Medina — Manuel d'utilisation

Cyril Grouin
LIMSI–CNRS, Rue John von Neumann, 91400 Orsay
cyril.grouin@limsi.fr

11 janvier 2014

Table des matières

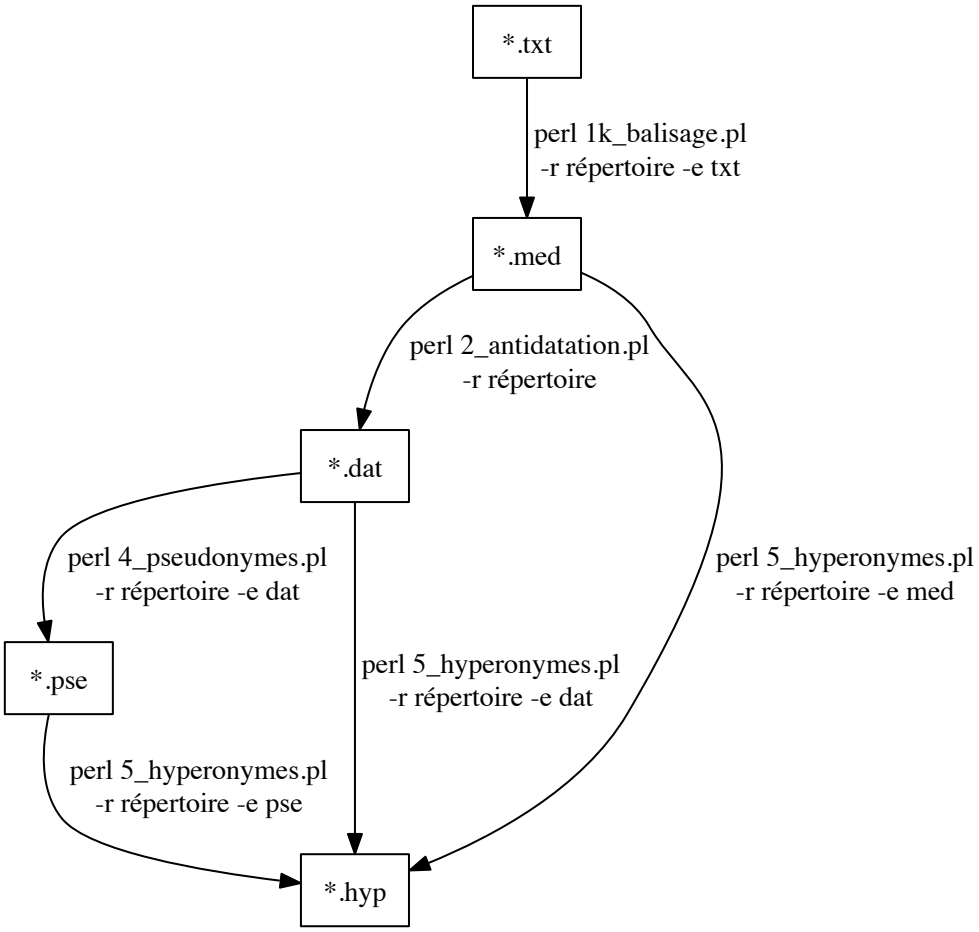
1 Présentation

Medina est un outil d'anonymisation des données personnelles présentes dans des documents textuels. Cet outil a été développé pour traiter des comptes rendus cliniques en cardiologie. L'outil se compose de plusieurs scripts, un premier permettant le balisage des informations à anonymiser, suivi de scripts de post-traitements pour procéder à l'anonymisation. L'outil a été développé entre 2008 et 2012 dans le cadre du projet Akenaton [1] pour anonymiser des comptes rendus médicaux en cardiologie. Si certains types d'informations personnelles sont transverses aux différentes disciplines médicales *(nom, prénom, adresse, téléphone, numéro de sécurité sociale...)*, d'autres sont spécifiques *(marques de défibrillateurs en cardiologie, références des dents en stomatologie, etc.)*. Les types de données traitées s'inspirent de la législation américaine HIPAA. Medina ne réalise pas la distinction entre médecin et patient mais conserve la distinction entre nom et prénom. Il est possible d'anonymiser par suppression des informations (hyperonymes) ou par remplacements (antidatation, pseudonymes).

