place à tort se trouve généralement dans un paragraphe entièrement dé-identifié (par exemple dans l'en-tête comprenant la liste des médecins correspondants associée à leur adresse) et sans contexte la compréhension d'une donnée est plus complexe.

Dans un deuxième temps, les auxiliaires ou verbes supprimés à l'issue de l'étape de suppression des motifs et civilités (voir Méthode, I.D, page 36) ne sont plus considérés comme faux positifs dans nos calculs. Nous pouvons donc exposer des

résultats moins sévères mais plus réalistes. Seule la précision est modifiée comme il est montré dans le Tableau 10. La précision passe alors de 79,6 % à 89,2 %, et la F-measure de 87,9 % à 93,4 %.

Mesures	Valeurs
Nombre de courriers de sortie	508
Nombre moyen de mots par courrier	510
FP	1 340
Rappel (R)	98,1%
Précision (P)	89,2 %
F-measure (F)	93,4 %

Tableau 10. Résultats de l'évaluation de l'efficacité de FASDIM, deuxième version

III. Évaluation de la perte d'information

Les 508 courriers précédemment évalués sont repris et évalués de nouveau sur le versant cette fois de la perte d'information engendrée par le traitement par FASDIM.

A. Résultats totaux

Dans les courriers non dé-identifiés, 15 563 concepts ont été dénombrés, pour 15 411 concepts dans les courriers dé-identifiés. Le taux de conservation de l'information médicalement pertinente, toutes catégories confondues est donc de 99,02 %. Le détail par catégorie est donné dans le Tableau 11.

Catégorie d'information	Proportion d'information conservée
Actes CCAM	99,66 %
Diagnostics CIM-10 et autres codes CIM-10	99,49 %
Actes CIM-10	98,92 %
Noms de médicaments	98,84 %
Résultats de biologie CIM-10	96,99 %
Toutes catégories	99,02 %

Tableau 11. Taux de conservation de l'information par catégorie

Un courrier de sortie contient en moyenne 30 données à visée médicale.

B. Concordance inter-experts

La concordance entre experts n'a pas été évaluée sur un jeu de courriers commun. Cependant, les résultats obtenus sur des courriers différents semblent montrer que les experts trouvent des résultats comparables (voir Tableau 12).

Catégorie d'information	Expert 1	Expert 2
Nombre de courriers évalués	169	340
Actes CCAM	99,65 %	99,67 %
Diagnostics CIM-10 et autres codes CIM-10	99,32 %	99,57 %
Actes CIM-10	100 %	98,81 %
Noms de médicaments	99,34 %	98,59 %
Résultats de biologie	96,56 %	97,2 %
Toutes catégories	99,04 %	99,015 %

Tableau 12. Évaluation de la perte d'information : détails des résultats entre experts

IV. Évaluation du temps de travail

Un des objectifs est de mettre au point une méthode qui puisse être aisément reproduite par tout hôpital. Le temps de travail requis pour installer et reproduire FASDIM sans matériel initial est présenté sur la Figure 15.

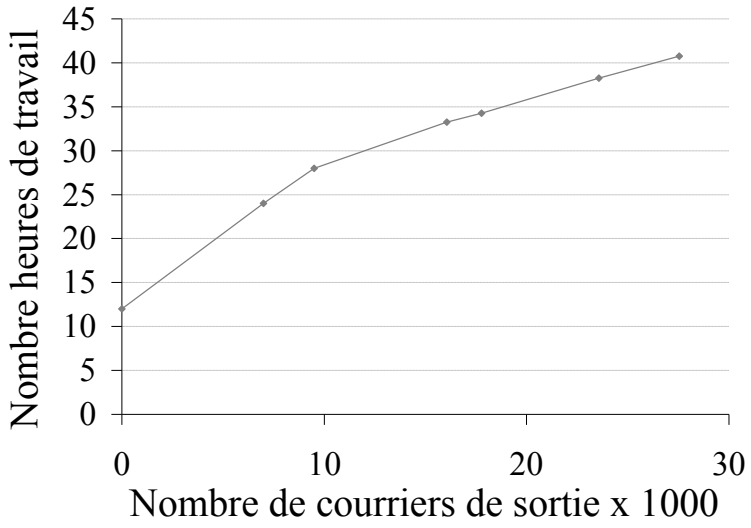

Figure 15. Nombre d'heures de travail en fonction du nombre de courriers à dé-identifier

Le temps requis comprend tout d'abord la conception et l'élaboration du programme (12 heures) et ensuite la création et la mise à jour de la liste de mots autorisés s'il est décidé de l'implémenter. Le temps nécessaire à la mise à jour de la liste de mots autorisés est dû au fait que chaque nouveau set de courriers est susceptible de contenir de nouveaux mots jamais attribués à une des deux listes. Les nouveaux courriers amènent de moins en moins de nouveaux mots non listés, et parmi ces nouveaux mots, la proportion de mots mal orthographiés augmente [Tu 2010]. Le nombre de mots en fonction du nombre de courriers est présenté dans la Figure 16.

