Populating the i2b2 database with data from the electronic medical record:

Ontology-Based Form Data Integration for i2b2

See you at MIE2011, Oslo!

Sebastian Mate[1], Thomas Bürkle[1], Hans-Ulrich Prokosch[1,2], Thomas Ganslandt[2]

Chair of Medical Informatics
Prof. Dr. Hans-Ulrich Prokosch

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

INTRODUCTION

Case report form (CRF) data capture systems and the electronic medical record (EMR) store precious data that is tempting to reuse in medical research [1]. However, while technical access to the systems' databases is easy, it is very difficult to reuse the data for research purposes. Although stored in a structured way, it is typically not encoded with standard terminologies like the UMLS or SNOMED. Furthermore, single medical concepts are often expressed by a combination of multiple CRF components (text input fields, checkboxes and radio button groups). These relationships are, however, only apparent to humans and are not stored explicitly in the database schema. Using standard data integration tools, the time-consuming work of remodeling data relationships for export is typically "buried" inside SQL scripts, which are difficult to maintain. What is even worse is that the tediously modeled knowledge is lost for further processing.

METHODS

We have developed a system [2] to extract, transform and load heterogeneous CRF/EMR data into an i2b2 research database [3]. In contrast to standard data integration approaches, our system is designed to address the knowledge-related issues mentioned above. Instead of working at the database level, relationships between form components, data transformation and filtering rules are expressed in „higher-level" OWL ontologies [4]. This allows easier maintenance and the semantic interpretation of the created knowledge from a technical, administrative and medical perspective.
The system works by describing both the *source system* (EMR) and the *target dataset* independently with ontologies. The source ontology is an abstract, technical description of the source system that provides information on how to access the data records behind each form component, done in a machine-interpretable way. The target ontology describes the collection of medical concepts to be made available in the i2b2 system, along with semantic features available in i2b2 (e.g. a mono-hierarchy of medical concepts, data type, medication and lab value ranges – if applicable). The target ontology is also used to automatically create the i2b2 ontology.

In order to link the source system (EMR) to the target system (i2b2) and to perform the export, a mapping ontology has to be created. This ontology contains manually created semantic relationships between the source and the target ontology, which can be processed by a software component to perform the data transfer. Relationships can either be „simple" one-to-one or „complex" mappings, with full data filtering and transformation rules. The latter ones are defined by using intermediate operation nodes, as illustrated with the ADD node on the image on the right side. Intermediate nodes can also be cascaded into full expression trees to allow comprehensive data filtering and arbitrary data transformation. By sequentially processing these nodes in a correct order, SQL statements are automatically generated to perform the data transfer from the source systems to the target i2b2 database.

RESULTS

We have evaluated the system's capabilities in a cross-institutional data sharing project between the university hospitals in Erlangen and Münster, Germany. Its objective was to establish a joint i2b2 research database with annotation data from prostate cancer patients for the *Deutsches Prostatakarzinom Konsortium* (DPKK), a large prostate cancer research network in Germany.

The challenge was to map two distinct EMR systems to a central i2b2 database with a common dataset. As illustrated in the table below, we were able to map almost all of the data elements available in the cancer documentation of the two EMRs (*Siemens Soarian*® in Erlangen and *Agfa Orbis*® in Münster) to the set of 166 medical concepts. Only one mapping was impossible, two mappings were impractical to create with our current system.

DISCUSSION

The current implementation must still be considered prototypical as it offers room for improvements. For example, the system is currently designed to import data to i2b2 only, while being limited to the functionality and semantics of i2b2. We are confident, however, that we will be able to describe generic target database systems and that we will be able to better differentiate between metadata and facts data. This would allow us to derive and reuse collected mapping information in other without transforming actual data. We also hope to extend the system to follow commonly accepted desiderata and standards for medical terminologies [5, 6, 7].

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.

[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] S. Mate et al., Populating the i2b2 Database with Heterogeneous EMR Data: A semantic Network Approach. MIE proceedings (2011), accepted for presentation at MIE 2011, Oslo.

