Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 104 (2017) 138 – 145

Procedia Computer Science

ICTE 2016, December 2016, Riga, Latvia

Executable Data Quality Models

Janis Bicevskis[a,]*, Zane Bicevska[b] , Girts Karnitis[a]

[a] *University of Latvia, Riga, Latvia,* [b] *DIVI Grupa Ltd, Riga, Latvia*

Abstract

The paper discusses an external solution for data quality management in information systems. In contradiction to traditional data quality assurance methods, the proposed approach provides the usage of a domain specific language (DSL) for description data quality models. Data quality models consists of graphical diagrams, which elements contain requirements for data object's values and procedures for data object's analysis. The DSL interpreter makes the data quality model executable therefore ensuring measurement and improving of data quality. The described approach can be applied: (1) to check the completeness, accuracy and consistency of accumulated data; (2) to support data migration in cases when software architecture and/or data models are changed; (3) to gather data from different data sources and to transfer them to data warehouse.

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016

Keywords: Data quality; Domain-specific languages; Business process execution

1. Introduction

Data quality can be defined as the degree to which a set of characteristics of data fulfills requirements[1]. Examples of characteristics are: completeness, validity, accuracy, consistency, availability, and timeliness. Requirements are defined as the need or expectation that is stated, generally implied or obligatory.

The ISO 9001:2015 standard considers data quality as a relative concept, largely dependent on specific requirements resulting from the data usage. It means the same data can be of good quality for one usage and completely unusable for another. For instance, to determine a count of students in a high school, only the status of students is of interest, not other data like students' age or gender.

* Corresponding author.
E-mail address: Janis.Bicevskis@lu.lv

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016
doi:10.1016/j.procs.2017.01.087

To evaluate the data quality for the specific usage, the requirements for data values must be described. The descriptions should be executable, as the stored data will be "scanned" and it's compliance to requirements will be checked. It should be emphasized that many conditions and requirements can't be checked during the data input as they are dependent on values of other data objects that are not entered into data base yet. For instance, at the time of student's enrollment not all information about his/her financial obligations is available and/or entered in the database as the data could be received later. This is the reason why high-quality data in practice occurs rarely. Therefore a topical issue is to analyze and to evaluate the data suitability for specific usages or tasks.

The described problem is topical since over 50 years. Traditionally, developers include data quality checking functionality into information systems by two separate types of control:

- Syntactic control – values of data objects are checked locally within one record (compliance of input data with the syntax), it also includes compatibility control on linked data of the record
- Semantic/ contextual control – checking whether the newly entered data is compatible with data previously entered and stored in database. The semantic control should be repeated every time when new values of data objects' attributes are entered or existing ones edited

The proposed approach intends creating of specific data quality model for each information system. The model is described by using means of a domain specific language, and it lets clearly define requirements for data objects attribute values and compatibility. The data quality model is executable: both the syntactic and the semantic controls are performed. The approach provides the possibility to use the data quality model for measurement and evaluation of data quality as well as for checking of the data to be imported in a data warehouse.

The paper deals with following issues: an identification of the problem to be solved and a brief description of basic ideas (Section 2), an overview about the related research (Section 3), a description of the proposed solution (Section 4), and a discussion about the possibilities to use data quality models in practice (Section 5).

2. Statement of the problem and ideas for solutions

The data quality management solution proposed by the authors of this paper is based on the usage of executable domain specific languages (DSL). Data quality management is complicated because data in the database is accumulated gradually, according to the flow of received documents or recorded events, and it does not always correspond to the sequence of events in the real life. Data about events entered into database can also be incomplete and in a logically wrong sequence. This research seeks the mechanism supporting the data quality evaluation even when only part of data objects is fully entered into the database.

Traditionally, data quality in databases is provided and controlled by setting constraints on the attribute values and relations that may be stored in the database, for example, by using the Object Constraint Language (OCL)[2]. The OCL allows you to specify in detail constraints on data/ attribute values to be stored in the database, and it preserves the database from an incorrect content. Nevertheless the OCL is not able to solve the problem formulated above – a gradual accumulation of data allowing incompatibility awhile (temporary) but guaranteeing correct data in definite checkpoints (a complete data set about a data object entered in database).

The proposed solution supports the gradual filling in the data with data quality controls in specific points (steps in the data quality model). Data quality models are described by using DSL that is suitable for description of data objects attribute values. A data quality model consists of graphical diagrams which are created in the specific DSL and resemble the traditional flow charts (in detail see in the next chapters). Furthermore, not one universal, the "right" DSL is proposed but a whole language family. The editor building platform DIMOD[3] ensures the defining of DSL syntax and creating of the editor suitable for editing of diagrams in the DSL. Data quality controls are the elements of the DSL. The models described by the DSL become executable as soon as the controls are implemented in software routines or by SQL statements.