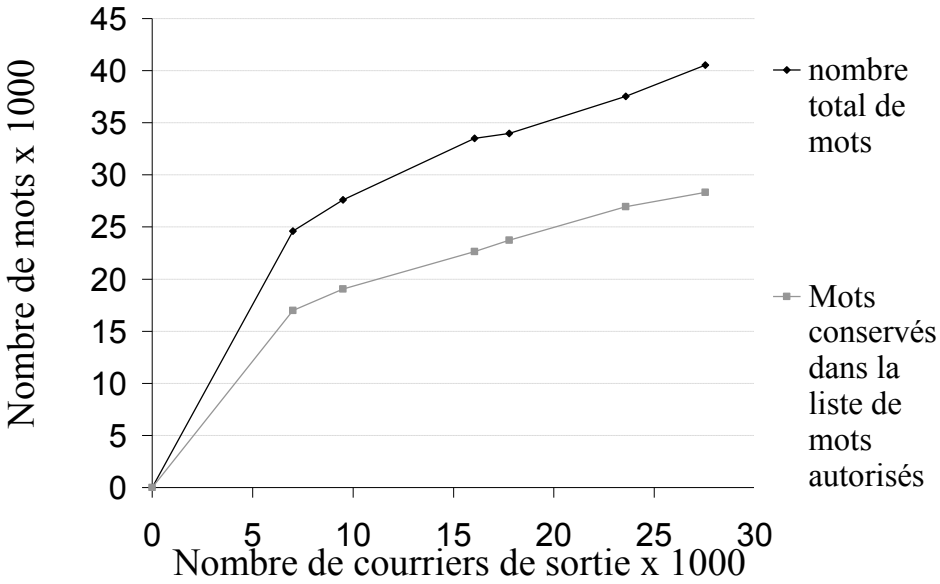

Figure 16. Nombre de mots contenus dans les courriers en fonction du nombre de courriers

Dans cette approche qui doit être indépendante du langage, les accords et les formes conjuguées des verbes sont considérés comme des mots différents.
La première extraction de mots a été faite sur 7 000 courriers. Ces 7 000 courriers nous ont fourni 24 600 mots différents dont 17 600 ont été conservés dans la liste de mots autorisés. La dernière itération effectuée montrait un nombre total de mots différents de 40 547 avec une liste de mots autorisés à 28 327, les nouveaux mots étant au nombre de 3 000 (la moitié soit environ 1 500 ayant été portés dans la liste de mots autorisés).

Discussion

Devant la nécessité d'assurer une lecture experte de dizaines de milliers de documents en texte libre contenus dans les dossiers médicaux hospitaliers, l'élaboration de FASDIM (voir section Méthode I, page 30), *a Fast and Simple De-Identification Method*, permet d'enlever de tous ces documents les données personnelles et identifiantes, respectant ainsi l'obligation légale de protection de la vie privée du patient, et de pallier l'absence de méthode efficace et libre de droits en langue française.
FASDIM est une méthode dont le concept repose sur la création d'une liste de mots autorisés à apparaître dans les courriers médicaux (voir Méthode, I.E, page 38), associée à la suppression des données numériques (excepté certains motifs définis) (voir Méthode, I.E, page 38) ainsi qu'à la suppression de motifs de titres et de civilités (voir Méthode, I.D, page 36).
Une évaluation de notre méthode de dé-identification, portant à la fois sur l'efficacité, la conservation de l'information médicale et la charge de travail requise (voir Méthode, II, page 45) montre d'excellents résultats (voir Résultats, II, page 53; III, page 54; IV, page 55) : ainsi le rappel est de 98,1 %, la précision de 79,6 %, la F-measure de 87,9 % ; le taux de conservation de l'information de 99,02 %.

FASDIM est une méthode simple de dé-identification : elle ne requiert aucune liste préalable à construire, à se procurer ou à mettre à jour. Plusieurs autres méthodes en effet nécessitent d'obtenir des listes exhaustives de noms de famille [Gupta 2004, Taira 2002, Aramaki 2006, Fielstein 2004, Beckwith 2006, Friedlin 2008, Neamatullah 2008, Sweeney 1996, Szarvas 2007, Thomas 2002], une liste de lieux [Aramaki 2006, Beckwith 2006, Fielstein 2004, Friedlin 2008, Guo 2006, Neamatullah 2008, Szarvas 2007] ou une liste de termes médicaux [Berman 2003, Friedlin 2008, Morrison 2009, Szarvas 2007, Thomas 2002], etc.
Il faut noter que la liste des noms et prénoms des patients fournie par le SIH est utile mais optionnelle ; le fait de ne pas la mettre à disposition ne compromet pas la confidentialité, grâce à la liste de mots autorisés.
De plus, contrairement à d'autres méthodes recourant au *machine learning*, aucun set d'entraînement n'a besoin d'être construit, et l'efficacité de la méthode est donc indépendante du nombre de courriers à anonymiser.
Un autre avantage à souligner est l'exportabilité de la méthode : en effet, l'approche considérant les mots comme étant indépendants les uns des autres, sans lien sémantique (on entend par là qu'aucun lien n'est établi entre deux formes conjuguées d'un même verbe par exemple, ou entre le singulier et le pluriel d'un mot), elle ne tient compte ni de la grammaire ni de la syntaxe et pourrait donc potentiellement être transposée dans une autre langue.

Cette méthode est rapide à mettre en place, puisque nous avons totalisé deux semaines de travail temps plein pour une personne à la phase initiale. Pour tout

établissement désireux de créer une méthode performante de dé-identification, ce temps est très écourté : le programme est disponible en *open source*, les listes de mots autorisés et interdits sont également mises à disposition du public.

