Development of an ETL Process for Bulk and Incremental Load of German Patient Data into OMOP CDM Using FHIR

Elisa Henke, Yuan Peng, Ines Reinecke, Michele Zoch, Martin Sedlmayr

Background

Within the Medical Informatics Initiative Germany (MI-I), the MIRACUM (Medical Informatics in Research and Care in University Medicine) focuses on its “Use Case 1: Alerting in Care – IT Support for Patient Recruitment”[1]. The aim of the use case is to develop a Clinical Trials Recruitment Support System (CTRSS), which suggests patients for clinical trials based on data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). For this purpose, we have developed an ETL (Extract-Transform-Load) process. The ETL process transforms data into OMOP CDM using Fast Healthcare Interoperability Resources (FHIR) as data source. To enable quick recruitment, it is necessary for the data to be up-to-date and analyzed quickly. This results in the requirement that the observational patient data must be loaded and uploaded not only once, but almost near real time or at least once a day. Therefore, an ETL process for bulk and incremental load was developed.

Methods

We mainly based our ETL process on the specifications of the FHIR profiles of the core data set of MI-I[2]. Sometimes it was necessary to use MIRACUM FHIR profiles[3] instead, because some MI-I FHIR profiles are still under construction or not completely suitable for the usage in MIRACUM. For example, the patient FHIR profile from MI-I is not anonymized. However, anonymized patient data should be used in MIRACUM. The following FHIR profiles are used to transform German patient data into OMOP CDM: 1) MI-I FHIR profiles: Medication, MedicationAdministration, Procedure, Observation; 2) MIRACUM FHIR profiles: Patient, Encounter and Condition. All FHIR resources are stored in a FHIR gateway, which is a PostgreSQL database. The gateway provides a way for all MIRACUM partner sites to have similar FHIR resources available so that the ETL processes (both input from core data set to FHIR gateway and FHIR gateway to OMOP) could be used quickly and easily.

The target of the ETL process are the tables in OMOP CDM v5.3.1[4]. The following tables can be loaded with the ETL process: PERSON, LOCATION, DEATH, VISIT_OCCURRENCE, OBSERVATION_PERIOD, VISIT_DETAIL, CONDITION_OCCURRENCE, FACT_RELATIONSHIP, PROCEDURE_OCCURRENCE, OBSERVATION, MEASUREMENT, DRUG_EXPOSURE. The basic semantic mapping from FHIR profiles to tables in OMOP CDM is included in predefined mapping tables.

The ETL process from FHIR to OMOP CDM is implemented with Java 11. Moreover, the architecture of the ETL process is based on the open source framework Java Spring Batch 2.4.3[5]. The ETL process consists of the three processing units: Reader (Extract), Processor (Transform) and Writer (Load). Furthermore, we used two different loading methods for the ETL process: bulk load and incremental load. The bulk load is intended for an initial loading of a target with all data from a source. In contrast, incremental loading only considers those data from a source, which have been newly added or have changed since the last time the ETL job was executed.

Results

Figure 1 shows the architecture of the developed ETL process from FHIR to OMOP CDM. Before the job is executed, a switch can be set to select whether the ETL job is executed as bulk load e.g. for an initial load or as incremental load e.g. for daily updates. Next, the reader reads the FHIR resources from FHIR gateway. This is followed by the processor. For each FHIR resource type there is a processor. Each processor is associated with a mapper that contains the business logic to map the elements in FHIR resources to OMOP CDM tables. When reading and processing FHIR resources, there is a significant difference between bulk load and incremental load. During bulk load, all FHIR resources are read in a certain order depending on their resource type to ensure referencing between FHIR resources. On the other hand, incremental loading does only load new or updated FHIR resources and in this context does not consider any specific order of FHIR resources. These aspects lead to major implementation challenges such as checking for data to be updated in OMOP CDM or creating dummy datasets in OMOP CDM for referenced FHIR resources that have not yet been processed. Finally, the writer writes all transformed data to OMOP CDM.

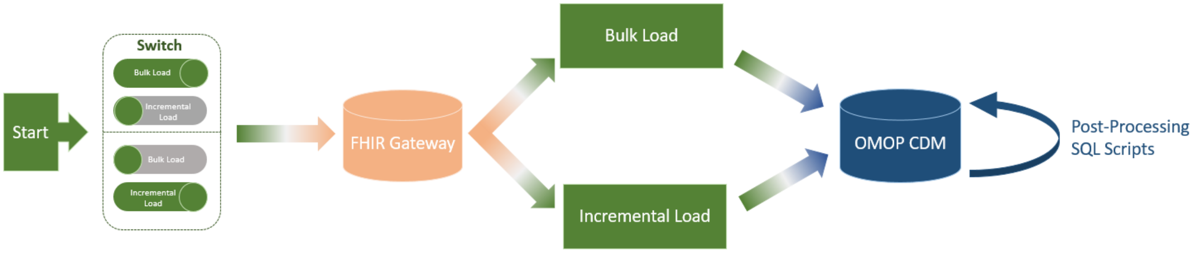

Figure 1: Architecture of the developed ETL process from FHIR to OMOP CDM (own figure)

After all FHIR resources are written to OMOP CDM, post processing takes place. Post processing uses SQL scripts to perform complex mappings at database-level.

Conclusion

The developed ETL process can be executed for transforming data from the core data set of MI-I and MIRACUM in FHIR format to OMOP CDM as bulk load or incremental load. Thus, patient data can be updated to enable rapid recruitment with the CTRSS based on OMOP.

In the future, the ETL process is continuously adapted: Firstly, we will connect the ETL process with a FHIR server and use meta data from FHIR resources and OMOP CDM to automate the ETL process. Secondly, we will adjust this ETL process to fit for the new profiles of the core data set of MI-I once they have been published, since some FHIR profiles are still under construction. Therefore, the ETL process can also be used by every partner side of the MI-I, so all German university hospitals will be enabled to load patient data into OMOP CDM.

References/Citations

1. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a Modular Clinical Trial Recruitment Support System Based on FHIR and OMOP. Stud Health Technol Inform. 2020 16;270:158-162. doi: 10.3233/SHTI200142.
2. Medical Informatics Initiative Germany. Basismodule des Kerndatensatzes der MII [Internet]. [cited 16 June 2021]. Available from: https://www.medizininformatik-initiative.de/de/basismodule-des-kerndatensatzes-der-mii.
3. MIRACUM. MIRACUM Core Implementation Guide – Table of Contents [Internet]. 2020 [cited 16 June 2021]. Available from: https://fhir.miracum.org/core/toc.html.
4. Observational Health Data Sciences and Informatics. OMOP CDM v5.3.1 [Internet]. [cited 16 June 2021]. Available from: https://ohdsi.github.io/CommonDataModel/cdm531.html.
5. Ward L, Syer D, Risberg T, Kasanicky R, Garrette D, Lund W, Minella M, Schaefer C, Hillert G, Renfro G, Bryant J, Hassine M B. Spring Batch – Reference Documentation [Internet]. 2021 [cited 16 June 2021]. Available from: https://docs.spring.io/spring-batch/docs/current/reference/html/index.html.