Reutlingen University

DBKDA 2017, Barcelona

Using the Graph-Model for Schema and Data Mapping

Fritz Laux
Prof. emeritus
Reutlingen University
Dept. of Informatics
Reutlingen, Germany

fritz.laux@reutlingen-university.de

© F. Laux

Reutlingen University

Aim of the Talk

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

- *Motivation to use the Graph Model for visualizing schema mapping and data transformation*
- *Contents*
  - Present the Graph Model and with relevant properties for our purpose
    - ⇨ Formally compact, yet sufficient for the target aim
  - Apply the model to typical situations (patterns)
    - ⇨ Show benefits and pitfalls
- ***Research challenges (open questions)***
  - Automate the matching
  - Finding the “best” possible mapping

2 /22
© F. Laux

Reutlingen University

Motivation

- *Increasing number of data sources need to be integrated to…*
  - gain added value (knowledge, insights, predictive analysis)
  - coordinate complex processes (e.g. traffic control, fight epidemic, industry 4.0)

3 /22
© F. Laux

Reutlingen University

Challenges

Data Integration problems[1)]

- Variety of systems/technologies
  - ⇨Incompatible platforms, systems, access technologies[2)]
- Logical and semantic reasons
  - ⇨Different data models, data structure/ representation, synonyms, homonyms[3)]
- Social and administrative hindrance
  - ⇨Data owners fiefdom, data privacy, performance reasons[2)]

4 /22
© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Idea for Solution

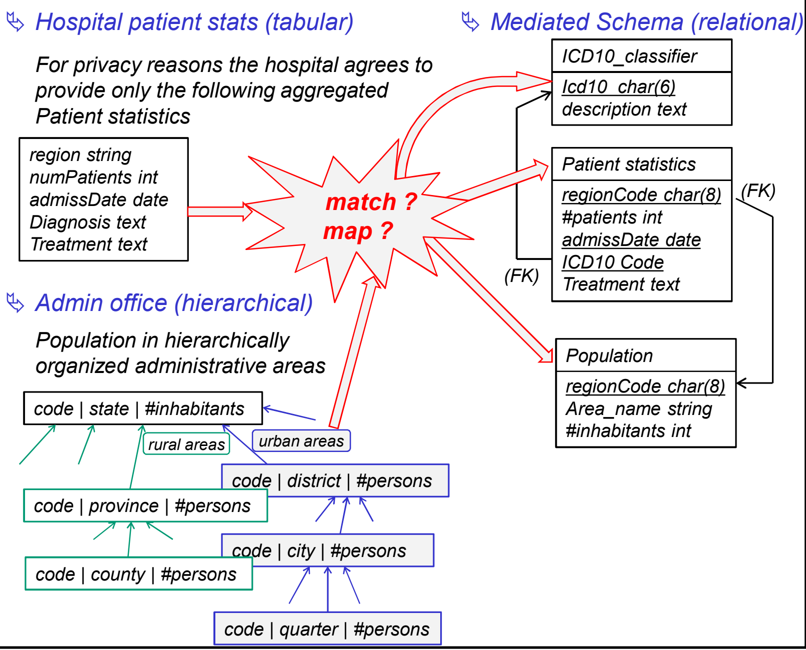
- *Even if we can create transformations to solve these problems, but ongoing maintenance will be difficult since the source systems keep evolving.*
- *There is a strong need to have visual help to understand the impact and interplay of any changes.*
- *The idea: Use the Graph Model (GM) to visualize, formalize, and support the data integration.*
  - Why? (1) GM is flexible and easy to understand
    (2) GM and Category theory allows to check the validity of the integration mappings
    (3) GM visualizes interdependencies

5 /22
© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Data Integration Scenario (Running Example)



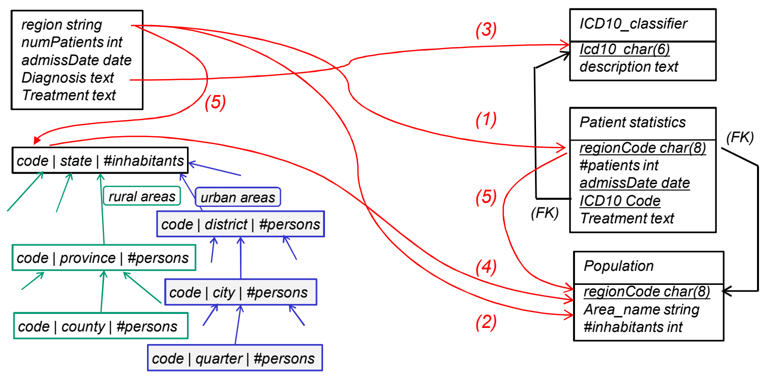
6 /22
© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Data Integration Scenario (Running Example)