A l'issue de cette évaluation, nous pouvons confirmer que les principales difficultés sont liées aux éponymes médicaux et aux fautes d'orthographe. Lors de la construction de la liste de mots autorisés, nous gardons ce que nous pensons être des fautes d'orthographe afin de conserver la lisibilité du texte; certaines sont flagrantes et relèvent certainement d'une erreur humaine.

> *Exemples :*
> - *« férosémide » pour « furosémide »*
> - *« karrdégic » pour « kardegic »*

Dans ces cas précis, la probabilité qu'il s'agisse d'un nom de famille ou de ville est très faible. Par contre, des mots que nous avons interprétés comme étant des fautes d'orthographe ou de frappe se sont avérés être des noms de famille ou de ville :

> *Exemples :*
> - *« mormal » (nom de famille) interprété comme le mot « normal »*
> - *« hergnie » (nom de ville) interprété comme le terme médical « hernie »*

Les éponymes médicaux posent le problème qu'ils peuvent être des noms de famille de patients ou de médecins, ainsi que des noms de rue.

Devant ces problèmes, il s'avère que la rentabilité de la mise à jour s'amenuise rapidement, étant donné l'augmentation importante de la proportion de fautes d'orthographe dans les nouveaux mots jamais listés.

> *Exemples :*
> - *Berger*
> - *Koch*

Les résultats obtenus quant à l'évaluation de l'efficacité de la méthode sont excellents et comparables à ceux obtenus par les méthodes en langue anglaise. Lors de l'évaluation d'un set de 508 courriers dé-identifiés par FASDIM, le rappel (R) est de 98,1 %, la précision (P) de 79,6 % et la F-measure 87,9 %. C'est la première fois qu'un système de dé-identification français atteint de tels scores (Grouin *et al.* ont obtenu R=65 %, P=23 % avec leur adaptation en français de DE-ID, et R=83 %, P=92 % avec leur système MEDINA [Grouin 2009]). De plus, nos résultats sont comparables avec ceux obtenus en anglais [Neamatullah 2008] (R=96,7 % ; P=74,9 %) ; [Beckwith 2006] (R=98,3 %) ; [Tu 2010] (R=80,2-86,7 % ; P=87,4-91,1 % ; F-measure=84-89 %). Les systèmes basés sur le *machine learning* donnent quant à eux de meilleurs résultats, avec une F-measure de 96 % pour [Szarvas 2007], 92 % [Hara 2006] et 86 % [Guo 2006] (voir le Tableau 2, page 23).
FASDIM est par ailleurs la première méthode à avoir été évaluée sur la conservation de l'information médicale dans les documents dé-identifiés. Le taux de conservation calculé à 99,02 % est en effet une mesure obtenue à partir de codes reconnus de manière internationale d'une part, puisqu'issus de la CIM-10 et nationale d'autre part

concernant les codes CCAM. Il existe toutefois une variabilité dans le codage effectué entre deux experts.
La méthode d'évaluation se veut rigoureuse, et cette rigueur s'en ressent dans le score de précision, à 79,6 %. En effet, une large proportion de faux positifs est constituée par les verbes (la plupart du temps des auxiliaires) et les prépositions retirés lors du traitement des motifs de civilités et de titres. Il s'agit d'un choix stratégique de départ assurant une meilleure garantie quant à la confidentialité du patient et nous avons choisi d'en tenir compte lors de notre évaluation. Cependant la perte d'information engendrée par ce choix est minime : la deuxième évaluation montre en effet un taux de conservation excellent de 99,02 %.
Nous avons également, au cours de notre évaluation de l'efficacité de la méthode, annoté dans le tableau le nombre de faux positifs dus aux verbes retirés dans le cadre de la suppression des titres et civilités. Cette annotation nous a permis de calculer rapidement sans charge de travail supplémentaire la précision lorsque les verbes retirés ne sont pas pris en compte. Le score atteint est donc de 89,2 % pour la précision, ce qui ramène la F-measure à 93,4 %, le rappel étant inchangé.
De même, notre corpus d'évaluation étant élaboré à partir d'un set de courriers provenant d'un même hôpital, une erreur (par exemple, un nom de médecin trié comme mot inclus alors qu'il aurait dû être attribué à la liste de mots interdits) peut être comptée un certain nombre de fois. Certains auteurs [Douglass 2005] ont choisi de ne compter qu'une fois un type d'erreur. Par ailleurs ces auteurs ont des scores très similaires aux nôtres, nous pouvons donc supposer que notre méthode est d'autant plus performante. Cependant, il faut noter que très peu d'auteurs communiquent avec précision sur leur annotation et la constitution de leur corpus d'évaluation. Deux articles mentionnent avec détails ces opérations. Concernant les autres méthodes présentées, au mieux l'article présente le classement par catégories des PHI recherchés sur les documents, le détail des faux positifs n'est pas fourni. Ce détail a peu d'importance au regard de l'évaluation de l'efficacité des méthodes, pourtant il renseigne en partie la perte d'information résultant du traitement des courriers. Cela renforce la pertinence de notre seconde évaluation se rapportant à la quantification de la conservation de l'information et à l'intelligibilité des courriers dé-identifiés.
Pour citer un autre exemple qui a pesé d'une manière certaine dans les résultats, le mot « peri » a été gardé, puisqu'il peut correspondre à une donnée anatomique. Lors de la revue des courriers, nous avons compté un concept restant comme portion d'adresse lorsque ce mot apparaissait pour une rue dont le nom est « Gabriel Peri ». Or, comme notre corpus d'évaluation provenait d'un même hôpital, une adresse de médecin généraliste comprenant ce nom de rue revenait régulièrement au cours de l'évaluation et donc était souvent comptée. Douglass *et al.* précisent dans leur méthode que seul le « type » d'erreur est compté, c'est-à-dire qu'ils ne dénombraient pas les redondances [Douglass 2005].

