Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200138

Data Integration into OMOP CDM for Heterogeneous Clinical Data Collections via HL7 FHIR Bundles and XSLT

Patrick FISCHER [1, a], Mark R. STÖHR [b], Henning GALL [b],
Achim MICHEL-BACKOFEN [c] and Raphael W. MAJEED [b]
[a] Institute of Medical Informatics, Faculty of Medicine, Justus-Liebig-University, Giessen, Germany
[b] UGMLC, German Center for Lung Research (DZL), Justus-Liebig-University, Giessen, Germany
[c] Data-Integration-Center, Faculty of Medicine, Justus-Liebig-University, Giessen, Germany

Abstract. Data integration is an important task in medical informatics and highly impacts the gain out of existing health information data. These tasks are using implemented as extract transform and load processes. By introducing HL7 FHIR as an intermediate format, our aim was to integrate heterogeneous data from a German pulmonary hypertension registry into an OMOP Common Data Model. First, domain knowledge experts defined a common parameter set, which was subsequently mapped to standardized terminologies like LOINC or SNOMED-CT. Data was extracted as HL7 FHIR Bundle to be transformed to OMOP CDM by using XSLT. We successfully transformed the majority of data elements to the OMOP CDM in a feasible time.

Keywords. ETL, OMOP, HL7 FHIR, LOINC, SNOMED-CT

1. Introduction

International collaborative research faces a variety of challenges. One includes integrating data from different heterogeneous data sources into a data repository or data registry. Such registries can serve a lot of different purposes such as drug safety analysis [1], longitudinal analysis of certain cohorts [2] or as in our, and many other cases [3] a disease specific registry to promote research in that field. As part of the Pulmonary Vascular Research Institute, a collaboration of healthcare professionals and researchers working in the field of pulmonary hypertension (PH), an international PH registry/data repository needs to be established.

We chose the OHDSI OMOP Common Data Model [1] to be the foundation of this registry. The OMOP CDM is a database specification, based on standardized vocabularies like SNOMED-CT, LOINC or RxNorm, with the aim to provide an analytic-friendly non-complex representation of medical data [1].

[1] Corresponding Author, Patrick Fischer, E-mail: Patrick.Fischer@informatik.med.uni-giessen.de

To populate a registry with data from existing databases, usually, Extract-Transform-Load tools (ETL) are used to transform and transfer data to the target schema. Popular tools in this regard are Talend Open Studio or Pentaho, which are used to define the ETL process as a dataflow from source to target system. However, by using such software a complete ETL process needs to be implemented for each source separately. This is also the recommended way of designing ETL processes for the OMOP CDM. Because we want to connect multiple heterogeneous databases, a common parameter set was predefined. We plan to use HL7 FHIR [4] as an intermediate format for the local data extraction. The FHIR resource will be fed into an XSLT processor to generate CSV files in the OMOP schema. Similar international and national approaches using intermediate formats or HL7 FHIR together with OMOP CDM have led to good results [5–7].

Therefore, our aim is to evaluate the feasibility of HL7 FHIR Bundle and XSLT as a generic ETL process to populate an OMOP CDM. Feasibility will be assessed by the computation time and coverage of source data in the target CDM.

2. Method

2.1. Common data set

To integrate data from multiple registries into a single data warehouse, a common data set is needed. A meeting of domain experts from two large PH registries from both the UK and Germany was scheduled to discuss all data items and produce a common data set definition. In a second step, this data set is annotated with international standard terminology (LOINC, SNOMED-CT, etc.) by documentation experts.

2.2. Source data

As source data, we used a German PH registry database. Its data is stored in a Microsoft Access database with 22 proprietary tables and 18 user forms. Additionally, two EXCEL sheets are used to collect data about specimen and echocardiography. All tables are exported to CSV format using MS Access's standard export functionality. For ETL-preparation, we used HIStream-import to convert the proprietary CSV files to a standard FHIR Bundle collection. For this purpose, an XML configuration file is written which includes mapping of source field names to standard terminology as defined by the common data set. The resulting HL7 FHIR Bundle resource contains Patient, Encounter and Observation resources and conforms to the latest FHIR standard 4.1.0 [4]. An example of the Bundle structure containing the source data is depicted in Figure 1.

