User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-502

Populating the i2b2 Database with Heterogeneous EMR Data: a Semantic Network Approach

Sebastian MATE[a,1], Thomas BÜRKLE[a] , Felix KÖPCKE[a], Bernhard BREIL[b], Bernd WULLICH[c], Martin DUGAS[b], Hans-Ulrich PROKOSCH[a,d], Thomas GANSLANDT[d]
[a] *Chair of Medical Informatics, University Erlangen-Nuremberg, Erlangen, Germany*
[b] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[c] *Department of Urology, Erlangen University Hospital, Erlangen, Germany*
[d] *Center for Medical Information and Communication, Erlangen University Hospital, Erlangen, Germany*

Abstract. In an ongoing effort to share heterogeneous electronic medical record (EMR) data in an i2b2 instance between the University Hospitals Münster and Erlangen for joint cancer research projects, an ontology based system for the mapping of EMR data to a set of common data elements has been developed. The system translates the mappings into local SQL scripts, which are then used to extract, transform and load the facts data from each EMR into the i2b2 database. By using Semantic Web standards, it is the authors' goal to reuse the laboriously compiled "mapping knowledge" in future projects, such as a comprehensive cancer ontology or even a hospital-wide clinical ontology.

Keywords. i2b2, electronic medical records, secondary use, semantics, controlled vocabulary, heterogeneous data integration

1. Introduction

Data collection for cross-institutional research projects or the annotation of biospecimens is often done by manual reentry of data into a shared database. This process is error-prone, time-consuming and may result in incomplete data collection. With the shift from paper-based to electronic documentation in recent years, much of this data is already captured in various subsystems of the hospital information system, for example in the *electronic medical record* (EMR). It is tempting to reuse this data for research purposes. However, while technical access to these databases is easy, it is very difficult to process this data in a semantically correct manner, especially if it's not encoded with a standardized coding system. This task is getting even harder when trying to merge medical data from different hospitals. The efficient reuse of these large pools of precious information has been declared as a major challenge for medical informatics in the near future [1].

The *Deutsches Prostatakarzinom Konsortium e. V.* (DPKK) is a German cross-institutional research network consisting of more than 70 urologists, pathologists and

[1] Corresponding author: Sebastian.Mate@imi.med.uni-erlangen.de

scientific researchers to fight prostate cancer. Similar to the CPCTR's efforts in the US [2], one of their goals is to establish a shared database of tissue specimen, containing annotation data from the patients' medical history, surgery and pathology. Recently, a new common dataset has been defined by DPKK experts in Erlangen and Münster, comprising 26 medical concepts (e.g. pTNM) with 154 atomic enumerable values (e.g. pN=0) and 12 medical concepts with non-enumerable values (e.g. the PSA value). The current web-based DPKK research database implementation, however, requires the reentry of such clinical annotation data, even though most of the data are already stored in the partners EMR systems. We therefore evaluate a new single source approach based on i2b2, a NIH-funded, open source clinical data warehouse and translational toolkit [3], as a pilot project between the university hospitals in Erlangen and Münster. i2b2 features a generic database schema and enables the easy and user-friendly construction of database queries to determine patient cohorts based on the combination of eligibility criteria [4].

In order to reuse the data elements already stored within the two hospitals' EMR systems, we had to implement ETL (extraction/transformation/loading) steps to load those data into the i2b2 database. Since i2b2 does not provide an integrated means for data loading, we had to establish those functions externally. For this purpose we decided not to use proprietary import/export programs between the respective EMR systems and i2b2, but to extend i2b2 with an ontology suite, which supports the generic mapping of heterogeneous EMR data to a set of common data elements. These mappings can then be processed to perform the data export into the i2b2 research database. By using Semantic Web standards [5] for the definition of machine processable, declarative mappings, it is our vision to reuse the now laboriously compiled "mapping knowledge" in future projects, combined with other freely available medical ontologies [6] in the context of a comprehensive cancer or hospital ontology.

2. Methods

We chose an approach in which all required information is represented with semantic networks in the flexible Web Ontology Language (OWL) [5], as illustrated by the two bold arrows in figure 1. The targeted DPKK dataset is defined inside a target ontology describing all data elements, which shall be exported into the i2b2 database. There, all concepts are stored in a taxonomy-like structure with attributes such as name, datatype, and a short textual description, plus, if applicable, i2b2 specific attributes such as medication and lab value ranges. To speed up the ontology editing process, we have developed *OntoEdit* for entering and editing those contents. In a similar manner, each source system's EMR data structure (i.e. data entry forms, data input fields, enumerable value lists, checkboxes and radio buttons within Soarian metadata) has to be defined in shape of a source ontology. This source ontology also contains technical information on how to access to the source system's database in order to retrieve the data records represented by each ontology concept. The creation of this ontology is custom to each source system. If direct access to the source system's EMR metadata is difficult (e.g. because of licensing issues), we have implemented *OntoGen* to support the import and use of CSV files instead. OntoGen publishes the data records from the CSV file in a temporary database and automatically derives the ontologies from the columns' headlines and by aggregating data values.