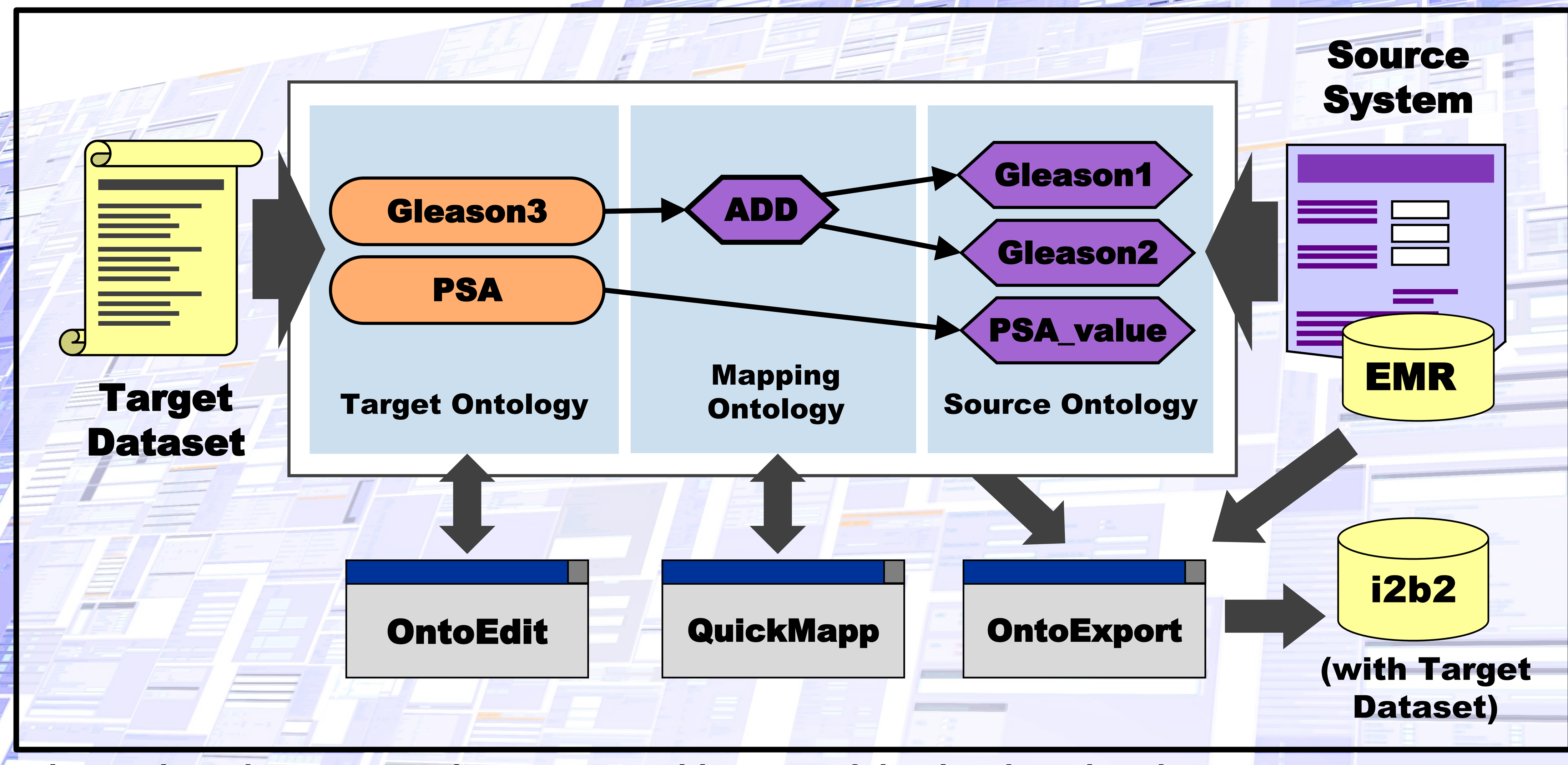
[3] S. N. Murphy et al., Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007 (2007), 548–552.