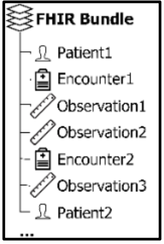

Figure 1. Example FHIR Bundle structure of provided input data

2.3. Mapping to OMOP domains

The OMOP CDM splits data into different domains depending on their scope. Condition, Drug, Procedure, Measurement are common medical domains, which are mapped to separate tables in the OMOP CDM. We used Athena [8] to browse the complete OMOP vocabularies to find the correct target table for the data elements of our defined data set. We did so by searching for the respective SNOMED-CT, LOINC or ATC code and retrieving the OMOP domain of the resulting concepts, as well as the concept ids.

2.4. Extending OMOP vocabularies

There are many subtypes of pulmonary hypertension depending on the etiology or comorbidities of the patient. Those subtypes are classified using the clinical classification of the World Symposium on Pulmonary Hypertension of 2013, which is also referred to as PH Nice Classification [9]. Since the Nice classification is not part of the standardized vocabularies already included to OMOP, we had to add it manually.

By using the hierarchical classification and the ER model of the OMOP standardized vocabularies [10], we generated CSV files to bulk load the Nice classification to OMOP.

2.5. Data transformation using XSLT and XPath

To transform the source data, provided as HL7 FHIR Bundle, to the OMOP schema we used XSLT (v2.0) [11] and XPath (v2.0) [12] to generate CSV for each used OMOP target table. The resulting CSV files contain all required information for the respective table, e.g. Measurement or Patient.

3. Results

3.1. Mapping to standard terminology

The predefined common dataset consists of 141 distinct data elements all represented by a specific code. Out of the 141 elements, 36 have been mapped to SNOMED-CT, 38 to LOINC, 27 to ATC and one to ICD-10. 27 of the remaining 39 data elements belong to the PH Nice classification, while no standard code was found for 12 parameters.

3.2. Algorithm

The transformation is split up into different files, one for each table that will be populated. Currently our approach populates the following OMOP tables with all required fields: *PERSON, OBSERVATION_PERIOD, CONDITION_OCCURRENCE, DRUG_EXPOSURE, PROCEDURE_OCCURRENCE, OBSERVATION and VISIT_OCCURRENCE*

The transformation structure is in principle the same for each table. A generic XSLT template is defined, to generate one row for each entry to the respective table. This generic template also extracts metadata about the entry like the date time or references to patients or encounters. Additionally, we defined parameter-specific templates, which map one parameter to its correct concept id in OMOP. Those parameter-specific templates call the generic template with the correct parameter specific data like concept id, status ids or type ids for the respective parameter.

The resulting CSV files are then bulk loaded to the database. A schematic illustration of the ETL process is shown in Figure 2.

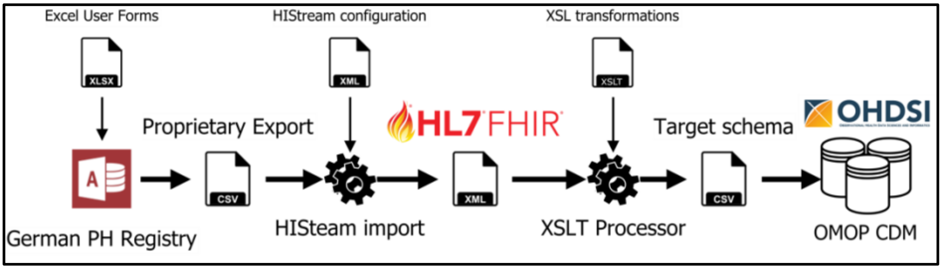

Figure 2. Schematic illustration of ETL process, showing involved software, including configurations and temporary result files

3.3. Transformation results and performance

Except for the 12 data elements of the common data set, where no standard concept in the standardized vocabularies could be found, every element was successfully mapped to the OMOP domains resulting in an overall coverage of about 91.5 %. We applied for new codes within LOINC for five data elements [13]. Additionally, we added 34 new concepts, 54 concept relationships and 68 concept ancestor entries to the vocabularies and therefore achieved 100 % coverage of the PH Nice classification in OMOP.

In total, data of 3,887 patients was transformed to OMOP in 15 minutes and 29 seconds using Saxon-HE v9.9.1 XSLT processor (on Intel Core i7-8550U with 256 GB SSD). To populate all required fields, we assigned default values for the parameters, which were not available in the source data. Those fields were for example: *visit_type, admitted_from, discharged_to* and others. In addition, fields like ethnicity and race have been set to default values for the first prototype, but those will be correctly mapped in the future as well.