When the source and target ontology have been defined in OWL, mappings between the two can be defined inside a flexible *mapping ontology*. Figure 1 illustrates two different types of mappings. In the first example, the target concept D is directly mapped to source concept B using the *hasImport* relation, because D exactly matches B. Therefore, the corresponding data records from B can be exported to i2b2 without any data transformation. In some cases, however, filtering and transforming of source concepts may be necessary in order to conform to the concepts in the target ontology. We express such operations with intermediate transformation nodes. This is illustrated in figure 1 with an *ADD* node between the concepts E, A and C, which means that the target concept E is the sum of the source concepts A and C. To keep operations "semantically atomic", nodes are limited to 2 operands; complex operations can be expressed by cascading multiple nodes into expression trees (an example will be given later in figure 2). We have developed *QuickMapp* for the easy creation of such mappings. In order to actually perform the data export from the source EMR to i2b2, an export software, *OntoExport*, automatically translates the information stored inside all ontologies into SQL statements. These extract the source systems' data records, transform them according to the mapping rules and write them into the target i2b2 database.

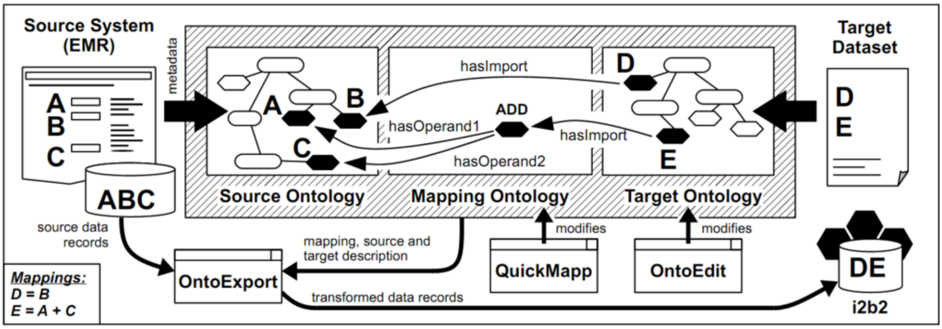


Figure 1. All information to perform a data export is described in semantic networks.

3. Results

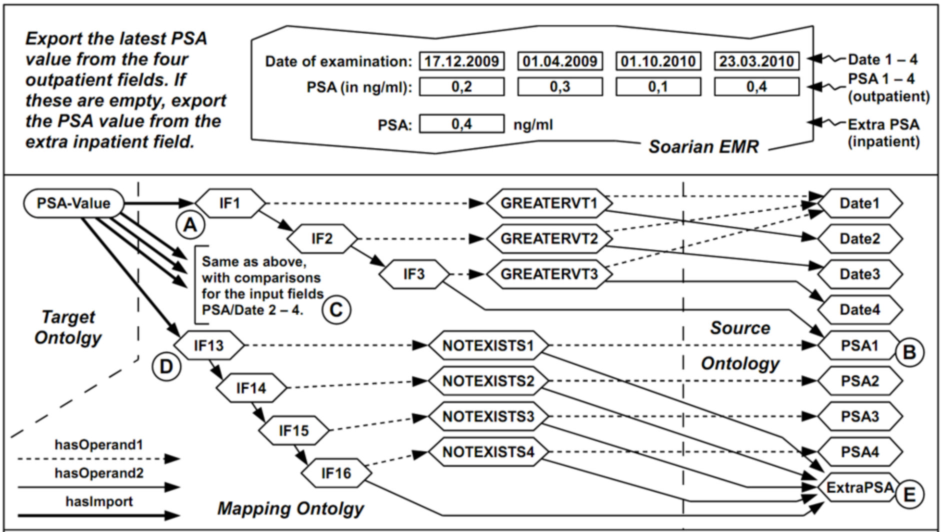
We have implemented this approach for the EMR systems Siemens Soarian Clinicals® in Erlangen and Agfa HealthCare ORBIS® in Münster. In Erlangen, we were able to derive 42,000 ontology elements from the Soarian EMR by processing its metadata tables. Because direct database access to ORBIS in Münster was not allowed, we had to use a CSV export from the EMR and post-process it with *OntoGen*.

Table 1 summarizes the achieved mapping results. More than 75% of the required DPKK data elements could be matched directly from the two EMR systems. For 10 data elements in Erlangen and one in Münster, transformation nodes had to be defined in the mapping ontology. In Erlangen, four of these data elements required checking whether a specific data entry form existed. This could only be implemented by using a workaround, which simulates another – in reality nonexistent – database table, which stores this abstract information. One mapping in Erlangen was not supported yet, because it required access to administrative data (date of birth) from the ADT system. *OntoExport* however, is currently limited to process data records from only one database connection at a time. At both sites, two mappings were impractical to create because our current implementation is limited to mappings at the value level only. Creating them with the current implementation would result in 108 distinct partial mappings.