1. *Automated Knowledge Extraction from medical records iN Association with a Telecardiology Observation Network*, financement ANR-07-TecSan-001-06

2 Lancement rapide

Deux options sont utiles :
-r : répertoire contenant les fichiers (obligatoire) ;
-e : extension des fichiers en entrée (inutile pour le script des dates).



2.1 Balisage des informations

```
perl 1k_balisage.pl -r <répertoire> -e <extension des fichiers>
```

2.2 Post-traitements

Modification des dates (antidatation aléatoire comprise entre 1 et 4 ans, ou antidatation selon le nombre fourni par l'utilisateur) ; fichier.med → fichier.dat :
```
perl 2_antidatation.pl -r <répertoire> [-n <nombre>]
```

Remplacement des noms et prénoms par un pseudonyme ; fichier.dat → fichier.pse :
```
perl 4_pseudonymes.pl -r <répertoire> -e <extension des fichiers>
```

Remplacement des données personnelles par un hyperonyme ; fichier.pse → fichier.hyp :
```
perl 5_hyperonymes.pl -r <répertoire> -e <extension des fichiers>
```

3 Utilisation détaillée

3.1 Architecture globale de l'outil

L'outil se compose des éléments suivants :
- un fichier de configuration : config ;
- un répertoire constitué des ressources linguistiques : data *(dictionnaire de mots communs, liste noire de mots ne devant pas être anonymisés, listes de noms de médecins, de noms d'hôpitaux, de noms de famille, de pays, de prénoms et de villes)* ;
- un script pour baliser les informations personnelles : 1k_balisage.pl
- des scripts annexes de post-traitements :
  - 2_antidatation.pl : remplace les dates précédemment identifiées par d'autres dates en conservant l'écart temporel entre chaque date à l'intérieur d'un même document ;
  - 4_pseudonymes.pl : remplace les occurrences de noms et prénoms par des pseudonymes (les 10 noms les plus portés en France et 10 prénoms mixtes) ;
  - 5_hyperonymes.pl : remplace les données précédemment identifiées par un hyperonyme (conservation des dates si script appliqué sur des fichiers *.dat ; conservation des noms et prénoms si appliqué sur des fichiers *.pse).

3.2 Configuration et lancement

3.2.1 Balisage des informations personnelles

L'outil d'anonymisation repose sur une première phase de repérage des informations à anonymiser. À l'issue de ce repérage, les informations seront encadrées de balises XML typant l'information identifiée. Le script produit des fichiers d'extension « *.med » dans le répertoire des documents.

1. Ouvrir le fichier de configuration avec un éditeur de texte et modifier les différents champs selon les besoins :
   - indiquer les informations qui doivent être anonymisées face à chaque catégorie *(adresses, âges, codes postaux, dates, hôpitaux, médicaments, mesures, noms, prénoms, numéro de sécurité sociale, référence des stimulateurs cardiaques, téléphones, unités hospitalières, villes)* ;
   - indiquer les listes de ressources linguistiques à utiliser ;
   - compléter les listes de déclencheurs ;
   - indiquer l'âge minimum au-delà duquel l'anonymisation de l'âge des patients est requise (la législation américaine HIPAA impose d'anonymiser les âges au-delà de 90 ans) ;
   - indiquer le format des balises à utiliser pour traiter les données.
2. Créer un répertoire contenant les documents au format textuel à anonymiser.
3. Lancer le script d'anonymisation au moyen de la commande suivante :
   `perl 1k_balisage.pl -r <répertoire> -e <extension des fichiers>`

3.2.2 Remplacement des informations identifiées

Une ou plusieurs étapes de post-traitements sont alors utiles pour procéder réellement à l'anonymisation.