4. Discussion

Based on the achieved results we conclude the feasibility for HL7 FHIR Bundle and XSLT as part of a generic ETL process. The majority of data elements were successfully transformed to the target scheme in reasonable time. Additionally, the customization of the available ETL process to connect new source databases reduces to only adjusting the data extraction process and HIStream configuration to provide data as HL7 FHIR Bundle. Therefore, providing a HL7 FHIR export for a local data source can increase interoperability of the included data, not only for data warehousing, but also for communication with other systems.

Limitations of this work include the necessary synchronization between HIStream configuration and XSL transformations in order to transform all available data elements to OMOP. Additionally, the mapping of standard concepts is currently hard coded, by leveraging the rich vocabularies provided as CSV the mapping can be designed in a more generic way. We also plan to reduce the computational time, even though 15 minutes for the transformation is already feasible because data will only be loaded periodically.

References

[1] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, and P.E. Stang, Validation of a common data model for active safety surveillance research, *J. Am. Med. Inform. Assoc.* **19** (2012) 54–60. doi:10.1136/amiajnl-2011-000376.

[2] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, and M.N. Zozus, Evaluating common data models for use with a longitudinal community registry, *J. Biomed. Inform.* **64** (2016) 333–341. doi:10.1016/j.jbi.2016.10.016.

[3] M. Coleman, D. Forman, H. Bryant, J. Butler, B. Rachet, C. Maringe, U. Nur, E. Tracey, M. Coory, J. Hatcher, C. McGahan, D. Turner, L. Marrett, M. Gjerstorff, T. Johannesen, J. Adolfsson, M. Lambe, G. Lawrence, D. Meechan, E. Morris, R. Middleton, J. Steward, and M. Richards, Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data, *The Lancet*. **377** (2011) 127–138. doi:10.1016/S0140-6736(10)62231-3.

[4] HL7 International, FHIR v4.1.0, *FHIR*. (2018). http://build.fhir.org/ (accessed September 30, 2019).

[5] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.U. Prokosch, and M. Sedlmayr, Towards Implementation of OMOP in a German University Hospital Consortium, *Appl. Clin. Inform.* **09** (2018) 054–061. doi:10.1055/s-0037-1617452.

[6] J. Guoqian, K. Richard, P. Eric, and S.H. R, Building Interoperable FHIR-Based Vocabulary Mapping Services: A Case Study of OHDSI Vocabularies and Mappings, *Stud. Health Technol. Inform.* (2017) 1327–1327. doi:10.3233/978-1-61499-830-3-1327.

[7] J. Guoqian, K.R. C, S.D. K, P. Eric, and S.H. R, A Consensus-Based Approach for Harmonizing the OHDSI Common Data Model with HL7 FHIR, *Stud. Health Technol. Inform.* (2017) 887–891. doi:10.3233/978-1-61499-830-3-887.

[8] OHDSI, Athena, (2019). http://athena.ohdsi.org/search-terms/terms (accessed September 25, 2019).

[9] G. Simonneau, M.A. Gatzoulis, I. Adatia, D. Celermajer, C. Denton, A. Ghofrani, M.A. Gomez Sanchez, R. Krishna Kumar, M. Landzberg, R.F. Machado, H. Olschewski, I.M. Robbins, and R. Souza, Updated Clinical Classification of Pulmonary Hypertension, *J. Am. Coll. Cardiol.* **62** (2013) D34–D41. doi:10.1016/j.jacc.2013.10.029.

[10] OHDSI, OHDSI/CommonDataModel, *GitHub*. (2019). https://github.com/OHDSI/CommonDataModel (accessed September 25, 2019).

[11] XSL Transformations (XSLT) Version 2.0, (n.d.). https://www.w3.org/TR/xslt20/ (accessed January 7, 2020).

[12] XML Path Language (XPath) 2.0 (Second Edition), (n.d.). https://www.w3.org/TR/xpath20/ (accessed January 7, 2020).

[13] LOINC, Submitting New Term Requests, *LOINC*. (2019). https://loinc.org/submissions/new-terms/ (accessed September 25, 2019).