IQUATE: MODELING THE CAUSES AND IMPACTS OF INFORMATION QUALITY

(Research-in-Progress)

Jochen Kokemüller
Fraunhofer IAO, Germany
Jochen.Kokemueller@iao.fraunhofer.de

Abstract: Information quality is a vital asset to organizations. Yet, no modeling technique exists that allows the structured analysis of causes of low information quality and its impacts on the organization's performance. This research aims at filling this gap by providing the parsimonious modeling technique IQUATE for the causes and impacts of low information quality. To bridge the gap between social and technical perspectives on information quality we based IQUATE on the Work System method. We show IQUATE's feasibility in the analysis of a real world scenario. In this we follow the design science approach.

Key Words: Data Quality Modeling, Information Quality Modeling, Information Product, Work System

INTRODUCTION

Information Quality has severe impacts on organizations, processes or more general work practices. Even disasters have been reported [11]. Organizations suffer high economic losses due to low information quality. Ballou et al. [2] therefore define Information Products and assign holistic manufacturing and quality assurance processes equally to that of physical products.

Organizations strive to improve the quality of information products in data quality initiatives. There is considerable knowledge among information quality professionals about the key factors to an initiative's success. Among the methodologies for data quality improvements the alteration of processes, decision criteria, and responsibilities are a high priority [4]. It has even been shown, that only the alteration of those factors in a data quality initiative has a long term effect, while projectized data quality efforts have no significant long term effect on perceived data quality [13].

Low information quality has severe impacts on an organizations performance. Especially the quality of master data is multiplied into transactional data [15]. Several modeling techniques have been developed, that aim at modeling optimized work practices related to data from the data producers, consumers, and custodians. However, before modeling optimized work practices one needs clear understanding, where data quality issues arise and what their impact on the organizations performance is.

This is an important prerequisite, as it allows:

- The structured description of data quality problems.
- The selection and prioritization of organizational enhancements and information quality initiatives.
- It further provides explanatory capabilities that help in gaining information quality sponsors from senior management.

The analysis of causes and impacts of low information quality may not restrict itself to a technical perspective as this would exclude the social interactions. Alter [1] defines the concept of work system, as a view on work occurring through a purposeful system. Work hereby is defined as effort applied to accomplish something. This allows a general view on organizations, where processes are not always executed in a well-defined way, but are often executed unstructured, unplanned, or uncontrolled. The definition of ideal business processes or ideal data flow charts may therefore only address parts of the real problem space. Nevertheless, modeling helps in analyzing a problem space. Yet, for the analysis of information quality in real world situations the focus on processes cannot suffice, it is necessary to focus on work systems.

Analyzing work systems and their relationship to information products, it is obvious that this relationship is bidirectional. While work systems and here executed work practices have an effect on the quality of information products, the quality of information products has an impact on work systems. We therefore propose in this contribution a modeling technique that helps in the analysis of this bidirectional relationship with a *cause model* and an *impact model*.

The paper is structured as follows: We start with the discussion of related work and develop then models for information quality analyses as cause and impact models of low information quality. We then discuss our implementation of the model and show its feasibility in a three step analysis of a situation common to the direct order industry. Therefore, we follow the classic approach of design science-oriented research as we first develop an artifact and we second provide an evaluation of the artifact.

RELATED WORK

Information quality is a multifaceted concept. Wang and Strong [25] identify 15 dimensions of information quality. These dimensions help in understanding the information's "fitness for use". The latter definition – although not very useful in practice – highlights the goal of information quality improvements: to improve the information quality to a suitable level as required by its use cases. The term information often refers to processed and meaningful data present in information systems; data itself only refers to raw facts. In spite of that in the context of information and data quality these two terms are commonly used interchangeably as it is often impossible to distinguish between the two [19]. We adhere to that practice and use these terms interchangeably.

For the modeling of causes and impacts of low information quality a fundamental understanding of information systems is necessary. Information Systems (IS) are used to create, read, update, and delete information. One may therefore follow the viewpoint of software developers or architects. According to Lyytinen [18] two types of modeling exist: a formalistic and a functional approach. Formalistic modeling aims at a formal, non-algorithmic specification of a system. To this end fundamental data structures, such as entities, attributes and relations are modeled. This allows the modularization of the system into static elements. Further abstractions and restrictions can be modeled. The dynamic of IS is modeled by triggering events and actions that result in alterations of the data.

Functional modeling allows a process oriented descriptions of systems. Here activities are put in relation to each other and to consequences on data. This results in a system with input and output relations to other systems.

