Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-929

Designing ETL Tools to Feed a Data Warehouse Based on Electronic Healthcare Record Infrastructure

Fabrizio PECORARO[a,1], Daniela LUZI[a] and Fabrizio L. RICCI[b]
[a]*Institute for Research on Population and Social Policies, National Research Council, Italy*
[b]*Institute for System Analysis and Computer Science, National Research Council, Italy*

Abstract. Aim of this paper is to propose a methodology to design Extract, Transform and Load (ETL) tools in a clinical data warehouse architecture based on the Electronic Healthcare Record (EHR). This approach takes advantages on the use of this infrastructure as one of the main source of information to feed the data warehouse, taking also into account that clinical documents produced by heterogeneous legacy systems are structured using the HL7 CDA standard. This paper describes the main activities to be performed to map the information collected in the different types of document with the dimensional model primitives.

Keywords. Data Warehouse, Extract, Transform and Load, HL7, Clinical Document Architecture, Electronic Health Record, Logical data map.

Introduction

One of the main issue in data warehousing is to design and develop Extract, Transform and Load (ETL) tools responsible for the integration of data provided by multiple information systems in a common target schema. This task is accomplished considering the distinct set of characteristics of each source system, such as database management system, operating system, communications protocols and data representation [1].

In healthcare, the Electronic Healthcare Record (EHR) represents an important infrastructure that provides an overview of the healthcare status of an individual to support physicians and other professionals in the delivery of care services. It collects standard clinical documents produced by heterogeneous legacy systems and repositories developed for different specialties and purposes and by different organizations (e.g. laboratory and hospital information systems, GP's record). In our vision the EHR can represent one of the main source of information to feed the data warehouse [2]. This approach simplifies the design of ETL procedures given that information collected in the EHR are structured using a common schema based on the HL7 CDA [3] that codes data using standard nomenclatures and dictionaries. However, the design and implementation of ETL procedures make it necessary to harmonize the different types of document structured using explicit CDA specifications, moving the integration issue from a source system to a document template point of view.

[1] **Fabrizio Pecoraro**, IRPPS-CNR, Via Palestro, 32, 00185 Rome, Italy (f.pecoraro@irpps.cnr.it).

Aim of this paper is to propose a methodology to design an ETL tool of a data warehouse architecture based on the different document templates stored in the Italian EHR system.

1. Methods

According to the dimensional model framework [4] the first step to design a data warehouse is to identify the business processes to be modelled. Once this task is accomplished the next steps are to identify the components of the data warehouse schema as well as the source of information necessary to describe them. To facilitate this task the conceptual framework to define the primitives of the data warehouse dimensional model based on the CDA concepts has been applied [5]. In particular, dimensions can be derived from: 1) HL7 Hierarchy defined by the triple <Participation, Role, Entity> that models subjects/objects involved in the process as well as the role played by them within the action [6]; 2) CDA Backbone composed by the specializations of the different Act classes as well as their relationships that models the actions documented in the CDA; 3) attributes of the Act class that identifies the type of action observed using coded vocabularies. These attributes can also be used to define quantitative or qualitative measures of the relevant business process. However, given that each business process can be described using information collected in different documents this task requires the analysis of the different CDA specifications to highlight similarities and differences in the representation of the main CDA concepts (e.g. cardinalities of relationships and attributes, vocabularies adopted to represent data). Finally, a logical data map is defined to specify how the information provided by the source documents are mapped in the data warehouse dimensional model.

2. Designing ETL tools based on the EHR

The following business processes have been selected to assess the proper provision as well as the outcome of healthcare services to patients with diabetes: 1) Hospitalization to analyze the number of patients hospitalized for specific diseases related to diabetes, such as stroke; 2) Laboratory tests to evaluate both the number of individuals that have carried out a given amount of clinical tests (e.g. glycated hemoglobin) and whether the result values are within a given range of values; 3) Medication therapy to study the number of patients treated with specific class of drugs, such as antihypertensive. They are modelled using a constellation schema as shown in Figure 1. It is composed by three facts (i.e. Hospitalization, LaboratoryTest and Medication) that share four conformed dimensions: Patient, Date, Facility and Provider. Moreover, three individual dimensions are selected: 1) Reason (i.e. disease(s) related to the hospitalization event); 2) TestType (i.e. type of measure captured during a laboratory test); 3) DrugType (i.e. specific pharmaceutical product administrated during a therapy). Both Reason and Provider dimensions require a bridge table to represent the diverse reasons related to a hospitalization and the different physicians involved in a specific event. Information that describe the chosen business processes are collected in the following documents stored in the EHR: 1) Report that collects information on laboratory results; 2) Discharge letter that gathers information relative to the patient's hospitalization; 3) Prescription that collects data on the request and supply of healthcare services, such as

a pharmacological therapy. These documents have been analysed and a portion of the CDA model is proposed in Figure 2. The CDA concepts defined in this analysis have been subsequently applied to develop the logical data map shown in Table 1 where, for each business process, the columns of each table of the data warehouse model are mapped with the attributes of the relevant CDA class.