Certains auteurs renforcent la sécurité de leur *de identifier* en « pseudonymisant », c'est-à-dire en échangeant un PHI original avec un pseudonyme. Ainsi, une faille dans la méthode est quasi indétectable, puisque le texte présente des données nominatives (non pertinentes car ce ne sont pas les vraies), pour révéler une identité, il faudrait savoir que la méthode s'est trompée, et ensuite pouvoir repérer la localisation de l'erreur, c'est-à-dire trouver parmi les pseudonymes lequel n'en est pas un. Toutefois cette précaution impose que la méthode de dé-identification

catégorise les PHI pour remplacer avec un pseudonyme réaliste (différencier les noms de famille des patients et les noms de ville par exemple, ou une date avec un code postal). Le concept de pseudonymisation requiert également des listes de pseudonymes.
La simplification typographique appliquée dans la méthode peut également être un frein à la lecture, en diminuant le confort de lecture. De plus, dans d'autres langues comme le vietnamien par exemple, à alphabet latin, une simplification typographique entraîne trop de changements de sémantique pour obtenir des documents interprétables après dé-identification.

Comme nous l'avons déjà souligné, plusieurs points de méthode sont optionnels : il s'agit de la liste de noms et prénoms des patients provenant du SIH et de la mise à jour de la liste de mots autorisés.
La liste de noms et prénoms des patients est une sécurité supplémentaire, son absence ne menace cependant pas la confidentialité des données puisque l'élimination des motifs comprenant des titres associée à la suppression des mots n'appartenant pas à la liste de mots autorisés assure déjà la suppression des données identifiantes. Cet aspect permet d'anonymiser les courriers, mais d'une manière partielle : en effet, il est fréquent que les noms ou les prénoms soient mal orthographiés dans le texte contenu dans les courriers. En plus des fautes de frappe usuelles, plusieurs cas entraînent des différences entre les noms fournis dans la base structurée du SIH et les noms contenus dans le texte libre des courriers : les noms d'origine étrangère, spécialement s'ils sont longs et/ou retranscrits à partir d'un alphabet non latin, les noms avec particules, les noms fournis par des personnes illettrées, les noms de naissance versus noms maritaux (parfois non contenus dans la base). Ce point est important à prendre en compte, puisqu'un nom, même mal orthographié, peut permettre de retrouver une personne. Le problème par ailleurs est connu et décrit [Uzuner 2007, Berman 2003]. C'est pourquoi la méthode de dé identification que nous avons développée pallie cette difficulté et rend optionnelle la disponibilité d'une base fournie par le SIH.
Ensuite, la mise à jour de la liste de mots autorisés permet dans les faits de conserver la précision (c'est-à-dire la proportion de faux positifs) et donc la lisibilité et la compréhensibilité du texte. Le rappel est conservé puisque les nouveaux mots, s'ils ne sont pas traités, n'apparaîtront pas dans le texte.

La dé-identification en Santé Travail

I. Introduction : Le Dossier Médical en Santé au Travail

A. Définition

1. Le Dossier Médical en Santé au Travail (DMST)

Le DMST est évoqué dans divers articles du Code du Travail [CT R.7214-20, CT R.7214-21, CT D.4624-46, CT R.4412-56]. La tenue du dossier médical du salarié est réglementairement obligatoire. Les dossiers médicaux constituent un support d'informations indispensables qui doit pouvoir permettre le lien entre les différentes expositions d'un salarié et la survenue, éventuelle, d'une pathologie, pendant toute la durée du travail des salariés. Les données consignées permettent également de repérer des états de santé nécessitant un aménagement du poste de travail ou des restrictions d'aptitude. C'est l'élément de base du suivi médical des salariés à leur travail, leur finalité étant de retracer un historique des expositions et d'effectuer un suivi adapté [ISTNF 2011].

La Haute Autorité de Santé (HAS) fournit également une définition du DMST dans sa recommandation sur le dossier médical en santé travail de janvier 2009 : « Lieu de recueil et de conservation des informations socio-administratives, médicales, et professionnelles, formalisées et actualisées, nécessaires aux actions de prévention individuelles et collectives en santé au travail, enregistrées, dans le respect du secret professionnel, pour tout travailleur exerçant une activité, à quelque titre que ce soit, dans une entreprise ou un organisme, quel que soit le secteur d'activité. Le DMST est individuel. Tenu par le médecin du travail, le DMST peut également être alimenté et consulté par les personnels infirmiers de travail collaborateurs du médecin du travail, sous la responsabilité et avec l'accord de celui-ci, dans le respect du secret professionnel et dans la limite de ce qui est strictement nécessaire à leur mission. » [HAS 2011].

La recommandation de l'HAS précise également les objectifs du DMST. Le DMST doit aider le médecin du travail à apprécier le lien entre l'état de santé du travailleur d'une part et le poste et les conditions de travail d'autre part. Le DMST est également un support pour proposer des mesures de prévention et faire des propositions en termes d'amélioration ou d'aménagement du poste ou des conditions de travail et de maintien ou non dans l'emploi. Un objectif important du DMST est de participer à la traçabilité des expositions professionnelles, des informations et conseils de prévention professionnels délivrés au travailleur.

