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

doi:10.3233/978-1-61499-512-8-702

Semantic integration of medication data into the EHOP Clinical Data Warehouse

Denis DELAMARRE [abc], Guillaume BOUZILLE [abc], Kevin DALLEAU [b], Denis COURTEL [d] and Marc CUGGIA [abc,1]
[a] *INSERM, U1099, Rennes, F-35000, France*
[b] *Université de Rennes 1, LTSI, Rennes, F-35000, France*
[c] *CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France*
[d] *CHU Rennes, DIFSI, Rennes, F-35000, France*

Abstract. Reusing medication data is crucial for many medical research domains. Semantic integration of such data in clinical data warehouse (CDW) is quite challenging. Our objective was to develop a reliable and scalable method for integrating prescription data into EHOP (a French CDW). Method: PN13/PHAST was used as the semantic interoperability standard during the ETL process, and to store the prescriptions as documents in the CDW. Theriaque was used as a drug knowledge database (DKDB), to annotate the prescription dataset with the finest granularity, and to provide semantic capabilities to the EHOP query workbench. Results: the system was evaluated on a clinical data set. Depending on the use case, the precision ranged from 52% to 100%, Recall was always 100%. Conclusion: interoperability standards and DKDB, document approach, and the finest granularity approach are the key factors for successful drug data integration in CDW.

Keywords. Drug data, Clinical Data Warehouse, Data integration

Introduction

With the intensive development of electronic health records (EHR) and the large use at the bedside of computerized provider order entry (CPOE) functionalities, data concerning patients' drug treatment is now widely available and potentially processable for secondary reuses. Combining and mining drug data with other health data obviously has many application domains, such as pharmacovigilance, epidemiology, clinical research or evaluation of care quality. One solution to efficiently exploit drug data is to integrate it, with all other sources, into a Clinical Data Warehouse (CDW). Such integration faces several issues specifically related to the complexity of drug data. A drug is an entity which can be described in different ways (tradename, ingredients, clinical components, dose, galenic, etc.).

In a hospital, drug data are not static data. They are generated and collected all along the care process (from prescription to administration). Time scales can vary enormously (from minutes, in an intensive care unit, to weeks or months for long-term treatment). In order to exploit all its potentiality, it is crucial to exhaustively integrate

[1] Marc Cuggia. LTSI, Campus de Villejean, Université de Rennes 1, 2 Avenue Du Professeur Léon Bernard - CS 34317 -35043 Rennes Cedex, France. marc.cuggia@gmail.com.

drug data with the finest granularity. The smallest medication-dispensing unit represents this atomic granularity. It is also necessary to provide semantic functionalities so that CDW end-users can easily query and process this data. For instance, a way to detect a cohort of patients treated with quinolone (which involves clustering all patients who have been administered any quinolone brand drug forms), or to find all patients treated with drugs contraindicated for asthma (which entails building a query with a list of all drugs contraindicated for asthma).

This paper describes a scalable method for medication data integration and exploitation into a CDW. This method addresses the questions of data granularity and semantic complexity, and is based on two key ideas: (i) Data are natively integrated in the CDW by leveraging the PN13 interoperability standard, which is commonly used as standard for drug data exchanges between the components of hospital information systems (HIS) in France. (ii) A Drug Knowledge DataBase (DKDB) is integrated to provide semantic functionalities to the CDW query builder.

1. Methods

EHOP (previously named R-oogle)[1] is a CDW developed by the Health Big Data research team of INSERM/LTSI (Rennes 1 university). This CDW is currently used in several academic hospitals in France. The database contains both structured (e.g. labs results, DRG data) and non-structured patient data (e.g. clinical or imagery reports). The query workbench allows performing simple to complex queries with semantic expansion, negation or uncertainty detection and temporal constraints.

Phast[2] is a non-profit SDO developing interoperability standards, especially for medication workflows in health care facilities. PHAST maintains and publishes the PN13 standard, whose interoperability specifications are close to the HL7 formalism. PN13 is built on 2 complementary components: A technical framework, describing the structures and formats of B2B messages used for drug data exchanges; and the CIOsp (medication Interoperable Coding), which is a standardized medication nomenclature describing various medication attributes, which provides common, understandable vocabulary. CIOsp describes possible values for the status of messages related to medication administration, as well as clinical medication properties, such as active ingredients, strength and form.

THERIAQUE[3] is a comprehensive DKDB developed and maintained by the French National Hospital Center for Drug Information (CNHIM). Theriaque contains highly structured information and is widely integrated in many hospital information systems. Each drug is encoded by a medication-dispensing unit called UCD (equivalent to the National Drug Code provided by the FDA in the USA). Each UCD is annotated with a great deal of information; reference terminology is used for some of them (ATC for drug composition, ICD-10 for indication and contraindication information, etc). The database is available in different formats and can be easily integrated into third party applications. The Theriaque database is updated once a week.