Table 1. Result after mapping the two EHRs to the common DPKK dataset with 166 concepts.

	Erlangen Hospital	Münster Hospital
No. of concepts directly mapped	138	127
No. of concepts mapped through transformations	10 (4 with a workaround)	1
No. of concepts not documented in source system	15	36
No. of mappings not supported / impractical	1 / 2	0 / 2
Generated SQL statements / execution time:	548 / ~15 seconds	284 / ~ 3 seconds
Number of facts / patients in source table:	29,721,416 / 161,512	5,100 / 500 (test data)
Obtained facts / patients for DPKK i2b2:	3,686 / 155	2,585 / 487 (test data)

Concerning our method's mapping capabilities, we have successfully implemented and tested various types of string manipulation as well as arithmetic, Boolean and comparison operations. Figure 2 shows a complex real-world example from Erlangen. The DPKK data set requires the latest PSA value from each respective EMR. In Soarian, four outpatient follow-up PSA measurements are stored in four data fields with attributed date fields. A fifth, extra PSA field contains the last inpatient value, if no outpatient follow-up was done so far. The export logic is comprised of five distinct partial mappings with conditional checks. The first mapping (A) checks whether *Date1* is greater than *Date4, Date3* and *Date2* and only then exports the *PSA1* field (B): Only if all *GREATERVT* nodes evaluate to "True" (the "*VT*" variant in particular allows the comparison of blank/null fields), the nodes *IF3*, *IF2* and *IF1* pass the data records abstracted by the *PSA1* concept into the i2b2 database. Likewise, three other mappings (C) process the fields for PSA/Date 2, 3 and 4. The fifth mapping (D) checks if all outpatient fields are empty and eventually exports the extra inpatient PSA value (E).



GREATERVT: Returns "True" if Operand 1 is greater than Operand 2; NULL values are allowed. ----- NOTEXISTS: Returns "True" if Operand 1 is NULL and Operand 2 is not NULL. ----- IF: If Operand 1 is "True", the node returns the values from Operand 2 as a result.

Figure 2. Example of a complex transformation: PSA mapping for Erlangen.

4. Discussion

There have been several prior projects to integrate and query heterogeneous medical data, e.g. [7, 8]. However, most of these implementations are stand-alone systems that require the formulation of complex queries in proprietary query syntax, while our approach reuses an existing platform (i2b2) for the final data integration that also acts as a proven, easy-to-use query interface.

One major advantage of our approach is that the transformation and loading processes between EMR source data structures and the DPKK target data set are not implemented with proprietary import/export program and SQL code, but defined on a higher, more generic and reusable ontology level. Thus, the mappings and domain knowledge can be reused in other i2b2 and warehousing projects and can be processed with standard tools such as Protégé. Furthermore, extending the data integration pilot project to further EMR systems (as it is planned in a next step) will reutilize the already defined target ontology and only require the definition of new source ontologies. These need to be compatible with the database's SQL syntax and the EMR's data schema in order to create proper SQL statements for the data extraction.

The current implementation must still be considered prototypical as it offers opportunities for improvement. We plan e.g. to improve the SQL code generation by optimizing the node's processing order. We further plan to extend the ontology suite to support mappings at different hierarchy levels instead of the value-level only. Currently, we have limited the target ontology's semantic features to the functionality of the i2b2 system. We are confident, however, that we will be able to expand or link the target ontology to a more powerful ontology that follows commonly accepted desiderata [9] and standards [10] for medical terminologies. By using the OWL format, our approach can act as a bridge between raw medical data, i2b2 and the Semantic Web, because it enables the linkage to other freely available medical ontologies [6].

Thus, by using i2b2 and extending it with our ontology suite we feel confident that we have made a step forward in efficiently accessing and reusing EMR data from routine care for a cross-institutional research database.

References

[1] Prokosch HU, Ganslandt T. Perspectives for Medical Informatics: Reusing the Electronic Medical Record for Clinical Research, *Methods of Information in Medicine* **48** (2009), 38–44.

[2] Patel AA, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: The Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. *BMC Cancer* **5** (2005).

[3] Murphy SN, et al. Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu Symp Proc. 2007* (2007), 548–552.

[4] Deshmukh VG, et al. Evaluating the informatics for integrating biology and the bedside system for clinical research, *BMC Med Res Methodol* **9** (2009).

[5] Ruttenberg A, et al. Advancing translational research with the Semantic Web, *BMC Bioinf* **8** (2007).

[6] Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support, *Yearb Med Inform* (2008), 67–79.

[7] Sujansky W. Heterogeneous database integration in biomedicine, J Biomed Inform 34 (2001), 285–298.

[8] Hernandez T, Kambhampati S. Integration of Biological Sources: Current Systems and Challenges Ahead, *SIGMOD Rec* **33** (2004), 51–600.

[9] Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century, *Methods Inf Med* **37** (1998), 394–403.

[10] Solbrig HR. Metadata and the Reintegration of Clinical Information: ISO 11179. *M.D. computing: computers in medical practice* **17** (2000), 25–28.