Le DMST a pour autre vocation d'aider le médecin du travail à participer à la veille sanitaire en santé travail [HAS 2011].

2. Cas du Dossier Médical Informatisé de Médecine du Travail (DMIMT)

Il n'y a aucune distinction déontologique entre le DMST « papier » et le DMIMT [ISTNF 2011]. La différence réside dans les moyens matériels d'appliquer les règles de confidentialité dues au salarié. La constitution du DMIMT passe par un logiciel permettant la saisie par le médecin des données médicales, personnelles et professionnelles. Il existe une multiplicité de l'offre de logiciels dans ce domaine, ce qui pose un problème d'harmonisation des pratiques au niveau régional entre les différents Services de Santé au Travail. En effet, les données conservées dans un logiciel donné ne sont pas exportables telles que dans un autre logiciel lambda, ce qui entrave la poursuite du DMST en cas de changement de SST pour le salarié, dans la mesure où ce dernier donne son accord à la transmission des données le concernant. Or, cette poursuite est à privilégier autant que possible car elle permet la traçabilité des expositions au cours de la carrière professionnelle.

B. Contenu

Tout comme le dossier médical du patient hospitalisé, le Dossier Médical en Santé au Travail comprend des données objectives (antécédents, plaintes, examens cliniques et complémentaires) qui sont communicables au patient et au(x) médecin(s) de son choix, et des données subjectives (notes personnelles du médecin). Ce contenu est détaillé dans la recommandation de la HAS de janvier 2009 [HAS 2011] et est présenté dans le tableau 13 (voir Tableau 13). Cette table n'est pas exhaustive : il convient d'ajouter certains champs spécifiques selon la profession du salarié.

Éléments correspondant aux données objectives du DMST
• Identification du salarié • Nom et adresse de l'employeur • Date d'embauche • Les différents postes de travail occupés précédemment dans l'entreprise actuelle et les entreprises précédentes • Éléments du poste de travail • Profil du poste de travail actuel et ses risques inhérents connus susceptibles d'avoir des répercussions sur la santé du salarié • Antécédents médicaux • Conclusions des examens des visites médicales initiales et successives • Résultats des examens complémentaires • Fiches d'exposition • Attestations d'exposition ouvrant droit au bénéfice du Suivi Post-Professionnel (SPP) • Courriers médicaux et divers • Avis d'aptitude, restrictions et réserves • Conseils de prévention donnés • Éventuels avis demandés au Médecin Inspecteur Régional du Travail (MIRT)
Éléments correspondant aux données subjectives du DMST
• Informations non médicales, sans relation avec l'activité de prévention, relevant de la confidence du salarié • Correspondance comprenant des éléments médicaux en confidence, sans relation avec l'avis d'aptitude et ne remettant pas en cause ce dernier • Notes personnelles du médecin du travail (au sens de l'arrêté du 5 mars 2004, JO 17 mars 2004) • Informations susceptibles de dévoiler un secret de fabrique ou des informations confidentielles de l'entreprise • Courriers émanant de l'employeur

Tableau 13. Contenu du DMST

La HAS précise qu'il est souhaitable que le DMST soit informatisé. Le choix d'un logiciel de gestion du DMST est soumis à certaines conditions, toujours selon la HAS. Ce logiciel doit permettre une exploitation collective des données issues des DMST par le médecin du travail ou en coopération avec d'autres médecins du travail. C'est-à-dire que le logiciel autorise la collection et l'analyse anonymisées de données de santé des individus et des données concernant l'emploi et les activités professionnelles, tout en permettant une utilisation aisée des thésaurus recommandés.

II. Matériel et Méthode

A. Services inter entreprises de Santé au Travail

Le premier point consiste à établir la façon dont sont gérés concrètement les dossiers médicaux en santé travail (les DMST) au sein des services de Santé-Travail. Après présentation du projet concernant la dé-identification des courriers médicaux par Monsieur le Professeur Frimat auprès du directeur du service inter entreprises AST 62-59, Monsieur Alain Cuisse, une réunion est organisée avec Monsieur Cuisse, Directeur Général, Monsieur Duflo, Directeur Opérationnel et moi-même.
A l'issue de cette réunion où des informations relatives à la gestion des DMST, à leur archivage et à leur informatisation, sont échangées, un contact par email est pris

avec un des médecins du travail investi dans l'aspect informatisation des dossiers et connaissant bien le logiciel d'exploitation des DMST. Par la suite, les entretiens sont menés par téléphone et par courrier électronique.

B. Logiciel de saisie et stockage du dossier médical informatisé en santé travail

Les informations concernant le logiciel sont complétées grâce à la communication avec un ingénieur commercial de la société Integral Data Santé (IDS) commercialisant le logiciel DINAMIT (Dossier Informatisé pour les Autonomes et Médecines INterprofessionnelles du Travail). Nous avons en effet pu mener par email un entretien avec M. Pirroux, ingénieur commercial chez IDS, qui a pu préciser certains points et nous fournir une documentation sur DINAMIT. Cette société est basée en région parisienne et est spécialisée dans la mise en place de solutions informatiques pour la santé au travail [IDS 2011]. D'après leurs sources, il existe 1 200 médecins utilisateurs gérant plus de 4 millions de salariés, répartis dans 150 services de Santé au Travail.