The drug data source is DxCare® (Medasys), which is the EHR of the academic hospital of Rennes. This software provides CPOE functionality for clinicians and nurses, and communicates with the other HIS components through standard messages (mainly based on HL7 standards). For medication data, DxCare® is able to natively receive and transmit PN13/CIOsp messages. We captured this data stream and reused it

to populate our CDW. TALEND Studio® was used as an ETL component in order to extract, transform and load prescription data from DxCare® into EHOP.

The EHOP's database model is based on a star model, which is slightly different from I2B2 in that it introduces a specific fact table (named EHOP_Document) dedicated to storing documents. A document is any set of patient data (clinical reports, pathology reports, DRG summaries, lab results). A second fact table (called EHOP_structured) contains all the structured entities related to these documents. With this model, whatever the document structure is (freetext report, list of values, forms), patient data, which are embedded in a document, are stored as a whole, i.e., a single entity, thus preserving the integrity and the context of the information, as it was stored, originally, at the source. The same data are also cut into single atomic information units and stored, as a series of entity-attribute-values, into the second fact table. This mechanism allows users to search specific information, either by free text query or by structured query, and retrieve this information in the context of the document. The EHOP database model is also structured using several dimension tables, intended to describe the meaning of the data stored in the fact tables. For instance, there is a table dedicated to describing patients' demographic information, one dedicated to describing hospitalization, another contains all nomenclatures that are used in coded raw data (e.g. ICD-10, LOINC, SNOMED 3.5, CCAM, ADICAP), etc.

The prescription raw data are extracted from the legacy EHR, using a b2b connector intended to format these data into standard PN13/XML messages encoded using the CIOsp nomenclature. These messages contain all the information concerning a drug's prescription, such as the drug's label, UCD code, date and time of administration (start and end data/time), posology, indication, etc. The transformation and loading stages involved inserting each PN13/XML file, as a whole, in the EHOP_document table, and then splitting these same files into Entity-Attribute-Values, and eventually loading them into the EHOP_structured table.

At this step, no semantic aggregation can be performed on the prescription data since the PN13/XML files described each drug using only its label and UCD code. Therefore, we extracted the most useful information from the THERIAQUE database, such as the UCD code, ATC code, indications (coded in ICD), contraindications (Coded in ICD), composition, price, adverse effects (Coded in MedRa), etc. This information was then inserted into the EHOP-dimension tables The UCD codes were used to bind each prescription data element on the fact table to a drug reference in the EHOP-dimension tables.

2. Results

The EHOP's query workbench regroups functionalities intended to build queries both on structure or free text data. For leveraging the different attributes of drug data, a specific component called "drug query builder" (DQB) (Figure. 1) was added to the workbench. DQB allows to search and select drugs in various manners: by browsing through the ATC hierarchy, by searching drugs with one or several specific active components, or by selecting drugs based on a specific indication or contraindication. DQB also offers the possibility of putting some constraints on specific attributes, such as dose, period of prescription or route of administration. The DQB output is a SQL query, which selects prescriptions corresponding to a list of one or several UCD codes, and which also takes into account the conditions fixed in the attributes. This query can

be combined with Boolean or temporal operators from other queries.

As data were stored in PN13/XML format, a specific parser was developed in order to display the drug data, in real time, in a human readable format. According to the PN13 standard, a prescription is an order for the supply of the medication, as well as a set of instructions for the administration of the medicine to a patient. Therefore, each PN13 prescription order for a specific medication is considered to be an independent entity. This approach is not particularly suitable for a clinician or a researcher, who often wants to know in which context and with which other drugs a medication has been prescribed. For a better readability, we had to logically regroup these elements in order to reconstruct prescription sets. These prescription sets are visualized over time—by minute or by hour, by days, by stay in a ward or by hospitalization. This approach gives the user a comprehensive view of all prescriptions during the patient care process, while never dissociating a specific drug prescription from the others.