Hospital patient stats (tabular) *Mediated Schema (relational)*



Possible problems

(1) Identify matching elements / metadata (e.g. region to regionCode)
(2) Resolve conflicts (e.g. when merging data)
(3) Preserve semantics (e.g. extract ICD10 code from diagnosis)
(4) Transform models and preserve its structure (e.g. mapping hierarchy to relation)
(5) Ensure consistency when multiple mapping paths exist (e.g. from hospital-region to population-regionCode,)

© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Property Graph Model

Model elements

- Nodes (Vertices) ≈ objects
- Lines (Edges) either directed or undirected ≈ related objects
- Properties (of vertices and/or edges) ≈ detail information about objects or relations

Definition: Graph

- A Graph G := (V,E) is a set of Vertices V and a set of Edges E.
- An Edge e ϵ E is defined by the pair of vertices (u,v), with u,v ϵ V, that connect u and v.

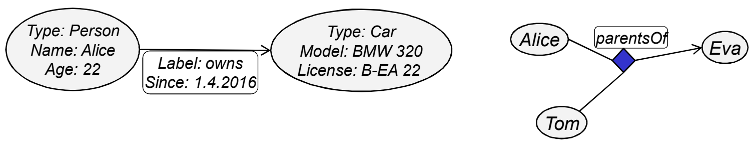
Definition: Property Graph

- A Property Graph PG = (V, E, P) is a Graph where any x ϵ V ∪ E can have a subset $P_x \subseteq P$ of properties (e.g. key-value pairs) attached to x.

Definition: Hypergraph

- A Hypergraph is a Graph G where the edges e can connect more than two vertices.

Property Graph *Hypergraph*



© F. Laux

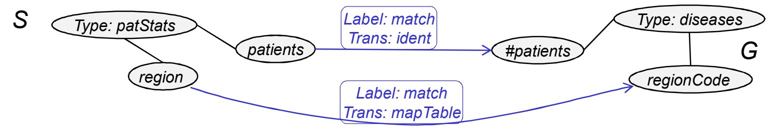
Reutlingen University

© F. Laux

Examples of Graph Mappings (1/2)

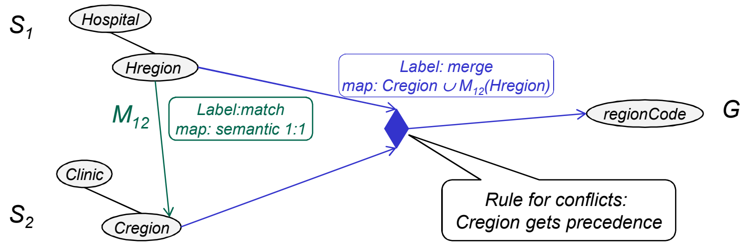
Match   problem (1, 3)

- Given schema S and G. A 1:1- or renaming mapping is called a Match. The mapping preserves the semantics.



Merge   problem (2)

- Given mapping $M_{12}: S_1 \rightarrow S_2$, $G = S_1 \cup S_2$



Reutlingen University

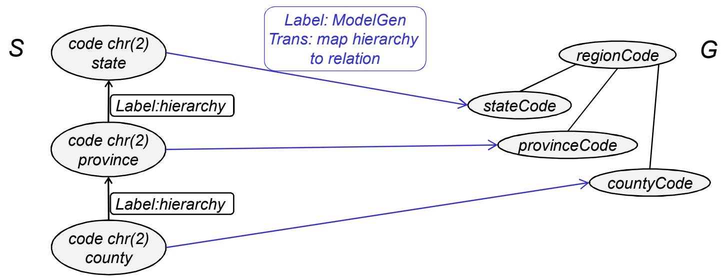
Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

© F. Laux

Examples of Graph Mappings (2/2)

Model Generation   problem (4)

- Given S and G of different meta-models, define mappings to transform S into G
- Goal: preserve semantics as far as possible
- Example: S hierarchy, G relational