1. Un script retranche à chaque date un nombre de jours aléatoirement tiré compris entre 365 et 1460 jours (soit entre 1 et 4 ans) ou retranche le nombre de jour fixé par l'utilisateur (option -n) ; ce nombre est le même pour toutes les dates d'un document, ce qui permet de conserver les écarts temporels entre deux dates tout en observant le principe d'anonymisation. Le format des dates est reproduit à l'identique. Le script produit des fichiers d'extension « *.dat » :
   `perl 2_antidatation.pl -r <répertoire> [-n <nombre>]`

2. Un second script remplace toutes les occurrences de noms et prénoms par des pseudonymes parmi l'un des 660 noms et prénoms mixtes les plus portés en France. [2] Toutes les occurrences d'un nom ou d'un prénom sont remplacées par le même pseudonyme à l'intérieur d'un document. [3] Le script produit des fichiers d'extension « *.pse » :
`perl 4_pseudonymes.pl -r <répertoire> -e <extension des fichiers>`
3. Un dernier script remplace les données personnelles balisées par un hyperonyme (la balise typant l'information) : <*ville*>*Versailles*</*ville*> → <*ville* />. Ce script est appliqué, soit sur les fichiers d'extension « *.med » (auquel cas toutes les informations sont anonymisées), soit sur les fichiers d'extension « *.pse » (toutes les informations autres que les noms, prénoms et dates seront anonymisées, les noms, prénoms et dates ayant été préalablement traités) :
`perl 5_hyperonymes.pl -r <répertoire> -e <extension des fichiers>`

4 Exemple

Le paragraphe d'exemple suivant est issu d'un compte rendu clinique en cardiologie. Toutes les informations personnelles (nom, prénom et dates) ont été modifiées par rapport à la version d'origine. Les dates qui figurent dans la version terminale anonymisée sont antidatées de 1377 jours (soit environ 3 ans 9 mois et demi) par rapport au fichier de base ; l'écart temporel entre chaque date est néanmoins conservé.

4.1 Fichier d'origine *.txt

Monsieur Théodore Bauche (21.07.53) est malheureusement revenu dans le service du 4 au 11 mai 2000 pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en octobre 99.

4.2 Fichier balisé *.med

Monsieur <prenom>Théodore</prenom> <nom>Bauche</nom> (<date>21.07.53</date>) est malheureusement revenu dans le service du <date>4 au 11 mai 2000</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>octobre 99</date>.

4.3 Fichier antidaté *.dat

Monsieur <prenom>Théodore</prenom> <nom>Bauche</nom> (<date>13.10.49</date>) est malheureusement revenu dans le service du <date>27 juillet au 3 août 1996</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>janvier 96</date>.

4.4 Fichier générique *.pse

Monsieur Claude Martin (<date>13.10.49</date>) est malheureusement revenu dans le service du <date>27 juillet au 3 août 1996</date> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date>janvier 96</date>.

2. Noms les plus portés : *Martin, Bernard, Dubois, Thomas, Robert, Richard, Petit, Durand, Leroy, Moreau*. Prénoms mixtes courants : *Alex, Camille, Charlie, Claude, Dominique, Louison, Maé, Maxime, Morgan, Stéphane*.

3. Dans le détail, tous les noms d'un document sont d'abord relevés puis triés par ordre alphabétique : le premier nom dans l'ordre alphabétique est remplacé par *Martin*, le second par *Bernard*, etc. Il en est de même pour les prénoms.

4.5 Fichier anonymisé *.hyp

Version post-traitée avec tous les scripts (antidatation, pseudonymes et hyperonymes)

Monsieur Claude Martin (13.10.49) est malheureusement revenu dans le service du 27 juillet au 3 août 1996 pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en janvier 96.

Version post-traitée directement après la sortie balisée (hyperonymes)

Monsieur <prenom /> <nom /> (<date />) est malheureusement revenu dans le service du <date /> pour la constitution d'un nouvel infarctus cette fois en territoire inférieur alors qu'il avait présenté un premier épisode d'infarctus en territoire antérieur en <date />.

5 Historique

18 novembre 2008 (version 1a). Création de l'outil dans le cadre du projet Akenaton pour anonymiser les comptes rendus cliniques en cardiologie.