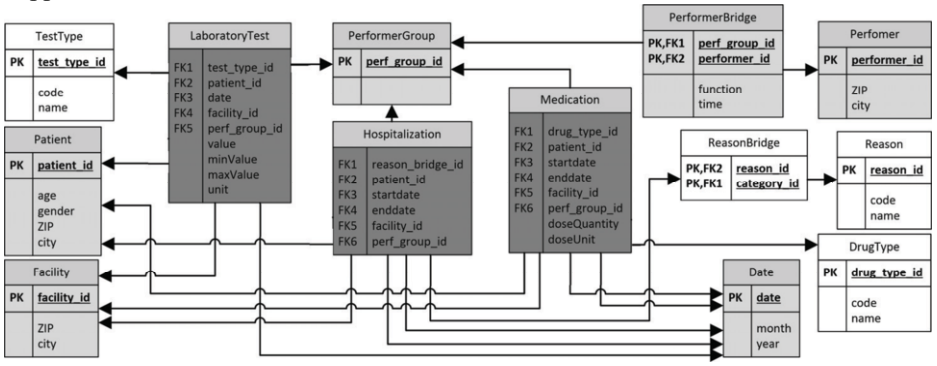

Figure 1. Constellation schema modeling the hospitalization, laboratory test and drug therapy business processes. It is represented by the three facts reported in dark grey, surrounded by four conformed dimensions reported in light gray as well as three individual dimensions reported in white.

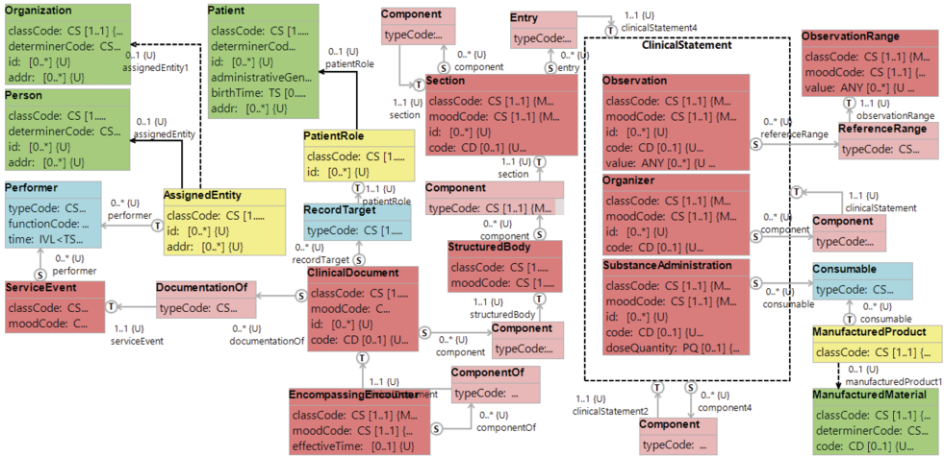

Figure 2. Portion of the of the CDA document describing an unified vision of the three document templates.

At the header level the three templates shared almost the same information with concepts that are semantically equivalent: the HL7 hierarchy <recordTarget, PatientRole, Patient> identifies the subject that receives the healthcare services described in the document; the HL7 hierarchies <performer, AssignedEntity, {Person, Organization}> outline both the physicians that carried out the ServiceEvent and their belonging Organization. Conversely, the Act encompassingEncounter that is not provided in the prescription has different meanings: in the report it describes the encounter in which the investigations documented in the CDA have been required (e.g. a specialist visit), while in the discharge letter it reports the hospitalization event itself to capture its start and end dates. Note that to correctly interpret the meaning of a CDA concept it is necessary to contextualize it depending on the document that contains it.

Table 1. Logical data map highlighting for each business process how the data warehouse model primitives (target table and column) are mapped with the CDA concepts (source document, path, class and attribute). Note that the symbol * specifies a recursive association, while ^ identifies an optional element.