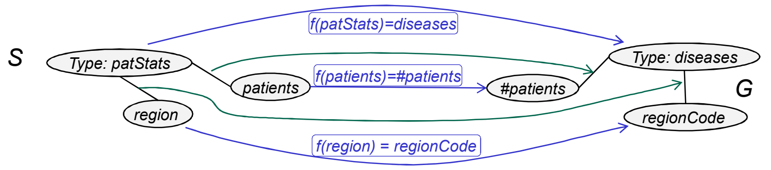
Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Important Graph Mapping Types

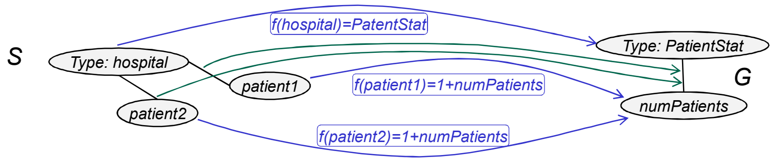
Isomorphism (Edge preserving bijection) problem (3, 4)

- Let $S=(V_1, E_1)$, $G=(V_2, E_2)$
  $f: (V_1) \rightarrow (V_2)$ is bijection and
  $\forall (v_1,v_2) \epsilon E_1 \Longleftrightarrow (f(v_1),f(v_2)) \epsilon E_2$



Homomorphism (Edge preserving map) problem (3, 4)

- $f: (V_1) \rightarrow (V_2)$ is mapping and
  $\forall (v_1,v_2) \epsilon E_1 \Longrightarrow (f(v_1),f(v_2)) \epsilon E_2$



© F. Laux

Reutlingen University

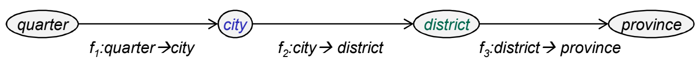
Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Mapping Composition

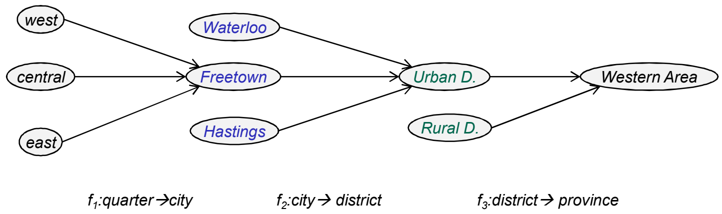
Function Chain

- Let $S_i = (V_i, E_i)$ (i=1,2,…n) be graphs and
  $f_j: (V_j) \rightarrow (V_{j+1})$ $(1 \leq i < n)$ be mappings.
  The composition (or chain) of functions
  $f_k \ldots o\ f_2\ o\ f_1$ (k< n) is defined as $f_k(\ldots f_2(f_1(v_1))\ldots)$ $(\forall\ v_1 \epsilon\ V_1)$

- Example:



- Each quarter is mapped to the city it is located, the city in turn is mapped to its district and finally the districts is mapped to the state.



© F. Laux

Reutlingen University

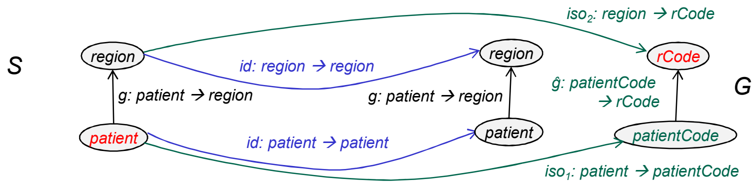
Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

13 /22
© F. Laux

Commutative Mappings

Commutative Mapping *problem (5)*

- A function chain is called commutative if and only if $f_2 \circ f_1 = f_1 \circ f_2$, i.e. $f_2(f_1(x)) = f_1(f_2(x))\ \forall x \in dom(f_1)$
- Example: $g \circ id = id \circ g$ (and more general $\hat{g} \circ iso_1 = iso_2 \circ g$)



- For a consistent mapping from patient to rCode it is irrelevant if the projection g to region is done first or the isomorphic mapping iso_1 to patientCode.

