dHealth 2023
B. Pfeifer et al. (Eds.)

doi:10.3233/SHTI230011

Transforming Documents of the Austrian Nationwide EHR System into the OMOP CDM

Raffael Lukas KORNTHEUER[a,1], Florian KATSCH[a,b] and Georg DUFTSCHMID[a]
[a] *Section of Medical Information Management, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria*
[b] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*

Abstract. The Austrian nationwide EHR system ELGA can contribute valuable data for research due to its high volume of data and broad population coverage. In order to be applicable in international research projects, transformation to a standardized, research-oriented data model such as the OMOP common data model is essential. In this paper we describe our experience with the corresponding transformation task. Using Python scripts, we implemented a prototypical process that extracts, transforms, maps, and loads fully structured sections of ELGA documents to an OMOP database.

Keywords. ELGA, HL7 CDA, OMOP CDM, electronic health records, ETL process

1. Introduction

The Austrian nationwide electronic health record (EHR) system ELGA [1] has been operational since 2015. Due to the fact that it is compulsory used by public care providers and about 97% of Austrian citizens participate in ELGA, it achieves high volumes of routine data with broad population coverage. In order to make use of ELGA data in international research projects, a transformation to a common data format is necessary. An example for such a format is the Observational Medical Outcomes Partnership (OMOP) common data model (CDM), which was developed by Observational Health Data Sciences and Informatics (OHDSI) [2]. The goal of the OMOP CDM is to standardize observational data and create a reproducible way to analyze the data and generate evidence of it.

In order to transform data from various sources to a common data format, an extract, transform and load (ETL) process is a commonly used method. The process consists of preparing the input data, transforming it to match the format of the target database and finally loading it into the target system [3].

In this paper we describe our prototypical implementation of an ETL process to integrate data from the currently normative ELGA document types (Physician Discharge

[1] Corresponding Author: Raffael Lukas Korntheuer, Section of Medical Information Management, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria; E-mail: raffi.korni@gmail.com

Summary, Nurse Discharge Summary, Medication Summary, Diagnostic Image Report, Laboratory Report, Immunization Summary Report) represented in Health Level 7 (HL7) Clinical Document Architecture (CDA) format into the OMOP CDM.

2. Methods

Figure 2 gives an overview of the implemented ETL process. Data were extracted from the XML based HL7 CDA documents and imported into a staging area in the form of a MySQL database. The latter's tables form a relational representation (Figure 1) of the CDA document types. Its main purpose is to serve as the origin for the OHDSI tools White Rabbit [4] and Rabbit-in-a-hat [5], which we used to analyze the extracted data and to define the structural mappings. These tools currently do not support XML documents to be used as data source.

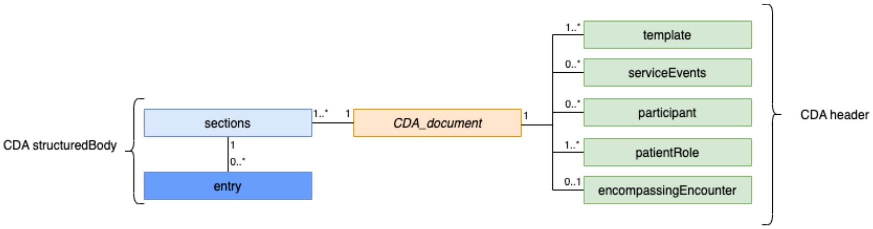

Figure 1. Relational representation of a CDA document

Next, the terminologies used in ELGA were mapped to the corresponding OMOP CDM standard terminologies. We applied OHDSI's tool Usagi [6] for manual mapping of proprietary codes. Data was then transformed to be conformant with structure and semantics imposed by the CDM tables definitions. Finally, the processed data was loaded into an OMOP CDM database.

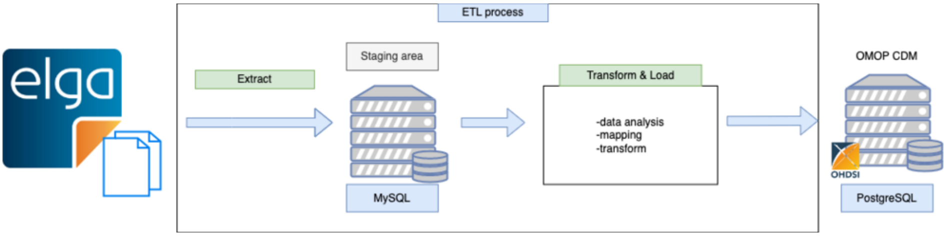

Figure 2. Overview of the implemented ETL process.

The automated parts of the ETL process were implemented using Python (version 3.9) and SQL in the dialect MySQL and PostgreSQL. As a database template version v.5.4 of the OMOP CDM was used.