III. Résultats

A. Entretien avec le service interentreprises AST 62-59

Après un entretien avec M. Cuisse, Directeur de l'AST62-59, et M. Duflo, Directeur opérationnel de l'AST 62-59, il apparaît que les dossiers médicaux en santé travail sont peu informatisés. La faible proportion de dossiers informatisés est stockée telle quelle dans le logiciel avec lequel ils ont été constitués (DINAMIT). Il n'y a pas d'extraction de fichiers, même sous un format allégeant la base de données. L'idée avancée par la dé-identification automatisée des courriers médicaux est qu'elle rendrait disponible une base de données conséquente permettant de faire des études statistiques entre les expositions professionnelles et certaines pathologies, sans passer par une démarche fastidieuse et coûteuse de recueil des consentements libres et éclairés des salariés.

B. Société Integral Data Santé (IDS)

L'ingénieur commercial de la société IDS qui commercialise le logiciel DINAMIT a communiqué des précisions sur l'exploitation des données issues des DMST.
Il est actuellement possible d'extraire des fichiers de format Excel. Les statistiques incluses dans DINAMIT (hors Excel), sont établies par période, par type (dangers, pathologie CIM10, résultats d'examens complémentaires, Accidents du Travail, pyramide des âges, etc.) et elles sont fonction d'une population de salariés (un poste de travail dans une entreprise, le service d'une entreprise, une entreprise, un groupe d'entreprises, (médecin(s),secteur(s), NAF, centre(s)) ou tous les salariés du service de santé (postes ouverts, postes clôturés, tous les postes).
Toutefois, ces statistiques peuvent nécessiter une relecture par un expert afin de valider certains résultats. Les informations utiles à cette relecture sont souvent contenues dans des courriers médicaux, leur mise à disposition est donc cruciale et prend tout son sens. A l'heure actuelle, ces possibilités ne sont pas exploitées par les SST.

IV. Discussion

Nous avons vu précédemment l'intérêt de dé-identifier les courriers (voir Introduction, II.C, page 15) : la quantité de données exploitables permise par la dé-identification offre des possibilités d'automatisation de fouille de données, fouille de données qui peut révéler des tendances et/ou des relations entre des expositions professionnelles et des pathologies. La première démarche qui pourrait être menée dans l'optique de recherche de ce type en santé travail peut être initiée au sein même de l'hôpital. En effet, dès que l'information concernant l'exposition professionnelle est renseignée, les méthodes de *Data Mining* permettent de parcourir des dizaines de milliers de séjours et de révéler des liens statistiques entre ces expositions professionnelles et certaines pathologies, grâce notamment au codage PMSI.
Dans cette approche, l'intérêt de la dé-identification revêt deux aspects : les documents sont disponibles tout d'abord pour la fouille de données, puis accessibles à la lecture experte permettant de valider les résultats ayant pu ressortir de cette fouille de données automatisée.
Par exemple, un pôle comme le service de traumatologie pourrait être étudié. On pourrait alors rechercher un lien statistique entre une pathologie pouvant être d'origine professionnelle comme la rupture de la coiffe des rotateurs et des expositions professionnelles ou un type de profession.

Au sein des Services de Santé Travail, pour lesquels nous avons constaté que l'informatisation des DMST en est au stade d'ébauche (voir La dé-identification en Santé Travail, III, page 64), la dé-identification automatisée des courriers médicaux mettrait à disposition un outil de validation indispensable à la relecture experte nécessaire à l'issue des processus de fouille de données. L'état des lieux en matière d'informatisation des DMST dans la région montre la faible proportion de DMST informatisés et l'hétérogénéité des logiciels utilisés dans les services, puisqu'à notre connaissance le service Pôle Santé Travail (regroupant les secteurs de Lille et Douai) utilise le logiciel STHETO. Cette faible utilisation de l'outil informatique et le stockage tel que au sein des logiciels limitent donc à l'heure actuelle une application directe de la dé-identification. Toutefois les problèmes soulevés par l'archivage des DMST sous forme papier amènent à développer leur informatisation. Selon le GISSET (Groupement Inter Services Santé et Travail), en 2005, l'archivage des dossiers médicaux pour la région Nord-Pas-de-Calais représente 1 600 mètres carrés d'archives (hors archives en entreprise), 4 000 mètres linéaires (toujours hors archives en entreprises) pour plus de 3 000 000 de dossiers archivés dans les services.

Une autre approche pourrait être l'évaluation automatisée de la qualité des DMST. Dans une démarche de mise en place d'indicateurs qualité, ces indicateurs pourraient être recherchés de manière automatisée dans une grande quantité de dossiers. Par exemple, un indicateur de traçabilité comme le fait que l'exposition professionnelle soit renseignée lors d'un changement de poste de travail pourrait être étudié de manière automatisée.

L’avantage indiscutable de l’outil FASDIM serait l’allègement des démarches nécessaires aux études sur un très grand nombre d’informations, études ne nécessitant pas de suivi de patients ou salariés dans le temps.

Conclusion

FASDIM est une méthode de dé-identification performante qui permet de respecter la confidentialité des patients tout en traitant une grande quantité de documents médicaux, jusqu'à plusieurs dizaines de milliers. Cette méthode est fiable et facilement transposable dans tout établissement de santé désirant dé-identifier des documents en texte libre.
FASDIM est une méthode novatrice, comblant un vide dans les besoins de dé-identification des documents médicaux en texte libre et en langue française. La méthode sera bientôt disponible en ligne gratuitement, permettant ainsi le partage inédit jusque maintenant d'une véritable méthode de dé-identification efficace et sûre. Les résultats obtenus (R=98,1 %, P=79,6 %, F-measure=87,9 %, taux de conservation de l'information médicale=99,02 %) montrent le respect de la confidentialité des données personnelles du patient, la conservation à la fois de la lisibilité et de l'interprétabilité des documents dé-identifiés. Ces résultats ont également un caractère exclusif puisqu'aucun calcul objectif de la conservation de l'information contenue dans les documents dé-identifiés n'a jamais été publié.