Desirable Mappings

- Projection π, Homomorphism hom, and Isomorphism iso are good candidates for commutative mappings. (e.g. $\pi \circ iso = iso \circ \pi$)

Reutlingen University

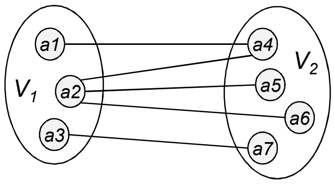
Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

14 /22
© F. Laux

Bipartite Graph and Graph Matching

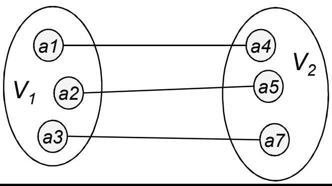
Bipartite Graph

- Let $G = (V,E)$ with $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$. If there are no edges within V_1 and V_2 then G is bipartite.
- Example 1:



Graph Matching *quality criteria for problem (1)*

- Let G be a bipartite Graph. A matching is a subset of edges where no two edges share an endpoint (node)
- Maximum matching = maximum number of vertices are matched
- Perfect matching = all vertices are matched
- Example 2:

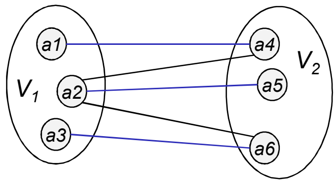


Reutlingen University

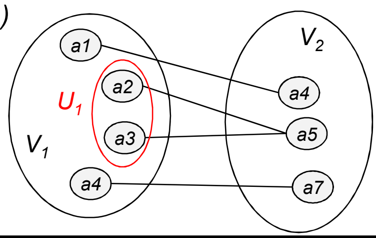
Theorem of Hall (Marriage Theorem)

- *Let $G = (V1 \cup V2, E)$ be a bipartite Graph. In G exist a perfect matching if $\forall U_1 \subseteq V_1$: $d(U_1) \geq |U_1|$.*
  $d(U_1) := |\{v \in V_2| u \in U_1 \wedge (u,v) \in E\}|$
  general criteria for data integration coverage/completeness

- *Example 3 (perfect match)*
  - All subsets U_1 of V1 have $d(U_1) \geq |U_1|$.
  - (a1,a4), (a2,a5), (a3,a6) is the only possible perfect matching.



- *Example 4 (no perfect match)*
  - Subset U_1={a2, a3} has $d(U_1)$=|{a5}|=1, but $|U_1|$ = 2.
  - perfect matching is not possible.



© F. Laux

Reutlingen University

Integration Framework

1. *Take source models and target model.*
2. *Make all data elements explicit (nodes) that must be matched or mapped.*
3. *Define a bipartite Graph with all elements from step 2 (sources = V_1, target = V_2).*
4. *Identify semantic matches between sources and target nodes by making edges.*
5. *Define mappings by giving transformation rules or formulae as properties.*

© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Checking rules for integration completeness

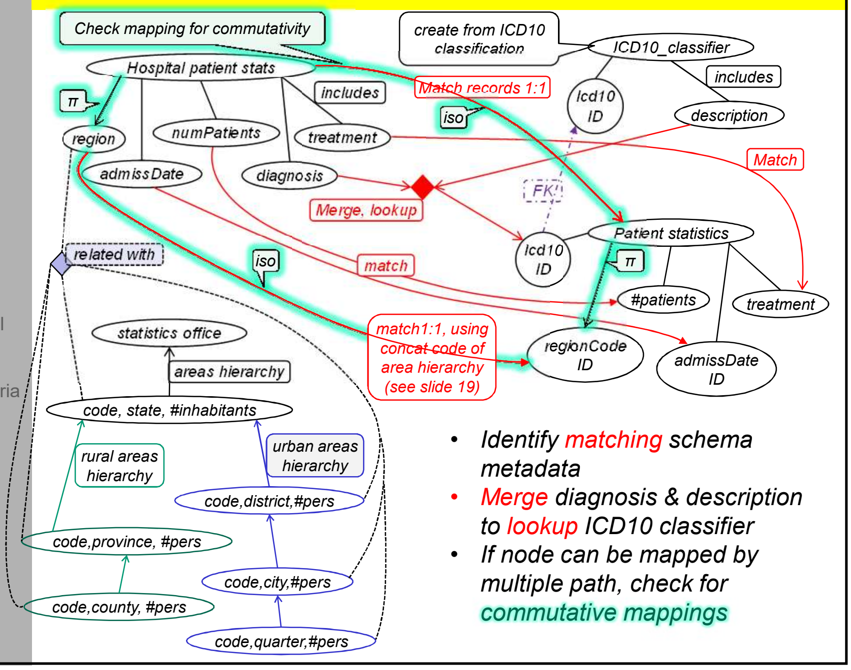
- *Use theorem of Hall to check for integration completeness resp. coverage*
- *Add relations in source and target models to check formal consistency.*
  - If a target node can be reached by more than one path, make sure that the mappings are commutative.
  - When the mapping is an aggregation, than the mapping should be a homomorphism.
  - If the mapping is an isomorphism, the mapping is lossless.