2.1. XQuery-based extraction of CDA data

As ELGA CDA documents follow the XML standard, data can be extracted with the XQuery language. Through Python scripts and tailored queries, relevant data were extracted from different parts of the document (header and sections from the body) and written into the staging area (MySQL database) for use with the OHDSI tools. Since there are currently no publicly available de-identified CDA documents from real patients, we utilized the ELGA test documents provided on GitLab [7]. Only highly structured and coded information from CDA level 3 sections was extracted.

2.2. Transforming and mapping of data

The extracted data were imported into White Rabbit, which returns detailed information on the source tables. Next, the generated scan report was imported into Rabbit-in-a-hat to create a manual mapping to OMOP CDM tables and fields, by dragging and dropping arrows between table fields. Based on these mappings, Rabbit-in-a-hat generates a textual documentation for the planned ETL process, which we then implemented by means of Python scripts. For the mapping of standard terminologies (e.g., SNOMED), ATHENA [8] was used to download the OMOP CDM vocabularies and as a lookup tool. For the mapping of a variety of proprietary terminologies specific to ELGA, Usagi was used.

2.3. Loading data into an OMOP CDM database

Depending on the CDA document type, the mapped data were written to the predefined OMOP CDM tables within a PostgreSQL database through Python scripts, using the database connector from the Psycopg2 package.

3. Results

The developed ETL process revealed that, even though this varies depending on the document type, a substantial part of the OMOP CDM can be fed from the ELGA documents.

3.1.1. CDA header

Metadata stored in the header of the documents about the patient, participants (e.g. primary care physician), the encompassing encounter and the health service provided (*serviceEvent*) were transformed into the corresponding OMOP CDM tables (PERSON, PROVIDER, VISIT_OCCURENCE and PROCEDURE_OCCURENCE). Concerning the CDM PERSON table, 12 of the 18 fields were filled with information from the patient information stored in the CDA *recordTarget* element. The remaining 6 fields contained information on the ethnicity and race, which is not available in ELGA.

3.1.2. Document specific sections

Depending on the CDA document type, different kinds of sections were transformed into the OMOP CDM. Table 1 shows the sections from the *structuredBody* of the different CDA document types and the corresponding CDM tables.

Table 1. Body sections from the different CDA document types and the corresponding CDM tables

ELGA document types and included sections	CDM tables
Laboratory Report	
Specimen section	SPECIMEN
Specialty sections (e.g. hematology)	MEASUREMENT
Physician Discharge Summary	
Hospital Discharge DX	CONDITION_OCCURENCE
Hospital discharge medications	DRUG_EXPOSURE
Medications on admission	DRUG_EXPOSURE
Enclosed findings / vital signs	MEASUREMENT
Nurse Discharge Summary	
Care diagnoses	CONDITION_OCCURRENCE
Vital signs	MEASUREMENT
Hospital discharge disposition	OBSERVATION
Immunization Summary Report	
History of immunizations	DRUG_EXPOSURE
Problem list	CONDITION_OCCURRENCE
History of past illness	CONDITION_OCCURRENCE
Laboratory studies	MEASUREMENT
Treatment plan	PROCEDURE_OCCURENCE
Medication Summary	
History of medication use	DRUG_EXPOSURE
Diagnostic Image Report	
Current imaging procedure descriptions	MEASUREMENT

Since only coded and highly structured sections were extracted and transformed, the number of sections, we were able to map, varies depending on the document type (see Table 2).

Table 2. Proportion of ELGA CDA body sections that could be used for our ETL process

Document type	Total
Laboratory Report	2/5
Physician Discharge Summary	4/27
Nurse Discharge Summary	3/26
Immunization Summary Report	5/5
Medication Summary	1/1
Diagnostic Image Report	1/16

3.2. Terminology mappings

Laboratory results and vital signs are mostly coded in LOINC in ELGA as well as in the CDM, which made direct mapping possible. Furthermore, units are coded as UCUM in both systems. Other coded elements had to be mapped either through the “Non-standard to Standard map (OMOP)” relationship in the OMOP vocabulary or through manual mapping. All currently mapped ELGA terminologies can be seen in Table 3.

Table 3. Mapped ELGA terminologies and the corresponding terminologies in the OMOP CDM

ELGA CDA terminology	OMOP CDM terminology	Mapping
LOINC	LOINC	Direct
UCUM	UCUM	Direct
ATC	RxNorm	Automatic
ICD-10 BMG	SNOMED CT	Automatic
HL7:AdministrativeGender	OMOP Gender	Manual
HL7:ActSite	SNOMED CT	Manual
HL7:SpecimenType	SNOMED CT	Manual
HL7:Act Code	OMOP Visit	Manual
HL7:ParticipationFunction	Medicare Specialty	Manual
DCM	SNOMED CT	Manual
ELGA_LaborparameterErgaenzung	LOINC	Manual
ELGA_HumanActSite	SNOMEC CT	Manual
ELGA_SpecimenType	SNOMED CT	Manual
ELGA_Dokumentenklasse	OMOP Type Concept	Manual
ELGA_MedikationsArtAnwendung	SNOMED CT	Manual
PraxisOrientierte Pflegediagnostik	SNOMED CT	Manual