Bibliographie

Aberdeen 2010	Aberdeen J, Bayer S, Yeniterzi R, et al. *The MITRE Identification Scrubber Toolkit: Design, training, and assessment.* International Journal of Medical Informatics 2010 Dec;79(12):849-59
Aramaki 2006	Aramaki E, Miyo K. *Automatic Deidentification by Using Sentence Features and Label Consistency*. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006
Baceanu 2009	Baceanu A, Atasiei I, Chazard E, Leroy N, the PSIP Consortium. *The expert explorer: a tool for hospital data visualization and adverse drug event rules validation.* Stud Health Technol Inform. 2009;148:85-94.
Beckwith 2006	Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. *Development and evaluation of an open source software tool for deidentification of pathology reports.* BMC Med Inform Decis Mak. 2006;6:12
Berman 2003	Berman JJ. *Concept-match medical data scrubbing. How pathology text can be used in research* Arch. Pathol. Lab. Med. 2003;127(6):680-86
CD 4	Article 4 du Code de Déontologie Médicale Article R.4127-4 du Code de la Santé publique http://www.conseil-national.medecin.fr/sites/default/files/codedeont_1.pdf
CD 72	Article 72 du Code de Déontologie Médicale Article R.4127-72 du Code de la Santé publique
CD 73	Article 73 du Code de Déontologie Médicale Article R.4127-73 du Code de la Santé publique
Chazard 2009	Chazard E, Merlin B, Ficheur G, Sarfati JC, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform. 2009;148:63-74.
CNIL 2011 (1)	http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17_definitive-annotee.pdf
CNIL 2011 (2)	www.cnil.fr/dossiers/sante/fiches-pratiques/article/communication-des-donnees-de-sante
CNIL 2011 (3)	www.cnil.fr/la-cnil/actu-cnil/article/article/letat-des-lieux-en-matiere-de-procedes-danonymisation
CP 226-13	Article 226-13 du Code Pénal
CP 226-14	Article 226-14 du Code Pénal
CSP L.1110-4	Article L.1110-4 du Code Pénal
CSP L.1111-4	Article L.1111-4 du Code Pénal
CSP L.1111-7	Article L.1111-7 du Code Pénal

CT D.4624-46	Article D.4624-46 - Décret n°86-569 du 14 mars 1986 modifiant le Code du Travail
CT R.4412-56	Article R.4412-56 du Code du Travail
CT R.7214-20	Article R.7214-20 – Décret n°86-569 du 14 mars 1986 modifiant le Code du Travail
CT R.7214-21	Article R.7214-21 - Décret n°86-569 du 14 mars 1986 modifiant le Code du Travail
Dalianis 2010	Dalianis H, Velupillai S. *De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields*. J Biomed Semantics. 2010;1(1):6
Dorr 2006	Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. *Assessing the difficulty and time cost of de-identification in clinical narratives* Methods Inf Med. 2006;45(3):246-52
Douglass 2005	Douglass MM, Clifford GD, Reisner A, Long WJ, Moody GB, Mark RG. *Computer-Assisted De-Identification of Free-text Nursing Notes.* 2005;32 :331-334
Fielstein 2004	Fielstein EM, Brown SH, Speroff T. *Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings*; Medinfo. 2004
Friedlin 2008	Friedlin FJ, McDonald CJ. *A Software Tool for Removing Patient Identifying Information from Clinical Documents.* J Am Med Inform Assoc. 2008;15(5):601-10
Gardner 2008	Gardner J, Xiong L. *HIDE: An Integrated System for Health Information De-identification.* Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems 2008;254-9
Grouin 2009	Grouin C, Rosier A, Dameron O, Zweigenbaum. *Testing tactics to localize de-identification.* Stud Health Technol Inform. 2009;150:735-39
Guo 2006	Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple R. *Identifying Personal Health Information Using Support Vector Machines*. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006
Gupta 2004	Gupta D, Saul M, Gilbertson J. *Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research.* Am. J. Clin. Pathol. 2004;121(2):176-86
Hara 2006	Hara K. *Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge* i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006a
HAS 2011	www.has-sante.fr/portail/jcms/c_757826
Hirschman 2010	Hirschman L, Aberdeen J. *Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts.* Text and Data Mining of Health Documents. 2010;72-5
Hripcsak 2005	Hripcsak G, Rothschild AS. *Agreement, the F-measure, and Reliability in Information Retrieval*. J Am Med Inform Assoc. 2005;12(3):296-98
IDS 2011	www.ids-fr.com

