Detecting Quality Problems in Research Data: A Model-Driven Approach

Arno Kesper
arno.kesper@uni-marburg.de

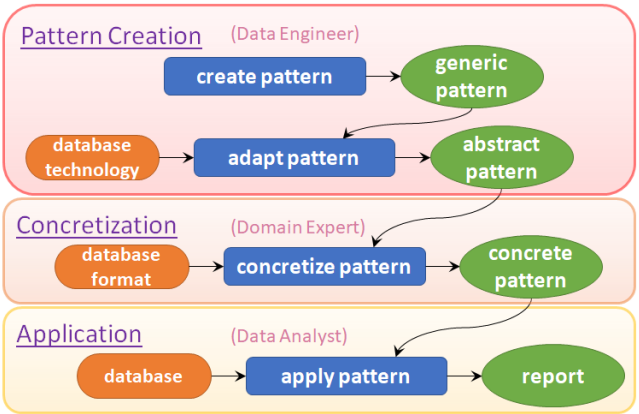
Viola Wenz
viola.wenz@uni-marburg.de

Gabriele Taentzer
taentzer@uni-marburg.de

Introduction

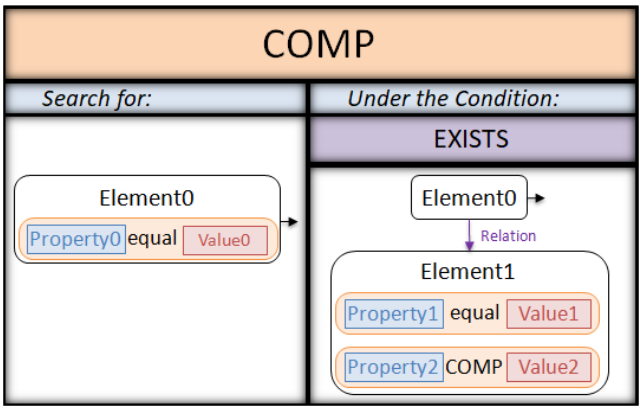
- Research data quality is essential for scientific progress.
- Challenge: Variety of DB technologies used
- ▷ *Model-driven* approach - *generic* wrt. DB technology - based on *anti-patterns* for data quality problems

Approach



Pattern	first-order logic expression over graphs
Generic pattern	independent of DB technology and format, parameterized
Abstract pattern	adapted to specific DB technology (e.g. XML), independent of DB format, parameterized
Concrete pattern	tailored to a concrete quality problem and DB format, parameters specified, translatable to query language
Report	problem occurrences (i.e. pattern matches) in the chosen DB

Example



The pattern detects *interval violations* of values in a specific structural context.

- Element identification: specific property has specific value
- Element0 returned if related Element1 exists and satisfies an additional comparison

Concretisation

- XML- adapted abstract pattern omitted
- Specification of relations omitted
- Concrete pattern detects “architect” elements with related “birthyear” elements with value greater than 2020

Property0, Property1	*TAG*
Value0	*“architect”*
Value1	*“birthyear”*
Property2	*DATA*
Value2	*2020*
COMP	*>*

Evaluation

RQ1: How far does our approach enable the detection of quality problems in research data?

- Strength: Detection of structural problems
- Integration with further analysis techniques could increase expressiveness

RQ2: What is the query response time of the problem detection?

- Based on the application of 43 patterns to two large cultural heritage databases
- Frequent application possible for most patterns

Patterns	Runtime
70 %	< 10 s
80 %	< 20 s
90 %	< 4 min
95 %	< 6 h

Conclusion

- *Model-driven* approach promising to develop tooling for data quality analysis independent of technologies and formats
- Higher level of abstraction than other pattern-based approaches to data quality analysis
- Proof-of-concept implementation for XML
- *Future work* includes empirical evaluation of the GUI and application to other kinds of (research) data

KONDA
Philipps Universität Marburg
Deutsches Dokumentationszentrum für Kunstgeschichte Bildarchiv Foto Marburg
Bundesministerium für Bildung und Forschung