4. Discussion

In this paper we presented our ETL approach to transform data, extracted from ELGA CDA documents to the OMOP CDM. Although the OMOP CDM has gained popularity over the last years, existing work on using HL7 CDA based data in this context is sparse. In [9], an automated processing of the HL7 “Continuity of Care Document” is described. Documents were extracted separately, mapped, transformed, and loaded into an OMOP CDM database through a python package called “CCD2OMOP” [10]. In [11], data from five different sections (header, problem, medication, laboratory data, and procedure) of Korean referral CDA documents were transformed and loaded into an OMOP CDM database. Ji et al. showed a high conversion rate, depending on the section contained in the document. Since both approaches use a different type of CDA document compared to the ones used in ELGA, comparison of the results is difficult.

The results show that we were also able to map highly structured and coded sections/elements from the source documents and populate the corresponding CDM tables. Document types Medication Summary and Immunization Summary Report allowed a high conversion rate with respect to the contained sections.

However, document types Laboratory Report, Physician Discharge Summary, Nurse Discharge Summary and Diagnostic Image Report contain a substantial number of CDA level 2 sections with free text content, which we could not cover in our ETL process. These sections would provide a suite of additional valuable information that would require natural language (NLP) processing methods for their exploitation.

Even for coded data, transforming from one terminology to another potentially comes with the problem of losing some level of detail. We especially noticed that when mapping ELGA medication data to RxNorm via the OMOP-internal mapping from ATC to RxNorm (e.g. ELGA drug “Diazepam Actavis 10mg, Box of 20” would be reduced in RxNorm to “diazepam”). Mapping of the laboratory results and corresponding units was mostly successful since both systems use LOINC and UCUM as their standard terminologies. Laboratory results coded in proprietary ELGA terminology could mostly be manually mapped to a corresponding LOINC code.

Nevertheless, without a great amount of real-world test data, which is currently not available, a detailed validation and testing of the implemented ETL process is impossible.

5. Conclusion

In conclusion, a first step was made to integrate data from the Austrian nationwide EHR system to an international data standard for observational data. The biggest challenges we were confronted with are the lack of available test data and the automatic mapping of proprietary ELGA terminologies to corresponding OMOP terminologies. Topics that remain open for future work are pseudonymization, testing and validation with real world documents, and handling of CDA level 2 sections through NLP. We aim to provide a solution, that can be reused in future projects targeting the integration of CDA documents into the OMOP CDM. Our sources can be found at GitLab [12].

References

[1] S. Herbek *et al.*, ‘The Electronic Health Record in Austria: a strong network between health care and patients’, *Eur. Surg.*, vol. 44, no. 3, pp. 155–163, Jun. 2012, doi: 10.1007/s10353-012-0092-9.

[2] OHDSI, *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019. [Online]. Available: https://books.google.at/books?id=JxpnzQEACAAJ

[3] IBM Cloud Education, ‘What is ETL (Extract, Transform, Load)?’, *IBM Cloud Learn Hub*, Apr. 28, 2020. https://www.ibm.com/cloud/learn/etl (accessed Dec. 19, 2022).

[4] OHDSI, ‘WhiteRabbit’. Feb. 01, 2019. [Online]. Available: https://github.com/OHDSI/WhiteRabbit

[5] OHDSI, ‘Rabbit in a Hat’. Feb. 01, 2019. [Online]. Available: https://github.com/OHDSI/WhiteRabbit#-rabbit-in-a-hat

[6] OHDSI, ‘Usagi’. Apr. 09, 2021. [Online]. Available: https://github.com/OHDSI/Usagi

[7] ELGA GmbH, ‘CDA Beispielbefunde’. https://gitlab.com/elga-gmbh/cda-beispielbefunde (accessed Nov. 22, 2022).

[8] Odysseus Data Services, ‘ATHENA – OHDSI VOCABULARIES REPOSITORY’. 2023. [Online]. Available: https://athena.ohdsi.org/

[9] H. Abedtash, ‘AN INTEROPERABLE ELECTRONIC MEDICAL RECORD-BASED PLATFORM FOR PERSONALIZED PREDICTIVE ANALYTICS’, p. 184.

[10] H. Abedtash and J. Duke, ‘CCD2OMOP: An Interoperable Extract-Transform-Load Package to Support the Implementation of OHDSI Software Tools Across Non-OMOP- based Electronic Health Records’, *2016 OHDSI Symp.*, p. 3, 2016.

[11] H. Ji, S. Kim, S. Yi, H. Hwang, J.-W. Kim, and S. Yoo, ‘Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM’, *J. Biomed. Inform.*, vol. 107, p. 103459, Jul. 2020, doi: 10.1016/j.jbi.2020.103459.

[12] Raffael Korntheuer, ‘ELGA2OMOP’, *GitLab*. https://gitlab.com/007korni/elga2omop (accessed Jan. 29, 2023).