23 novembre 2008 (version 1b). Le programme bouclait sur les expressions régulières des dates en raison des séparateurs définis dans la variable `$sep` (le point nétait pas suffisamment déspécialisé : `$sep="(\.|\-)"` au lieu de `$sep="(\\\\.|\-)")` ; en conséquence, le point était interprété comme n'importe quel caractère et non comme un point (exemple du numéro de série *TCA020372V* dans le document 4088104749.txt).
En revanche, impossible de spécifier que le séparateur doit être le même entre plusieurs éléments d'une date : `[0-9]{2}\$sep[0-9]{2}\$3[0-9]{2,4}` renvoie comme message d'erreur *Use of uninitialized value in concatenation (.) or string*.

24 novembre 2008. Les listes utilisées ont généralement été nettoyées des mots ambigüs (càd ceux également présents dans le dictionnaire de noms communs) ; en conséquence, de véritables noms de villes (Rennes), prénoms (Sylvain) peuvent ainsi avoir été extraits de ces listes car ambigüs. En tenir compte lors de la compréhension des erreurs.
Le package `use encoding 'utf8'` permet d'indiquer à Perl que les expressions régulières contenues dans le code doivent être interprétées en UTF-8 ; il faut combiner les use open pour que les entrées/sorties soient encodées en UTF-8.

8 décembre 2008. Résolution du problème lié au mois d'août dans les expressions régulières permettant l'anonymisation des dates : lors de la récupération de la liste des mois depuis le fichier de configuration, substitution de le forme *aoÃ»t* par la forme *août* (pour rappel, le code est enregistré en UTF-8). Les packages `encoding 'utf8'`, `open ':utf8'`, `open ':std'` et `Encode 'decode_utf8'` sont restés en commentaires). Avant : *a été hospitalisé du 11 au 12 août 2004 pour complément*, après : *a été hospitalisé du <date /> pour complément*
Modification réinitialisation de la variable `@tableau=()` au lieu de `@variable=""` : évite d'avoir un enregistrement vide en début de tableau et le remplacement des espaces par une balise <hopital />

10 décembre 2008. Dans le fichier de configuration, production de deux listes de déclencheurs pour les hôpitaux : une liste longue ("Centre hospitalier") et une liste courte ("Centre") pour éviter que les déclencheurs courts prennent le dessus sur les déclencheurs longs.

12 janvier 2009. Lors de la réécriture du fichier anonymisé, supprime les espaces autour des tirets (*un traitement associant TENORMINE - ALDACTAZINE - ASPEGIC - LODALES - LEVOTHYROX.* devient *un traitement associant TENORMINE-ALDACTAZINE-ASPEGIC-LODALES-LEVOTHYROX.* dans le document 4088107098_ano.txt). Cette modification pose problème si un alignement de corpus est effectué (entre référence et résultat anonymisé) pour évaluer la qualité des résultats. Lignes commentées.

19 janvier 2009 (version 1c). Les tableaux de stockage des données ont été remplacés par des tables de hachage (`@tableau` devient `%tableau`, `@noms` devient `%noms`, etc). L'anonymisation des données par comparaison avec les références contenues dans ces tableaux s'en trouve beaucoup plus rapide (on passe de 11min 40 à seulement 3 secondes pour traiter 23 fichiers !).

11 février 2009. Améliorations ponctuelles diverses : complétion de la liste des déclencheurs de noms (Madame, Mademoiselle, Monsieur), intégration d'éléments supplémentaires lors de la seconde anonymisation, etc.

13 février 2009. Suppression de la vérification de la présence des mots dans la liste noire lorsque ces mots sont précédés d'un déclencheur (Pr, Dr, etc.) ; permet d'anonymiser *Pr Weber* alors que *Weber* figure dans la liste noire.
Ajout dans la trace du nom du fichier anonymisé pour chaque info.

21 juillet 2009. Petites retouches sur les dates qui ne se terminent pas par un séparateur mais par une fin de ligne.
La ville *Marseille* n'est pas anonymisée : ajout dans la liste des villes mais anonymisation toujours pas réalisée (pas de concordance avec la table des villes). Résolu le 23 juillet : la liste utilisée est lst_villes_sur (*Marseille* ajoutée dans cette liste).

