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

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 often leads to lazy and incomplete contributions of data. 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 access and process the completely heterogeneous data elements in a semantically correct manner. This task gets even more complicated when trying to merge medical data from different hospitals. The efficient reuse of these giant pools of precious information has been declared as a major challenge for medical informatics in the near future [1].

[1] Corresponding Author.

The *Deutsches Prostatakarzinom Konsortium e.V.* (DPKK) is a German cross-institutional research network consisting of more than 70 urologists, pathologists and basic researchers to fight prostate cancer. Similar to the CPCTR's efforts in the US [2], one of their goals is to build 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 DPKK research database implementation, however, requires the above mentioned reentry of annotation data and has certain limitations regarding the extensibility of the implemented dataset. As a pilot project between the university hospitals in Erlangen and Münster, we are evaluating *Informatics for Integrating Biology and the Bedside* (i2b2) [3] as a new DPKK research database, because it features a generic database schema and allows for easy construction of powerful database queries with Boolean set operations [4]. Since most of the data elements required by the DPKK are already captured in the EMRs in Münster and Erlangen, the project aims to reuse this existing data. Because i2b2 is mainly designed for querying and processing purposes, it does not provide any powerful means for the loading of data into its own database. In a subproject described in this paper, it was our goal to develop a system for the mapping of heterogeneous EMR data to a set of common data elements, which can then be exported into an 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. We have implemented this approach for the EMR systems Siemens Soarian Clinicals® in Erlangen and Agfa HealthCare ORBIS® in Münster.

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 shared DPKK dataset is specified by a target ontology that describes all data elements to be exported into the i2b2 database. All concepts are stored in a taxonomy-like structure and linked to other properties denoting their name, datatype, a short textual concept description and other i2b2 specific attributes such as medication and lab value ranges. To speed up the editing process, we have developed *OntoEdit* for the quick entry of these properties into the ontology. Similarly, each source system's metadata has to be represented in shape of a source ontology that describes the system's overall document structure in hierarchical form (i.e. forms, attributes and enumerable values). The source ontology also describes how the source system's databases can be accessed in order to retrieve the facts data. This includes information about the data tables and how these tables have to be queried in order to fetch the desired data records. The creation of this ontology is custom to each source system. For cases in which direct access to the system's metadata is difficult (e.g. because of licensing issues), we have implemented *OntoGen* to allow the use of simple CSV exports instead. OntoGen publishes the data from the CSV file in a database and automatically derives the three ontologies from the columns' headlines and by aggregating data values (see figure 1, dotted arrows).

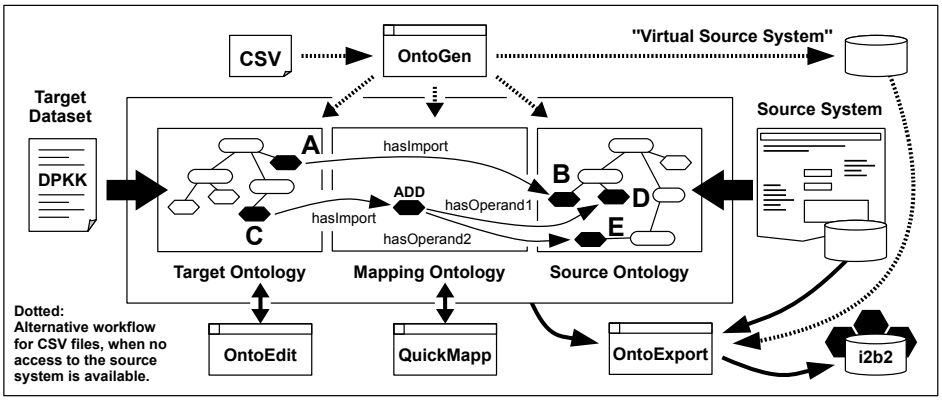

Figure 1. The mapping and data integration concept.

Once all required information is available 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, concept A is mapped to concept B by a *hasImport* relation. This means that the data element in the source system can be exported to i2b2 without any data transformation, since it matches the desired concept in the target ontology. However, there might be a need for filtering and transforming data. We express such operations with intermediate transformation nodes. This is shown with an ADD node between the concepts C, D and E, which means that the concept C is the sum of the concepts D and E. To keep operations "semantically atomic", nodes are limited to 2 operands; more complex operations can be expressed by cascading multiple nodes into trees. We have developed *QuickMapp* for the easy creation of such mappings. In order to actually create the i2b2 metadata and perform the data export to i2b2, an export software, *OntoExport*, translates the ontologies into SQL statements. This is done by processing all intermediate nodes in correct order, passing the interim results from node to node. For every step, an SQL statement specific to the database and the node's operation is created by using information from another ontology.