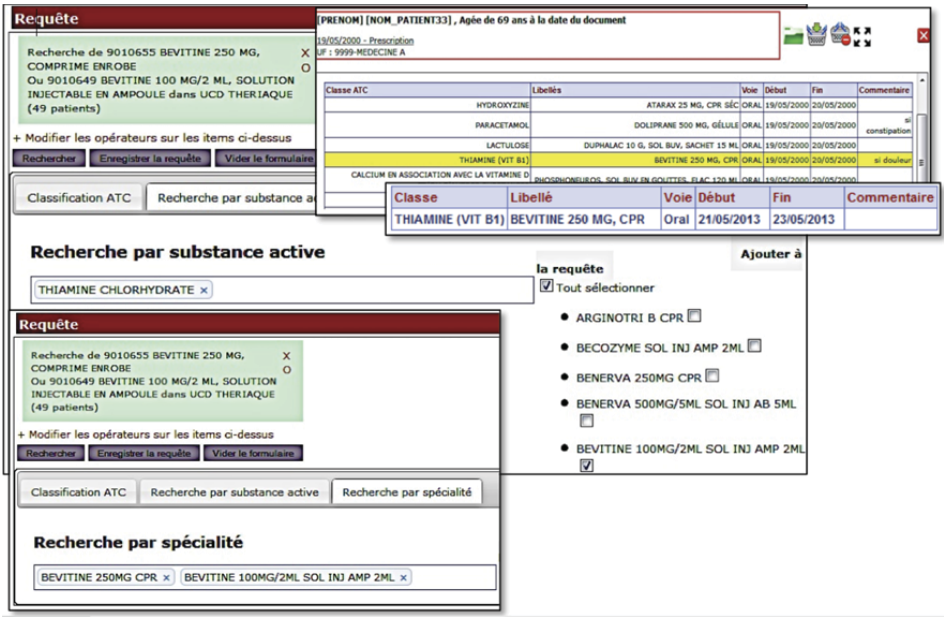

Figure 1. GUI of EHOP Drug Query Workbench (DQW)

We evaluated our system on a dataset consisting in 17722 various clinical documents, including 55 PN13 prescriptions of "Bevitine 100mg/2ml". A pharmacist previously validated this dataset. Three searching strategies were tested to retrieve these prescriptions: (i) Searching by brand name: by label (Bevitine 100mg/2ml) or [UCD: 9010649]: precision=100%; recall=100%. (ii) Searching by active ingredient: we searched all prescriptions containing: [Thiamine/Vitamin B_1 ; ATC: A11DA01]. 127 documents were found. Among them, we retrieved all 55 prescriptions of Bevitine 100mg/2ml (precision = 43,3%; recall=100%). Unsurprisingly, the low precision rate comes from the other documents retrieved by EHOP. All of them concerned other prescriptions of branded drugs which contained thiamine as their active component. (iii) Searching by string: in freetext research mode, when "Bevitine 100mg/2ml" was not misspelled, EHOP has found all 55 prescriptions (precision 100%; recall 100%).

3. Discussion

Drug data is probably one of the most complex data types to integrate, at a large scale, into CDWs, since there is actually no international interoperability standard or terminology system dedicated to this kind of data. Thus, each country develops and uses its own standards, which are more or less widely implemented. For instance, PN13 is actually a pre-standard, still under development, and part of the semantics of PN13 messages still depend on proprietary vocabularies defined and used by each EHR vendor. This interoperability issue is highly important, since there is a critical need for medical research to integrate and reuse drug data at multicentre and cross-border scales (EHR4CR, PCORNET, SHRINE[4,5]). Concerning the terminologies used, ATC lacks sufficient granularity and has a structural limitation insofar as accurately representing—by itself—the full complexity of drug data[6]. This is the reason we chose a DKDB (here Theriaque) rather than a unique terminology. Many other DKDB exist (e.g. in France: Banque Claude Bernard, Vidal) but these systems are not as open as Theriaque. RxNorm [7], which is based on UMLS, appears to us as the "ideal" DKDB for integrating and exploiting medication data, since it integrates multi-domain terminologies related to drugs (ranging from the drug itself, to its physiological effects, or metabolic pathways). In this work, we have only integrated prescription data, which was probably the simplest step to achieve. Integrating drug administration data is much more challenging. It implies, among other things, taking into account the different administration status (e.g postponed, stopped, on going, cancelled). It is however important to identify research use cases that justify integrating such complexity in CDW. CIOsp and CNHIM are currently developing web services dedicated to interacting with the EHR components. These WS aim to deliver updated drug information to third-party applications. We envision using these WS in our CDW, rather than integrating these terminological resources in a hard coded way. This method will contribute to achieve better and more sustainable quality in clinical dataset annotation.

References

[1] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F *et al.* Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform* 2011; **169**: 584–588.
[2] Phast - Information de santé standardisée (CIO - PN13 - MIO). http://www.phast.fr/index.php (accessed 24 Feb 2015).
[3] Husson M-C. [Theriaque: independent-drug database for good use of drugs by health practitioners]. *Ann Pharm Fr* 2008; **66**: 268–277.
[4] Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc JAMIA* 2014; **21**: 576–577.
[5] Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C *et al.* Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; **274**: 547–560.
[6] Winnenburg R, Bodenreider O. A framework for assessing the consistency of drug classes across sources. *J Biomed Semant* 2014; **5**: 30.
[7] Hernandez P, Podchiyska T, Weber S, Ferris T, Lowe H. Automated Mapping of Pharmacy Orders from Two Electronic Health Record Systems to RxNorm within the STRIDE Clinical Data Warehouse. *AMIA Annu Symp Proc* 2009; **2009**: 244–248.

Acknowledgements. We warmly thank PHAST and the CNHIM, who helped us to achieve this scientific work by providing their advice and by allowing us to use their technologies. Many thanks to Pr. Catherine Duclos for her expertise in pharmacy and CPOE technologies. The French National Agency of Research (ANR) funded this research work (RAVEL ANR-11-TECS-012).