28 juillet 2009. Lors de la récupération des noms de médicaments depuis la liste, on enregistre également la version désaccentuée du médicament (on utilise le code hexadécimal de chaque accent pour réaliser la désaccentuation : `tr/\xE8\xE9/ee/` par exemple).
Lors du test mot à mot des noms de médicaments, on teste également le mot mis en minuscules avec initiale en capitale.
Ces deux améliorations permettent de traiter efficacement le document 4088107098 dans lequel figurent des noms de médicaments en majuscules désaccentués : *ALDACTAZINE* testé sous la forme *Aldactazine* est anonymisé, de même que *ASPEGIC* testé comme *Aspegic* est trouvé comme tel dans la table de hachage des médicaments après avoir enregistré *Aspégic* sous la forme *Aspegic*.

31 juillet 2009. On crée une bijection sur les noms de médicaments composés (uniquement ceux intégrant une espace) de manière à appliquer cette bijection sur les lignes. Permet de traiter *Di Antalvic, Insuline NPH*, etc.

16 janvier 2010. Les tableaux de médicaments, prénoms et hôpitaux sont triés par tailles décroissantes des noms (même principe que dans Cokaine (COrpus and Knowledge-bAsed INformation Extraction), outil d'extraction des prescriptions médicamenteuses développé pour i2b2 2009, Deléger, Grouin, Zweigenbaum).

26 janvier 2010. Bonne gestion des passages d'arguments dans les routines (plus aucun message d'erreur).

29 janvier 2010 (version 1d). Dans le second passage, un mot commençant par une capitale suivant une balise <nom /> est remplacé par une balise <prenom /> uniquement si ce mot n'est, ni "prénom", ni "docteur" (permet d'éviter les cas : *Nom : <nom /> Prénom : <prenom />* qui devient *Nom : <nom /> <prenom /> : <prenom />* et *C. DUPONT Docteur F. DURAND* qui devient *<nom /> Docteur <nom />*).
Création de tableaux de bijection sur chaque liste de délencheurs (permet de trier par taille décroissante chaque élément).
Possibilité de trouver les expressions régulières listées en fin de ligne.

8 février 2010. Bloque sur certains fichiers, a priori en raison des parenthèses qui sont mal interprétées dans les expressions régulières.

5 mars 2010. Extension des fichiers de sortie changée de XML en SGML car pas vrai XML. Oui mais bof...

29 avril 2010. En seconde anonymisation, prise en compte des mots commençant par une capitale précédent une balise <nom /> ou <prenom />. Ces mots sont anonymisés uniquement si ils ne sont pas présents dans la liste des déclencheurs de noms.

1 septembre 2010. À partir du corpus clef en stomatologie, ajout de nouvelles entités à anonymiser (grades, numéros de dossier/acte médical) et ajout de déclencheurs supplémentaires pour les noms (métier : anesthésiste, opérateur, aide).
Les fichiers anonymisés ont pour extension « *.med » comme Medina tandis que « *.ano » est réservée comme extension de sortie de la chaîne par apprentissage (Wapiti ou CRF++).

20 décembre 2010. Modification des boucles `if` en `while` avec option "g" pour anonymiser tous les éléments pour une ligne et pas seulement le premier.
Amélioration des patrons et ajout de nouvelles règles. Permet de traiter presque tous les noms et prénoms du corpus.
Le traitement mot à mot est remonté dans la hiérarchie des opérations.

23 décembre 2010. Ajout anonymisation complémentaire sur la base de ce qui a déjà été anonymisé. permet de traiter les entités absentes des listes mais déjà traitées par des règles ou des déclencheurs (prénom *Nenci*).

