MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-1266

Generation of *open*EHR Test Datasets for Benchmarking

Samar El Helou[a], Tuukka Karvonen[a], Goshiro Yamamoto[b], Naoto Kume[c], Shinji Kobayashi[c], Eiji Kondo[d], Shusuke Hiragi[b], Kazuya Okamoto[b], Hiroshi Tamura[b], Tomohiro Kuroda[b]

a Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan
b Division of Medical Information Technology & Administration Planning, Kyoto University Hospital, Kyoto, Japan
c Department of Electronic Health Records, Graduate School of Medicine, Kyoto University, Kyoto, Japan
d Department of Gynecology and Obstetrics, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Abstract

openEHR is a widely used EHR specification. Given its technology-independent nature, different approaches for implementing openEHR data repositories exist. Public openEHR datasets are needed to conduct benchmark analyses over different implementations. To address their current unavailability, we propose a method for generating openEHR test datasets that can be publicly shared and used.

Keywords:

Electronic Health Record; Benchmarking

Introduction

OpenEHR is a technology-independent, open-source specification for Electronic Health Records' (EHR) architecture adopting a two-level modeling approach [1]. The choices of database technologies and approaches for repository implementations are left for the developers. Benchmarking performance of different repository implementations in different use-case scenarios is needed [2]. Usually, benchmark analyses include the comparison of query response times and thus require access to shared openEHR datasets, often unavailable due to strict medical privacy laws.

The effectiveness of a benchmarking dataset is affected by its level of accessibility, realism and evaluation capabilities. In the case of openEHR benchmarking datasets, the structure of the data is constrained by the reference model and archetypes' definitions, thus the evaluation capabilities can be defined and artificially simulated. As for realism, some could be potentially sacrificed in favor of accessibility in cases where real data is difficult to come by.

This work provides a method to generate open application-specific openEHR test datasets. The resultant datasets should comply with openEHR's information and archetype models and allow queries applicable in real world scenarios.

Methods

First, we identified the clinical concepts involved in a pregnancy home-monitoring application and determined a list of realistic data entries. Next, we mapped the clinical concepts to openEHR archetypes available in the openEHR Clinical Knowledge Manager (CKM) and created data value sets corresponding to the possible data entries. We applied an Object Relational Mapping (ORM) approach to design a relational schema allowing the persistence of the required archetypes over classes from the openEHR Reference Model. We created data generation plans using the archetypes' structure and data value sets, as shown in Figure 1. The plans were executed using Microsoft Visual Studio 2010 to populate an SQL Server database. Finally, we identified application-specific search scenarios for which we formulated and executed SQL queries solely using the archetypes' definitions.

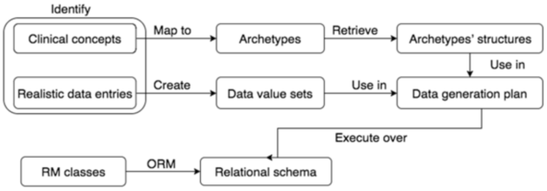

Figure 1 - openEHR dataset generation process

Results

The method was applied to generate datasets simulating a pregnancy home-monitoring repository. A set of seven queries were executed and generated non-empty result sets. Datasets of 10k and 100k records in CSV and JSON formats can be accessed via github.com/samarhelou/data. Cypher queries are also provided to allow dataset import, visualization, and testing in Neo4j, a labeled property graph database.

Conclusions

We proposed and tested a method for generating test openEHR datasets. Future work requires the inclusion of data generation rules to reflect the real distributions of medical cases in the population. The generated datasets will be used to benchmark an openEHR repository implementation using Neo4j.

References

[1] D. Kalra, T. Beale, and S. Heard, The *open*EHR foundation, Studies in health technology and informatics **115** (2005), 153-173.

[2] S. Frade, S.M. Freire, E. Sundvall, J.H. Patriarca-Almeida, and R. Cruz-Correia, Survey of *open*EHR storage implementations, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, *IEEE* (2013), 303-307.

Address for correspondence

Samar El Helou, email: samar@kuhp.kyoto-u.ac.jp