



Estudi d'anonimització de dades mèdiques utilitzant ontologies

AUTOR: Xavier Salvadó Martí

DIRECTORS: Dra. Aïda Valls Mateu, Dr. Sergio Martínez Lluís

Adreça electrònica: xavier.salvadam@estudiants.urv.cat

Titulació: Enginyeria Tècnica Informàtica de Gestió

Data Presentació: Juny 2013

Resum

La publicació de dades mèdiques és una font molt valuosa per realitzar recerca mèdica i d'aquesta forma poder millorar el tractament dels pacients, però les dades mèdiques que poden identificar a un pacient no poden ser utilitzades sense el seu consentiment previ. En canvi, segons l'Acta de Protecció de Dades de l'any 1998 d'Espanya, es poden utilitzar aquestes dades en recerca mèdica si no identifiquen unívocament els pacients.

Per protegir la privadesa de les dades mèdiques, s'utilitzen mètodes d'anonimització que generen una nova versió de la base de dades per assegurar la privadesa dels pacients minimitzant la pèrdua d'informació per assegurar la seva utilitat.

Aquest PFC forma part de les tasques del projecte ARES-CONSOLIDER. L'objectiu d'aquest PFC és desenvolupar una aplicació anomenada SAM (Semantic Anonymization Methods), que permeti anonimitzar un fitxer de dades mèdiques codificat en l'ontologia mèdica Snomed-CT.

Primer de tot, degut a que les dades mèdiques de test poden estar codificades en CIE-9-CM, s'ha realitzat un aplicatiu de pre-processament de dades mèdiques per re-codificar dades en CIE-9-CM segons la nomenclatura emprada a Snomed-CT.

A continuació, per realitzar l'aplicació SAM, s'ha partit d'unes quantes funcionalitats implementades pel grup de recerca ITAKA que permeten anonimitzar i avaluar fitxers de dades codificats en una altra ontologia anomenada WordNet. Concretament, la feina que s'ha realitzat en aquest PFC és la següent: s'ha desenvolupat una API per poder realitzar el procés d'anonimització amb dades codificades en l'ontologia SNOMED-CT, s'ha creat una interfície gràfica per l'aplicació, s'han modificat els processos d'anonimitzar i avaluar per a que s'executin com a fils, s'ha adaptat l'aplicatiu per a nous formats dels fitxers i s'ha creat un nou fitxer de sortida que conté estadístiques sobre el procés d'anonimització.

Els programes desenvolupats s'han provat amb dades mèdiques reals proporcionades per l'Oficina de Planificació i Desenvolupament de la Salut de Califòrnia.

Resumen

La publicación de datos médicos es una fuente muy valiosa para realizar investigación médica y de esta forma poder mejorar el tratamiento de los pacientes, pero los datos médicos que pueden identificar a un paciente no pueden ser utilizados sin su consentimiento previo. En cambio, según el Acta de Protección de Datos de 1998 de España, se pueden utilizar estos datos en investigación médica si no identifican unívocamente los pacientes.

Para proteger la privacidad de los datos médicos, se utilizan métodos de anonimización que generan una nueva versión de la base de datos para asegurar la privacidad de los pacientes minimizando la pérdida de información para asegurar su utilidad.

Este PFC forma parte de las tareas del proyecto ARES-CONSOLIDER. El objetivo de este PFC es desarrollar una aplicación llamada SAM (Semantic Anonymization Methods), que permita anonimizar un fichero de datos médicos codificados en la ontología médica SNOMED-CT.

Primero de todo, debido a que los datos médicos de prueba pueden estar codificados en CIE-9-CM, se ha realizado un aplicativo de pre-procesamiento de datos médicos para re-codificar datos en CIE-9-CM según la nomenclatura utilizada en SNOMED-CT.

A continuación, para realizar la aplicación SAM, se ha partido de unas cuantas funcionalidades implementadas por el grupo de investigación ITAKA que permiten anonimizar y evaluar ficheros de datos codificados en otra ontología llamada WordNet. Concretamente, el trabajo que se ha realizado en este PFC es el siguiente: se ha desarrollado una API para poder realizar el proceso de anonimización con datos codificados en la ontología SNOMED-CT, se ha creado una interfaz gráfica para la aplicación, se han modificado los procesos de anonimizar y evaluar para que se ejecuten como hilos, se ha adaptado el aplicativo para nuevos formatos de los ficheros y se ha creado un nuevo archivo de salida que contiene estadísticas sobre el proceso de anonimización.

Los programas desarrollados se han probado con datos médicos reales proporcionados por la Oficina de Planificación y Desarrollo de la Salud de California.

Summary

The publication of medical data is a valuable resource to perform medical research and thus to improve the treatment of patients, however, the medical data that could identify a patient cannot be used without his prior consent. Although, according to the Data Protection Act 1998 in Spain, these data can be used in medical research if it does not identify patients.

To protect the privacy of medical data, anonymization methods are used to generate a new version of the database in order to ensure the privacy of patients while minimizing the information loss and ensuring its utility.

This PFC is part of project tasks ARES-CONSOLIDER. The objective of this PFC consist on develop an application called SAM (Semantic Anonymization Methods), which allows the anonymization of medical data files according the nomenclature of the medical ontology SNOMED-CT.

First, because of the medical test data can be encoded in ICD-9-CM, an application for pre-processing medical data to re-encode data according to nomenclature used in SNOMED-CT has been made.

Then, to make the application SAM, we started from several functionalities implemented by the research group Itaka which allow anonymizing and evaluating data files encoded in an ontology called WordNet. Specifically, the work which has been done in this PFC is: an API has been developed to perform the anonymization process with data encoded in SNOMED-CT ontology, a graphical user interface has been created for the application, the processes to anonymize and evaluate have been modified in order to run as threads, the application has been adapted for the new file formats and a new output file containing statistics about anonymization process has been created.

The programs developed have been tested with real medical data provided by the California Office of Health Planning and Development.