10 janvier 2011. Le patron `$ligne=~/mod.le ([^\)]+)?/` pour les informations de pacemaker est trop large et anonymise des portions entières (je le commente) : *un modèle double chambre (qui permettra une stimulation de l'oreillette en cas de bradycardie sinusale liée à la majoration du traitement β-bloquant).* est anonymisé en *un modèle <info />).*

19 janvier 2011. Adaptation du programme au corpus d'anatomopathologie :
- En seconde anonymisation, on vérifie que le mot trouvé ne figure pas dans le lexique avant de le considérer comme un nom ou un prénom ;
- Le code postal doit obligatoirement être suivi par une espace et des caractères ; on ne peut pas le rencontrer en fin de ligne ou suivi par des étoiles (numéro de dossier à 5 chiffres) ;
- Les indices de grades (interne, externe) doivent commencer par une capitale pour éviter les anonymisations des adjectifs : *face, interne*. Problème également présent en stomatologie.

23 février 2012 (version 1e) et 24/02/12 (version 1f). Adaptation du script aux expériences pour le papier AMIA2012. *Les informations anonymisées sont désormais encadrées des balises typantes et non plus remplacées* comme auparavant. La précédente version remplaçant les informations par des balises, une évaluation au moyen du script de scoring nécessitait un réalignement (7e) entre le fichier « *.nom » d'origine et le fichier « *.med » anonymisé, produisant un fichier « *.enc » à évaluer. Des problèmes d'alignement ont conduit à suspendre cet alignement.

25 février 2012 (version 1g). Adaptation du script aux guidelines d'anonymisation définis pour le papier AMIA2012 et réintroduction de la seconde anonymisation.

27 février 2012 (version 1h). Ajout d'une fonction `etudePortion()` qui étudie le contexte dans lequel se trouve un patron passé en argument. Si le patron figure dans une portion déjà annotée, la fonction renvoie 1, sinon 0. Permet d'éviter d'annoter des entités à l'intérieur de portion déjà annotée ; par exemple, un prénom dans une adresse.

28 février 2012 (version 1i). Une seule règle pour les stimulateurs cardiaques : une marque de stimulateur suivie de un à cinq mots commençant par une capitale ou un chiffre et absent du dictionnaire de mots communs.
Ajout des déclencheurs *l'hôpital* et *l'Hôpital* dans le fichier de configuration (l'article est intégré à la portion annotée).

29 janvier 2012 (version 1j). Le corpus de test a été entièrement revu (1h de travail), des entités ayant été oubliées. Idem pour le corpus d'apprentissage. Modifications mineures.

12 mai 2012. Création d'un script de post-traitement : le script 2_antidatation.pl permet d'antidater toutes les dates d'un document d'un nombre de jours aléatoirement tiré entre 365 et 1460 (soit entre 1 et 4 ans). Garantit une anonymisation et une conservation des écarts temporels entre deux dates d'un document.

16 mai 2012 (version 1k). Meilleure prise en compte des dates (après application du script d'antidatation qui a révélé des erreurs dans le balisage des dates).

2 juin 2012. Création de deux scripts de post-traitements : le script 4_pseudonymes.pl remplace les occurrences de noms et prénoms par des pseudonymes, le script 5_hyperonymes.pl remplace toutes les données balisées par un hyperonyme (aucun remplacement si appliqué sur des fichiers *.dat ou *.pse).

Références

[Grouin *et al.*, 2009a] GROUIN, C., ROSIER, A., DAMERON, O. et ZWEIGENBAUM, P. (2009a). Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. *In* FIESCHI, M., STACCINI, P., BOUHADDOU, O. et LOVIS, C., éditeurs : *Risques, technologies de l'information pour les pratiques médicales*, volume XVII de *Informatique et santé*. Springler-Verlag, France.

[Grouin *et al.*, 2009b] GROUIN, C., ROSIER, A., DAMERON, O. et ZWEIGENBAUM, P. (2009b). Testing tactis to localize de-identification. *In Stud Health Technol Inform*, volume 150, pages 735–739.

[Grouin et Zweigenbaum, 2011] GROUIN, C. et ZWEIGENBAUM, P. (2011). Une approche à plusieurs étapes pour anonymiser des documents médicaux. *RSTI–RIA, Intelligence Artificielle et santé*, 25(4):525–549.