17 /22
© F. Laux

Reutlingen University

Aim
Motivation
Challenges
Idea
Example
Graph Model
Mappings
Quality Criteria
Framework
Solution
Conclusion

Putting everything together (Running Example 1st part)

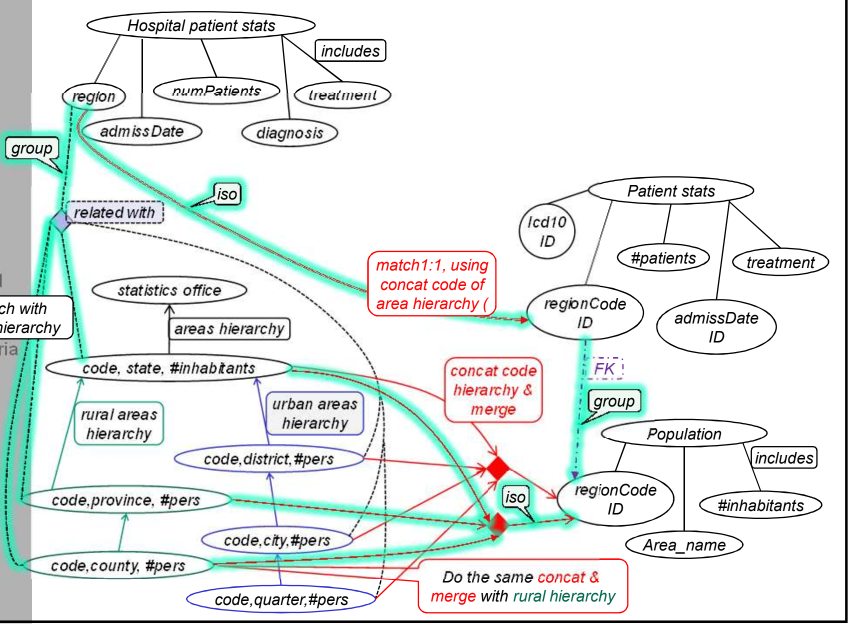


- *Identify matching schema metadata*
- *Merge diagnosis & description to lookup ICD10 classifier*
- *If node can be mapped by multiple path, check for commutative mappings*

18 /22
© F. Laux

Reutlingen University

Putting everything together (Running Example 2nd part)



© F. Laux

Reutlingen University

Lessons learned

- *Use the GM on the data/object type level*
  - Use different colors for node/edge types
  - Only use GM for instances if special details need to be visualized (e.g. aggregation of instances of the same object type)
- *In real world scenarios the GM tends to be confusing*
  - Model partial data structures separately
  - In extreme cases use only 1 source element and model all edges from and to this element only. This visualizes all influencing factors and dependencies.
- *Some GM theorems allow (formal and automated) quality checks of the data integration*
  - Theorem of Hall: coverage/completeness check
  - Commutative mappings: consistency checks
  - Hypergraph links need detailed description for mappings

© F. Laux

Reutlingen University

References

- 1) *A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration, Morgan Kaufmann, Elsevier, 2012, ISBN: 978-0-12-416044-6.*
  - cover comprehensively the different concepts of data integration using conjunctive queries as formal representation.
- 2) *M. Crowe, C. Begg, F. Laux, M. Laiho, “Data Validation for Big Live Data”, DBKDA 2017, pp. 30 - 36.*
  - Propose a REST and ReadCheck for uniform relational access and query
- 3) *Ch. Pinkel et al., “IncMap: A Journey towards Ontology-based Data Integration”, in Mitschang et al. (ed.) BTW 2017, pp. 145 – 164.*
  - IncMap can detect and leverage semantic-rich patterns in the relational data sources and use them for data integration.

21 /22

© F. Laux

Reutlingen University

Discussion

22 /22

© F. Laux