REPUBLIQUE FRANCAISE

Ministère de l'Enseignement Supérieur et de la Recherche

Université Paris 5 René Descartes

Mémoire présenté par Yasmina LALEG pour obtenir :

Le Diplôme de MASTER SCIENCES DE LA VIE ET DE LA SANTE
à finalité RECHERCHE mention MATHEMATIQUE ET INFORMATIQUE
spécialité INFORMATIQUE BIOMEDICALE

Intitulé :

Modélisation et implémentation d'une architecture sécurisée d'accès à un entrepôt de données cliniques

Soutenu le 02 juillet 2009
au SPIM (Laboratoire de Santé Publique et d'Informatique Médicale)
devant le jury suivant :

- Pr Richard Beuscart, CERIM, Université Lille 1, Lille. (Rapporteur)
- Pr Alain Venot , LIM&BIO, Université Paris 13. (Rapporteur)
- Eric Zapletal, HEGP, (Encadreur)
- Natalia.Grabar SPIM, HEGP (Encadreur)
- Pr Patrice Degoulet, SPIM,Université Paris 6
- Mme Marie-Christine Jaulent, SPIM, Paris
- Pr Pierre Zweigenbaum, LIMSI-CNRS, Paris

Sommaire

Abstract

Keywords

1. Introduction
    - Contexte
    - Objectif
2. Etat de l'art
    2.1. Législation en vigueur
    2.2. Travaux de repérage pour les données identificatrices dans les documents non structurés
    2.3. les travaux de chiffrements existants pour les données structurées
        2.3.1. Notions générale de chiffrement (cryptage)
            - Cryptage symétrique
            - Cryptage asymétrique
            - Cryptage hybride
3. Matériels
    3.1. Les lois en vigueur : Ethique et confidentialité des données
    3.2. Base DxCare et données cliniques (comptes rendus)
    3.3. L'entrepôt I2B2
    3.4. Outils Talend studio (extraction, transformation, chargement)
4. Méthode
    4.1. Modélisation UML
        4.1.1. Identification des cas d'utilisation
        4.1.2. Cryptage des données structurées
        4.1.3. Repérage des informations identificatrices dans les documents non structurés
5. Résultats et discussion
    5.1. Modèles des cas d'utilisation
        5.1.1. Identification des acteurs (utilisateurs)
        5.1.2. Identification des cas d'utilisation de chaque acteur
        5.1.3. Diagrammes de séquences

5.2. Cryptage des données structurées

5.3. Repérage des informations nominatives dans les documents non structurés

5.4. Architecture sécurisée d'accès

6 Conclusion et perspectives

Bibliographie

Annexes

Annexe 1 : Synthèse des lois (française et européenne) concernant le traitement automatisé des données nominatives

Annexe 2 : Algorithmes de chiffrement

Annexe 3 : Langage UML et modélisation multidimensionnelle

Annexe 4 : Les scénarios du processus d'anonymisation

Annexe 5 : Modélisation des contraintes applicables à l'entrepôt

Annexe 6 : Tableau des acteurs et leur cas d'utilisation

Modélisation et implémentation d'une architecture sécurisée d'accès à un entrepôt de données cliniques

Yasmina LALEG

INSERM UMR S 872, Université Paris 5 René-Descartes, France

Abstract

Objective: *Design a secure architecture for accessing a data warehouse with clinical data, which nominative elements are anonymized and encrypted in structured and unstructured documents while respecting the law in force.*

Material & Method: *We perform a UML modelization of use cases within the data warehouse and associate them with legal constraints. We then propose a reversible anonymization and encryption of the data and documents. We use for this natural language processing methods and encryption methods. The implementation is performed within an ETL module connecting the hospital information system with the data warehouse.*

Results and discussion: *We achieved the modelization of use cases of the clinical data warehouse and integrated legal constraints. The nominative information within unstructured documents is being de-identified with DEID software, which has been adapted to French documents and which performs now a reversible de-identification of patient personal information. All data are encrypted through a secure hybrid approach for their storage and transmission. Only accredited users can have access to nominative data, otherwise it remains encrypted. Further work will allow to improve the de-identification of nominative information and to take into account the European legislation.*

Keywords

De-identification, Security, Encryption, Patient identification, CNIL, Natural Language Processing, Health Information.