# A META-MODEL FOR DATA QUALITY MANAGEMENT SIMULATION

(Research-in-progress)

**Boris Otto, Kai M. Hüner**
University of St. Gallen, Switzerland
boris.otto@unisg.ch, kai.huener@unisg.ch

**Abstract:** Data quality management initiatives could both help to prevent the occurrence of data defects and repair their effect. While such initiatives can reduce overall costs, they also cause costs for their development and implementation. Therefore, the overall aim is not to improve data quality by any means, but to ensure cost-efficiency. The paper proposes a meta-model for simulating data quality management, which can be used for planning of cost-efficient initiatives.

## MOTIVATION AND SIMULATION APPROACH

Companies use data in their operating business, e.g. when they produce goods, or when they render services. The quality of the data used is always a critical factor regarding the outcome of business processes (e.g. process lead time, customer satisfaction, product quality). In order to be able to work with data of good quality, data quality (DQ) requirements need to be clearly defined. When they do so, companies need to be aware of the fact that both using poor data and creating good data brings about considerable costs [1].

Initiatives for data quality management (DQM) could help prevent or reduce occurrence of data defects. While such initiatives can reduce overall costs, they also cause costs for their development and implementation. Therefore, the overall aim is not to improve DQ by any means, but to ensure cost-efficiency when implementing DQM initiatives. Basically, the budget of DQM initiatives is limited by costs (i.e. monetary losses) that are caused by poor data without the expected effect of the DQM initiatives (cf. Figure 1.a). Taking a systems theory perspective, Figure 1.b illustrates these interrelatedness as a closed loop aiming at reducing costs arising from business problems and DQM initiatives. In this model, the system is constituted by business processes and by data used for doing operating business (including potential data defects and business problems), the sensor determines DQ and the costs arising from business problems, and the controller determines the scope and character of DQM initiatives.

Literature covers costs of poor DQ [2] and the classification of DQ and DQM costs [1]. However, the effort to ensure context specific data quality and the balance between related costs and benefits are rarely

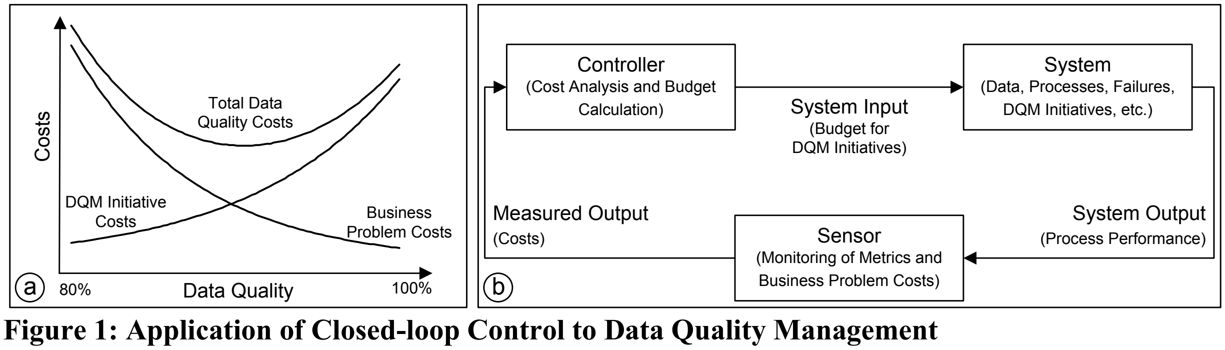

**Figure 1: Application of Closed-loop Control to Data Quality Management**

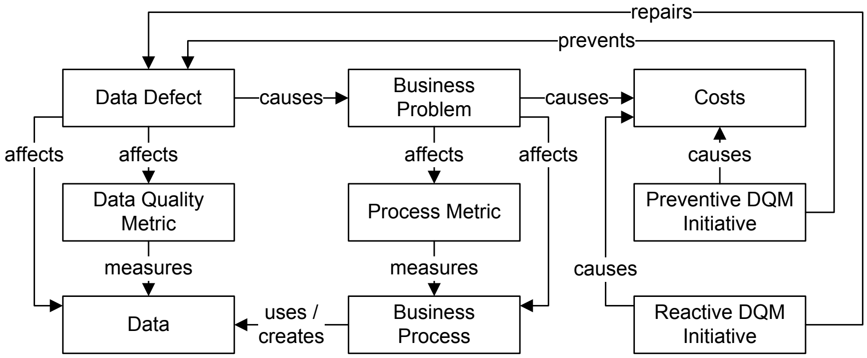

**Figure 2: Entities and relations for Data Quality Management Simulation**

investigated. Lee et al. [3] suggest the theory of real options for DQM cost/benefit analysis and refer to simulation models as a general solution. Orr [4] applies system theory to data use, but does not consider DQ or DQM costs.

Since neither the occurrence of a business problem nor its consequences or the effect of a certain DQM initiative can be taken for granted, the controller's (cf. Figure 1.b) decision in favor of or against a certain DQM initiative is always based on assumptions and estimations resulting from previous experiences or expert knowledge. To support such estimation, Figure 2 shows a meta-model allowing simulation of the relations specified in terms of calculating the system's output based on assumptions regarding the system (e.g. estimated probability of occurrence of data defects, or estimated impact of data defects on process performance).The overall aim of such a simulation is not to allow for automatic control of DQM initiatives or prediction of process performance, but to explicate assumptions (e.g. estimated impact of a certain DQM initiative) and support decision-making of the controller (which in real business scenarios is not an information system but typically a person responsible for DQM).

## Further Research

Challenges in using the approach mainly refer to identifying realistic and quantified cause-effect chains between data defects, business problems and costs (cf. Figure 2). In real-world scenarios necessary information (e.g. frequency of occurrence of a certain data defect) must be gathered from various sources in a company (e.g. by interviews and expert assessments), and the data gained from this must be mapped onto the meta-model's elements (e.g. parameters for distribution of a certain data defect).

One next step to be taken in the research process may be to verify the results of such DQM simulation against companies' actual development. Even if a simulation model takes up selected aspects only, it aims at describing realistic phenomena as well as possible. Such verification would require having real measuring values for DQ metrics and process metrics over a certain period of time and proving the effect of DQM initiatives for such real measuring values.

## References

[1] Eppler, M. J., Helfert, M. "A Classification and Analysis of Data Quality Costs," in *Proceedings of the 9th International Conference on Information Quality*, Cambridge, 2004, pp. 311-325

[2] Fisher, C. W., Kingma, B. R. "Criticality of data quality as exemplified in two disasters." *Information & Management,* 39 (2). 2001. pp. 109-116

[3] Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. *Journey to Data Quality*. MIT Press. Boston, 2006

[4] Orr, K. "Data Quality and Systems Theory," in *Proceedings of the 1996 International Conference on Information Quality*, Cambridge, 1996, pp. 15-29