André Silva Pedrosa

Desenvolvimento de uma arquitetura escalável para extrair metainformação de bases de dados médicas distribuídas

Development of a scalable architecture to extract metadata from distributed medical databases

André Silva Pedrosa

Desenvolvimento de uma arquitetura escalável para extrair metainformação de bases de dados médicas distribuídas

Development of a scalable architecture to extract metadata from distributed medical databases

"*You can't have happiness without pain, you need a little bit of rain to have a little bit of rainbow.*"

— Felix Lengyel

Universidade de Aveiro
2021

André Silva Pedrosa

Desenvolvimento de uma arquitetura escalável para extrair metainformação de bases de dados médicas distribuídas

Development of a scalable architecture to extract metadata from distributed medical databases

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Informática, realizada sob a orientação científica do Doutor José Luís Guimarães Oliveira, Professor catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

o júri / the jury

presidente / president

Prof. Doutor Joaquim Arnaldo Carvalho Martins
Professor Catedrático, Universidade de Aveiro

vogais / examiners committee

Prof. Doutor Paula Alexandra Gomes da Silva
Professora Auxiliar, Universidade de Coimbra - Faculdade de Ciências e Tecnologia

Prof. Doutor José Luis Guimarães Oliveira
Professor Catedrático, Universidade de Aveiro

agradecimentos / acknowledgements

Começo por agradecer a todas as pessoas que conheci por meio da minha estadia na residência. Dos veteranos aos caloiros um grande obrigado por me darem uma razão de continuar a trabalhar e por criarem um ambiente de família.

Aos meu orientador, o professor José Luís Oliveira, e coorientador, João Almeida, por acreditarem no meu trabalho e não desistirem de mim, para além das várias vezes que possa não ter atingido as suas expectativas.

Agradeço aos vários amigos e amigos que fui conhecendo ao longo do meu percurso académico da Universidade de Aveiro. Em especial, obrigado aos fundadores do grupo SASA LELE, Filipe Pires e João Alegria, por suportarem o meu feitio e por estarem ao meu lado durante o nosso curso.

À malta de Leiria que para além de me ter descolado para um local distante, continuarem a receber-me como um amigo chegado. Agradeço especialmente ao Leandro pela paciência e preocupação.

Por último, mas não menos importante, queria agradecer à minha família por estarem lá nos meus altos e baixos e por me darem força para eu seguir em frente.

Palavras Chave

meta-dados, registo eletrónico de pacientes, EHDEN, OHDSI, atualização, automatização, extração

Resumo

Para realizar estudos médicos, tais como o impacto de um fármaco numa determinada população, os investigadores precisam de acesso a dados reais de organizações médicas, tais como hospitais, para que os seus estudos tenham resultados fiáveis. No entanto, para os investigadores, este é um processo difícil e demorado até conseguirem encontrar o conjunto de dados que melhor se adapta ao seu estudo, dado que muitas vezes não têm acesso direto aos dados reais. Para ajudar neste processo, existem plataformas centralizadas que descrevem o conteúdo dos conjuntos de dados através de meta-dados. Atualmente, em muitas destas plataformas, estes meta-dados são atualizados manualmente, um processo lento e tedioso, o que leva a que se tornem muito facilmente desatualizados.

O objetivo deste trabalho é desenvolver um sistema que extrai automaticamente meta-dados diretamente dos dados reais das bases de dados e os envia para várias plataformas que estão a expor meta-dados de modo a que estes se mantenham atualizados. O sistema pretende trabalhar em bases de dados ligadas ao grupo EHDEN (European Health Data and Evidence Network), que pretende criar uma rede federada de bases de dados médicas na Europa, seguindo os mesmos princípios e ferramentas utilizados pela comunidade OHDSI (Observational Health Data Sciences and Informatics). A extração de meta-dados é feita por agentes que são instalados juntamente com o sistema de produção das bases de dados. Os meta-dados são então enviados para um sistema de gestão de meta-dados, construído sobre uma framework de fluxo de dados chamada Kafka. Este sistema de gestão é composto por um conjunto de componentes que foram desenvolvidos de acordo com uma filosofia de microsserviços, tendo sempre em mente a possibilidade de escalar cada um deles.

Keywords metadata, electronic health records, EHDEN, OHDSI, update, automation, extraction

Abstract

To conduct medical studies, such as the impact of a drug on a certain population, researchers need access to real data from medical organizations such as hospitals for their studies to have reliable results. However, for researchers, this is a difficult and time-consuming process until they can find the dataset that best fits their study, as they often do not have direct access to the actual data. To help in this process, there are centralized platforms that describe the content of datasets through metadata. Currently, in many of these platforms, this metadata is updated manually, a slow and tedious process, which leads to it becoming outdated very easily.

The goal of this work is to develop a system that automatically extracts metadata directly from databases' actual data and sends it to various platforms that are exposing metadata so that they are kept up to date. The system intends to work on databases linked to the EHDEN (European Health Data and Evidence Network) group, which intends to create a federated network of medical databases in Europe, following the same principles and tools used by the OHDSI (Observational Health Data Sciences and Informatics) community. Metadata extraction is done by agents that are installed along with the production system of the databases. The metadata is then sent to a metadata management system, built on top of a data streaming framework called Kafka. This management system is composed of a set of components that were developed according to a philosophy of microservices, always keeping in mind the possibility of scaling each one.