Results

In Erlangen, we were able to derive 42,000 ontology elements from the Soarian EMR (including full attribute and value versioning) by processing its metadata using SQL scripts. To access ORBIS in Münster we used a CSV export with OntoGen, while replacing the generated target and mapping ontologies with the target DPKK target ontology and a custom mapping ontology. Table 1 summarizes the mapping results. For Erlangen, four mappings required checking whether a form has been filled out; this could only be implemented by using a small trick. Another mapping was impossible because it required administrative data from a different system. At both sites, two mappings were impractical to create because our current implementation is limited to mappings at the value level only. Concerning our method's mapping capabilities, we have successfully 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 PSA outpatient follow-up exams are stored in four data fields plus respective date time fields. An extra field is used, if no outpatient follow up has happened so far. Thus, the export logic is split into five mappings for conditional checks. The first one checks whether Date 1 is greater than Date 4, 3 and 2 (nodes GREATERVT3 to 1, the "VT" variant allows the comparison of blank fields). If this is true, the IF-nodes pass the data records from the PSA 1 field into the i2b2 database. Three similar mappings have to be created for the fields PSA/Date 2, 3 and 4. Finally, the last mapping checks whether the above fields are empty and then exports the extra PSA field.

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

	Erlangen Hospital	**Münster Hospital**
Directly mapped (*hasImport* only) concepts:	138	127
Through transformations mapped concepts:	10 (4 required a small trick)	1
Concept is not documented in source system:	15	36
Currently impossible / impractical mappings:	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)

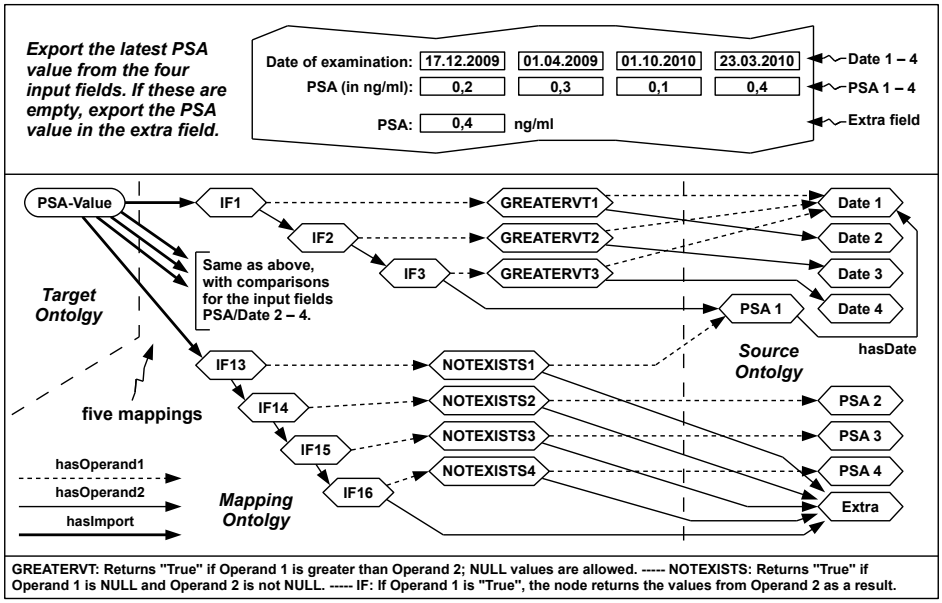

Figure 2. PSA mapping for Erlangen.

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 queries in custom 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. Our approach is specifically designed to persistently map time-

stamped medical data elements, i.e. from the EMR or the laboratory system. This restriction allowed us to keep the system simple, but adequate for the targeted i2b2.

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 need to extend the abilities to support mappings at different granularities instead of the value-level only. This would allow us to recreate the above-mentioned two impractical mappings with ease. Although the implemented manual mapping process is obligatory in order to achieve correct results, we plan to integrate techniques to support the user with matching suggestions, similar to [9, 10]. 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 to a custom, more powerful ontology that follows commonly accepted desiderata [11] and standards [12] 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]. The captured mapping and domain knowledge can be reused in other i2b2/warehousing projects and processed with standard tools such as Protégé. Furthermore, the global scope of identifiers (URIs, [5]) for the definition of medical concepts is crucial in networked research, as it is conducted by the DPKK.

Because of temporal aspects (individual patients' data being stored with a timestamp), calculations between clinical data elements from different source systems are impossible. This is not a limitation of our method; in fact it can be overcome with i2b2's query model and its timeline visualization. Thus, by using i2b2 with our data integration approach, we were able to achieve synergetic effects. In combination with Semantic Web standards to manage knowledge, we feel confident that we have made a step forward in efficiently accessing and reusing EMR data from routine care.

References

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

[2] A. A. Patel 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] S. N. Murphy 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] V. G. Deshmukh et al., Evaluating the informatics for integrating biology and the bedside system for clinical research, *BMC Med Res Methodol* **9** (2009).

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

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

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

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

[9] Y. Sun, Methods for automated concept mapping between medical databases, J Biomed Inform 37 (2004), 162–78.

[10] A. Ghazvinian et al., Creating mappings for ontologies in biomedicine: simple methods work, *AMIA Annu Symp Proc* (2009), 198–202.

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

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