Target		**Source**			
Table	**Column**	**Document**	**Path**	**Class**	**Attribute**
All business process					
Patient	id	All	recordTarget. PatientRole	Patient	id.extension
	age				birthTime.value
	gender				genderCode.code
	ZIP				addr.postalCode
	city				addr.city
Facility	id	All	ServiceEvent. performer. AssignedEntity	Organization	id.extension
	ZIP				addr.postalCode
	city				addr.city
Performer Bridge	function	All	ServiceEvent. performer	performer	functionCode
	time				time
Performer	id	All	ServiceEvent. performer. AssignedEntity	Person	id.extension
	ZIP				addr.postalCode
	city				addr.city
Hospitalization business process					
Hospitaliz.	startDate	Letter	ClinicalDocument	encompEncount	effectiveTime.low
	endDate				effectiveTime.high
Reason	code	Letter	Section	Observation	code.code
	name				code.displayName
Laboratory test business process					
Laboratory Test	value	Report	Section*.Organizer^	Observation	value.value
		Letter	Section		
	unit	Report	Section*.Organizer^		value.unit
		Letter	Section		
	date	Report	Section*.Organizer^		effectiveTime.value
		Letter	Section		
	minValue	Report	Section*.Organizer^. Observation	ObservRange	value.minValue
		Letter	Section.Observation		
	maxValue	Report	Section*.Organizer^. Observation		value.maxValue
		Letter	Section.Observation		
Drug therapy business process					
Medication	doseQuantity	Prescription & Letter	Section	SubstanceAdmin	doseQuantity.value
	doseUnit				doseQuantity.unit
	startDate	Prescription & Letter	ClinicalDocument	encompEncoun	effectiveTime.low
	endDate		Section	SubstAdmin	effectiveTime.high
TestType	code	Report	Section*.Organizer^	Observation	code.code
		Letter	Section		
	name	Report	Section*.Organizer^		code.displayName
		Letter	Section		
DrugType	code	Prescription & Letter	Section.SubsAdmin. consumable	manuLabelDrug	code.code
	name				code.displayName

At the body level information describing the clinical event are organized in different perspectives using the Section, Organizer and Act classes. For instance, the discharge letter and the prescription structure data in different Sections each one containing a set of inter-related Observations, while a report can have multiple self-associated Sections each one further structured in different Organizers to sub-group the events performed during the healthcare service provision. Such diversification requires the definition of different methods to extract the same information from each document template. This aspect is highlighted in the path column of the logical data map (Table

1) that specifies how to navigate the XML document to extract the specific attribute collected in the relevant class. Conversely, the three document specifications model the clinical event using an Act class of the CDA ClinicalStatement choice. In particular, the SubstanceAdministration and the related HL7 Hierarchy <consumable, manufacturedProduct, LabeledDrug> models the information about the drug delivered to the patient both in the prescription and the discharge letter. Similarly, both the report and the discharge letter describe a laboratory test through an Observation and the related ObservationRange. This similarity simplifies the design of transformation procedures as highlighted both in the class and attribute columns of the logical data map reported in Table 1. Particular attention has been posed in the mapping procedures of the Date dimension considering that each document type reports this information in more than one process and the same process can be described by different templates.

3. Discussion

The use of EHR as a source of information in a data warehouse architecture makes it easier to design and develop ETL tools given that clinical information provided by heterogeneous information systems are stored in a unique infrastructure and structured using standardized documents independently from the features of the source legacy systems. However, the integration of clinical information captured in different CDA implementations requires: 1) to disambiguate the meaning of the same concept that may change depending on the type of document where it is contained; 2) to implement specific procedures to extract the same information from different types of documents. In this paper we propose a methodology to accomplish these issues based on a conceptual framework [5] that maps the dimensional model primitives with the CDA concepts. This approach reduces the resources to be invested to implement ETL tools simplifying the identification of the dimensional model primitives that describe the business process to be analysed. Moreover, it makes it easier to implement transformation tools to load HL7 CDA messages in the data warehouse.

This approach is going to be tested using the EHR developed within the Smart Health 2.0 project that collects CDA documents produced by heterogeneous systems, such as GP's records, laboratory information systems, radiology information systems.

References

[1] P. Ponniah, *Data Extraction, Transformation, and Loading. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals,* (2001)

[2] F. Pecoraro, D. Luzi, F.L. Ricci, A Clinical Data Warehouse Architecture based on the Electronic Healthcare Record Infrastructure, *Proceedings of 7th International Conference on Health Informatics (HEALTHINF)* (2014).

[3] R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, A. Shabo, HL7 Clinical Document Architecture, Release 2, *J Am Med Inform Assoc* **13** (2006), 30-39.

[4] R. Kimball, M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling*, Wiley. Indianapolis. 2nd edition, (2009)

[5] F. Pecoraro, D. Luzi, F.L. Ricci. A conceptual Framework to Design a Dimensional Model Based on the HL7 Clinical Document Architecture. *Proceeding of Medical Informatics in a United and Healthy Europe (MIE),* (2014), IOS Press, 278-282.

[6] D. Luzi, F. Pecoraro, F.L. Ricci, G. Mercurio, A medical device Domain Analysis Model based on HL7 Reference Information Model, *Proceeding of Medical Informatics in a United and Healthy Europe (MIE)* (2009), IOS Press, 162-166.