Populating the i2b2 Database With Heterogeneous EMR Data: A Semantic Network Approach

Sebastian Mate\(^a\), 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

28.09.2011, Oslo
Introduction: Replacing Double Data Entry With “Single Source” – The DPKK Example

- EMR DPKK Partner A
- EMR DPKK Partner B
- DPKK Researcher
- Original DPKK database
- Query
- ETL
- i2b2 System
- EMR database
- NEW!
- Erlangen
- Münster
Problem 1: How to align and merge data from independent sources?

- Different data structures (different vendors)
- Similar information (full prostate cancer documentation in EMR)
Introduction: Developing the Data Integration Approach

Problem 2: How to efficiently access and process the huge piles of data elements?

Data:
- Structured
- Uncoded

Erlangen as an example: > 42,000 data elements (including data element versioning)
Methods: The “Ontology Layer”

Ontology / Knowledge Base Layer

Tools / Data Access Layer

Data / Storage Layer
Express the target dataset with a target ontology.

Express the source system(s) with one or more source ontologies.

Create a mapping ontology to connect concepts from the target ontology to concepts from the source ontology.

Some concepts can be mapped directly, others require transformations and/or data filtering, which can be expressed with intermediate nodes.
To allow complex data transformations, nodes can be cascaded into "expression trees".

OWL-to-SQL translation: For each mapping node, an SQL statement, which performs the operation expressed by the node, can be constructed automatically.
Methods: OWL to SQL Translation

INSERT INTO TEMPTABLE(E, A, V)
(SELECT E, A, V FROM …) OP1
FULL OUTER JOIN
(SELECT E, A, V FROM …) OP2
ON OP1.E = OP2.E
WHERE OP1.V IS NOT NULL AND ...

How to process nodes
How to access data
Results: Developed Prototypical Tools
QuickMapp

Universitätsklinikum Erlangen

PSA-Wert_postoperativ: IF (GREATERVT Value14712_14400_1 Value14712_14398_1) (IF (GREATERVT Value14712_14400_1 Value14712_14397_1) (IF (GREATERVT Value14712_14400_1 Value14712_14399_1) Value14712_14392_1))
Results: Ontology Mapping for DPKK
Results in the Münster and Erlangen Mapping

- Described i2b2 as a target system in OWL (incl. DPKK dataset)
- Created source ontologies for both EMRs (Soarian and ORBIS)
- Expressed many types of Oracle database operations inside the ontology (as it was illustrated with the blue “cloud”)
- Preliminary mapping results for Erlangen and Münster (from a total of 166 concepts):

<table>
<thead>
<tr>
<th></th>
<th>Erlangen Hospital</th>
<th>Münster Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directly mapped (hasImport only) concepts:</td>
<td>138</td>
<td>127</td>
</tr>
<tr>
<td>Through transformations mapped concepts:</td>
<td>10 (4 required a small trick)</td>
<td>1</td>
</tr>
<tr>
<td>Concept is not documented in source system:</td>
<td>15</td>
<td>36</td>
</tr>
<tr>
<td>Currently impossible / impractical mappings:</td>
<td>1 / 2</td>
<td>0 / 2</td>
</tr>
<tr>
<td>Generated SQL statements / execution time:</td>
<td>548 / ~15 seconds</td>
<td>284 / ~3 seconds</td>
</tr>
<tr>
<td>Number of facts / patients in source table:</td>
<td>29,721,416 / 161,512</td>
<td>5,100 / 500 (test data)</td>
</tr>
<tr>
<td>Obtained facts / patients for DPKK i2b2:</td>
<td>3,686 / 155</td>
<td>2,585 / 487 (test data)</td>
</tr>
</tbody>
</table>
Discussion
Limitations and How to Tackle Them: TODOs

- **Re-implementation** of the current proof-of-concept
  - **Client/server** architecture to improve administrative aspects
  - Integration with other data integration tools (e.g., Talend OpenStudio) to simplify access to multiple databases

- **Expression of complex** and **abstract relationships** between data elements, e.g., between data elements at different hierarchy levels

- Allow certain calculations between data elements from multiple forms or different source systems (=> **handling of temporal aspects**)

- Strong desire to export EMR data to other systems besides i2b2: allow description of **arbitrary target systems** and database models
Discussion: Benefits

- **Abstraction of data records** and database operations to single ontology nodes
  - **Simplifies** data identification, extraction and transformation from source systems (e.g. EMRs)
  - Allows **re-use** of created “ETL knowledge” (compared to SQL code)
  - OWL format enables easy **linkage to medical ontologies** (SNOMED, NCIt, ...)

- **Adds** the missing “easy-to-use” **ETL part to the i2b2**
- Comprehensive solution covers
  1. **Access to heterogeneous data**
  2. **“Future-proof”** (ontology-based) data processing
  3. Interactive translational **querying** of clinical data (i2b2)

=> Step forward in implementing “secondary use / single source”
Thank you for your attention!

sebastian.mate@imi.med.uni-erlangen.de

Sebastian Mate would like to thank the “Arbeitskreis Software-Qualität und – Fortbildung e. V.” (ASQF) and the commission of computer science of the University of Erlangen who have awarded this work with the 500€ ASQF diploma prize. Thank you!
Figure 1. All information to perform a data export is described in semantic networks.
Export the latest PSA value from the four outpatient fields. If these are empty, export the PSA value from the extra inpatient field.

Date of examination: 17.12.2009, 01.04.2009, 01.10.2010, 23.03.2010
PSA (in ng/ml): 0.2, 0.3, 0.1, 0.4
PSA: 0.4 ng/ml

Soarian EMR

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

<table>
<thead>
<tr>
<th>Source Ontology</th>
<th>Mapping Ontology</th>
<th>Target Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSA1</td>
<td>PSA2</td>
<td>PSA3</td>
</tr>
<tr>
<td>PSA4</td>
<td>ExtraPSA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GreaterVT:** Returns "True" if Operand 1 is greater than Operand 2; NULL values are allowed. **NoteExists:** 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.