Anne Koziolek, Ina Schaefer, Christoph Seidl (Hrsg.): Software Engineering 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021 61

Detecting Quality Problems in Research Data: A Model-Driven Approach[1]

Arno Kesper[2], Viola Wenz[2], Gabriele Taentzer[2]

Abstract: The quality of research data is essential for scientific progress. A major challenge in data quality assurance is the localisation of quality problems that are inherent to data. Based on the observation of a dynamic shift in the database technologies employed, we present a model-driven approach to analyse the quality of research data. It allows a data engineer to formulate anti-patterns that are generic concerning the database format and technology. A domain expert chooses a pattern that has been adapted to a specific database technology and concretises it for a domain-specific database format. The resulting concrete pattern is used by a data analyst to locate quality problems in the database. As a proof of concept, we implemented tool support that realises this approach for XML databases. We evaluated our approach concerning expressiveness and performance.

The original paper has been published at the International Conference on Model Driven Engineering Languages and Systems 2020 [KWT20].

Keywords: Data quality; Model-driven development; Pattern matching

1 Introduction

As scientific progress highly depends on the quality of research data, there are strict requirements for data quality coming from the scientific community. Relevant quality dimensions include consistency, completeness and precision. In a qualitative study on cultural heritage data we observed a variety of data quality problems. Examples include imprecise, redundant and semantically incorrect data. A major step to data quality assurance is to analyse the inherent quality problems. Due to the dynamic digitalisation in specific scientific fields, different database technologies and data formats may be used. In the digital humanities, for example, a shift from relational to XML and further to graph databases can be observed. To cope with this challenge, we present a model-driven approach to data quality analysis. Given a large variety of data quality problems, which may have various concrete forms, a *model-driven approach* is promising to develop technology- and format-independent concepts and tooling for data quality analysis.

[1] This work was partially funded by the German Federal Ministry of Education and Research (BMBF) grant 16QK06A.

[2] Philipps-Universität Marburg, Marburg, Germany, {arno.kesper, viola.wenz, taentzer}@uni-marburg.de

doi:10.18420/SE2021_19

2 Approach

Our model-driven approach to data quality analysis is based on the observation that many quality problems show anti-patterns. In contrast to related approaches, it allows data engineers to specify parameterised anti-patterns for data quality problems that are *generic* concerning the underlying database technology and format. Such a generic pattern can be adapted to several database technologies, resulting in *abstract* patterns. Depending on the database technology this can be done (semi-)automatically by a data engineer. A domain expert chooses an abstract pattern as template and concretises it to a domain-specific database format and to a concrete quality problem. The resulting *concrete* pattern is then automatically translated into a corresponding query language. A data analyst can apply the concrete pattern to localise occurrences of the quality problem in a database. The data analyst can then initiate an improvement process.

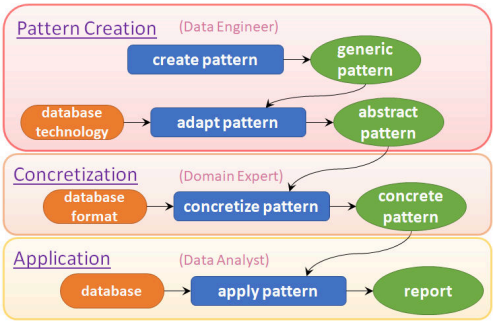

Fig. 1: Workflow of pattern creation and application

The core of the approach is a metamodel for patterns. It currently supports generic, XML-adapted abstract and concrete patterns. The metamodel is designed to allow extensions for further database technologies. Patterns are defined as first-order logic conditions over graphs. We use directed graphs to interpret data independently of the database technology. Our implementation includes a mapping between graphs and XML data as well as a translation of XML-adapted concrete patterns to XQuery. Mappings for other database technologies are outlined in our paper.

The evaluation of our approach based on cultural heritage data revealed that its strength lies in detecting structural problems. The expressiveness can be expanded by integrating further techniques for quality analysis, such as similarity metrics.

Our overall goal is to develop a framework for quality assurance of research data, where the detection of quality problems is the first essential step. In general, there is a need for powerful tools that analyse the quality of data. Data quality assurance is also of particular importance in the context of data-intensive software systems, which have gained interest in recent years. Hence, this topic affects not only data engineering but also software engineering.

Bibliography

[KWT20] Kesper, Arno; Wenz, Viola; Taentzer, Gabriele: Detecting Quality Problems in Research Data: A Model-Driven Approach. In: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems. MODELS '20, Association for Computing Machinery, New York, NY, USA, p. 354–364, 2020.