ISTNF 2011	http://www.istnf.fr/_admin/Repertoire/Fichier/2011/14-110315102815.pdf
Meystre 2010	Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. *Automatic de-identification of textual documents in the electronic health record: a review of recent research.* BMC Med Res Methodol. 2010;10:70
Morrison 2009	Morrison FP, et al. *Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?* J Am Med Inform Assoc. 2009;16(1):37-9
Neamatullah 2008	Neamatullah I, Douglass MM, Lehman LH, et al. *Automated de-identification of free-text medical records.* BMC Med Inform Decis Mak. 2008;8:32
Miro 2011	www.omiro.free.fr
Pantazos 2011	Pantazos K, Lauesen S, Lippert S, De-identifying an EHR Database – Anonymity, Correctness and Readability of the Medical Record. European Federation for Medical Informatics. 2011
PHP 2011	www.php.net
PSIP 2011	http://www.psip-project.eu
Ruch 2000	Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. *Medical document anonymization with a semantic lexicon*. Proc AMIA Symp. 2000:729-33
Sweeney 1996	Sweeney L. *Replacing personally-identifying information in medical records, the Scrub system.* Proc AMIA Annu Fall Symp. 1996:333-37
Szarvas 2007	Szarvas G, Farkas R, Busa-Fekete R. *State-of-the-art anonymization of medical records using an iterative machine learning framework*. J Am Med Inform Assoc. 2007 Oct;14(5):574-580
Taira 2002	Taira R, Bui A, Kangarloo H. *Identification of patient name references within medical documents using semantic selectional restrictions*. Proc AMIA Symp. 2002;757-61
Thomas 2002	Thomas SM, et al. *A successful technique for removing names in pathology reports using an augmented search an replace method*. Proc AMIA Symp. 2002;777-81
Tu 2010	Tu K, Klein-Geltink J, Mitiku TF, Mihai C, Martin J. *De-identification of primary care electronic medical records free-text data in Ontario, Canada*. BMC Med Inform Decis Mak. 2010;10:35
UMLS 2009	http://www.nim.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html
Uzuner 2007	Uzuner Ö, Luo Y, Szolovits P. *Evaluating the State-of-the-Art in Automatic De-identification.* J Am Med Inform Assoc. 2007;14(5):550-563
Uzuner 2008	Uzuner Ö, Sibanda TC, Luo Y, Szolovits P*. A De-identifier for Medical Discharge Summaries.* Artif Intell Med. 2008 Jan;42(1):13-35

Velupillai	Velupillai S, Dalianis H, Hassel M, Nilsson GH. *Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial.* International Journal of Medical Informatics. 2009 Dec;78(12):e19-e26
Wellner 2007	Wellner B, Huyck M, Mardis S, et al. *Rapidly Retargetable Approaches to De-identification in Medical Records*. J Am Med Inform Assoc. 2007;14(5):564-73
Wikipédia 2011 (1)	www.fr.wikipedia.org/wiki/Data_mining
Wikipédia 2011 (2)	www.fr.wikipedia.org/wiki/Précision_et_rappel
Yeniterzi 2010	Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. *Effects of personal identifier resynthesis on clinical text de-identification*. J Am Med Inform Assoc. 2010;17(2):159-68

Figures & Tableaux

Auteur : MOURET-KUBIAK Capucine
Date de soutenance : 17 janvier 2012
Titre de la thèse : Dé-identification automatisée de courriers médicaux : proposition et évaluation de la méthode FASDIM
Thèse, Médecine, Lille, 2012
Cadre de classement : DES de Médecine du Travail
Mots clés : anonymisation, dé-identification, confidentialité, secret médical, Natural Language Processing, Traitement Automatisé de Langage

Titre : Dé-identification automatisée de courriers médicaux : proposition et évaluation de la méthode FASDIM

Contexte : Les données médicales contenues dans les courriers médicaux constituent une source importante d'information pour la recherche clinique et les actions de Santé Publique. Ces données ne sont exploitables que si la vie privée du patient est respectée, ce qui peut être obtenu par la dé-identification des courriers médicaux, aussi appelée « anonymisation ». La dé-identification consiste à supprimer dans un document toutes les données directement ou indirectement nominatives permettant d'identifier une personne. Afin de traiter un grand nombre de courriers, la méthode de dé-identification doit être automatisée, cependant aucune méthode n'existe en langue française.
Matériel et méthode : FASDIM est une méthode simple et rapide de dé-identification de courriers médicaux en texte libre, permettant de supprimer les données identifiantes (*Protected Health Information* ou "PHI"). La méthode repose sur la création d'une liste de mots autorisés, associée à la création d'une liste de motifs dans lesquels les nombres sont protégés. Ensuite, les mots non autorisés ainsi que les nombres non protégés au préalable sont supprimés du texte. Enfin, FASDIM est évalué sur trois aspects: 1) efficacité à supprimer les données identifiantes, 2) perte d'information médicale engendrée par la méthode, 3) temps de travail nécessaire à sa mise en place et sa mise à jour.
Résultats : FASDIM permet de dé-identifier les courriers médicaux non structurés en langue française avec un rappel de 98,1 % et une précision de 79,6 %, soit une F-measure à 87,9 %, ce qui garantit la confidentialité des documents. L'évaluation de la perte d'information montre un taux de conservation de l'information médicale de 99,02 %. Le temps requis pour implémenter la méthode est d'une quarantaine d'heure. FASDIM est donc un outil simple, rapide à mettre en place et permettant de préserver la vie privée du patient.

Composition du jury :

Président :	Monsieur le Professeur BEUSCART
Assesseurs :	Monsieur le Professeur FRIMAT
	Monsieur le Professeur DUHAMEL
	Monsieur le Docteur CHAZARD
Directeur de Thèse :	Monsieur le Docteur CHAZARD

Adresse de l'auteur :
[redacted]