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

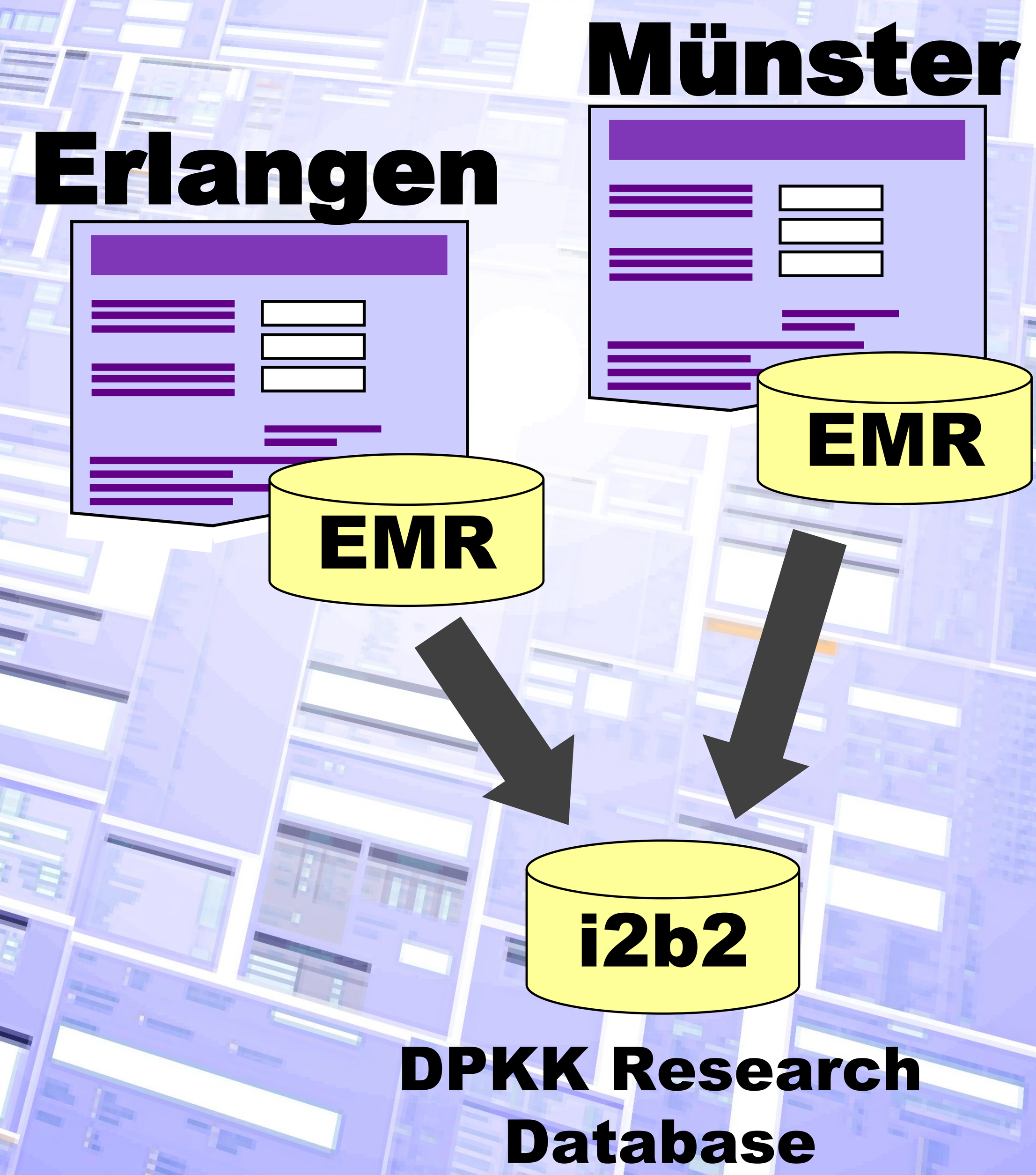
[5] J. J. Cimino, Desiderata for Controlled Medical Vocabularies in the Twenty-First Century, Methods Inf Med 37 (1998), 394–403.

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

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



Above: the Erlangen mapping concept with some of the developed tools.



	Erlangen Hospital	Münster Hospital
No. of concepts directly mapped (simple 1:1)	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)

Table above: Result after mapping the two distinct EMRs to a set of 166 common medical concepts. Our approach has proven to work well, because for almost all of the target concepts a mapping could be created.

Background image: visualization of the 780 forms in the Erlangen EMR with a total of 42,000 (potentially) semantically distinct data elements (including versioning).

[1]Chair of Medical Informatics, University of Erlangen-Nuremberg, Erlangen, Germany
[2]Center for Medical Information and Communication Technology, Erlangen University Hospital, Erlangen, Germany