These two fundamental modeling approaches build the basis for current technical modeling tools such as UML [22]. Nevertheless, for the analyses of the causes and impacts of low information quality these technical perspectives are insufficient. While UML allows the modeling of interactions of users with IS, it focuses on the IS. Yet, the core concept of Information Quality, the information's fitness for use, does not focus on the IS, but on stakeholders as data collectors, data custodians, and data consumers [17].

Therefore Bostrom and Heinen [6] take the social aspects into account by designing two subsystems: the social and the technical subsystem. By this "socio-technical" perspective the subsystems are evaluated separately and are then optimized respecting the requirements of the other subsystem. This approach helped in the development of the understanding of interweaved dependencies between the social and technical subsystems. In principal the socio-technical perspective is apt for the modeling of causes and impacts of low information quality, but as Land [16] recognized: "socio-technical methods focus on design of work systems to improve the welfare of employees. The prime aim of redesigning work systems is the improvement of the quality of working life". Yet, while information quality may be related to the quality of working life it cannot be reduced to it. The socio-technical perspective would therefore need improvements to include additional dependencies.

The Work System theory aims at understanding, analyzing, and improving work systems in organizations [1]. It is not necessary, that information technology plays an essential role in it. The elements of the work system are shown in Figure 1. Its building blocks are Work Practices that interact with its participant, information, and technologies. Work systems do not stand separated; they are surrounded by the environment, strategies, and the infrastructure in order to provide the products & services demanded by the customers. The Work system theory describes the interaction between the social and the technical perspectives in a sufficient fine granularity to build the basis for information quality analyses.

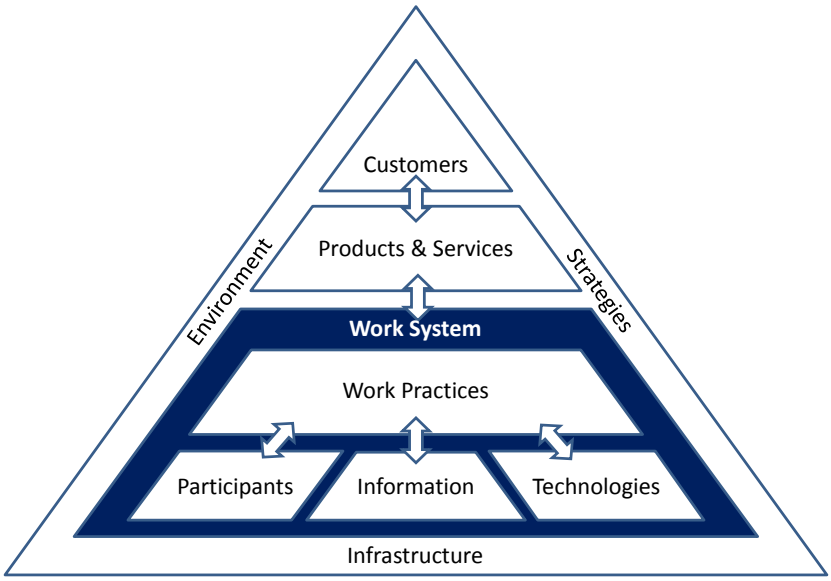


Figure 1: The Work System Framework [1]

To the supply chain of physical goods several modeling techniques are known, applying them to information products resulted in a modeling technique with similarities to data flow diagrams, called IP-MAP [17]. It is used to depict the creation and usage of information products. To this end numerous elements are provided that allow the modeling of the ideal handling of an information product. IP-MAPs thus do not allow the modeling of the influence (positive or negative) of certain activities on the resulting information quality.

To IP-MAPs a technical adoption based on UML, IP-UML, has been developed [23]. IP-UML allows that quality requirements are related to "quality data" by "quality associations", yet still the alteration of the quality is not modeled. Moreover, this association does not take into account the perspective of a user, as a use case agnostic set of quality requirements is modeled.

Other approaches understand data quality as a static metadata problem [5], simply analyze the deterioration of data quality as a risk using Ishikawa diagrams [24], or focus on certain user groups [7].

We therefore design the modeling language IQUATE following the design science approach. Design science research contributions present novel IT-artifacts and suitable evaluation approaches that address the artifact's appropriateness to contribute to the problem's solution [21]. These two facets of rigorous design science-oriented research contribute to the foundations and the methodologies pool of information systems research, i.e. they contribute to its knowledge base [12].

A MODEL FOR INFORMATION QUALITY ANALYSES

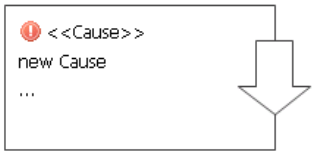
The purpose of IQUATE is to direct attention to causes and impacts of problems with data quality. As these may be of very punctual and distributed nature in an organization or IS landscape, it is designed to specifically highlight the punctual deficiencies, while at the same time providing an overview of major causes and impacts.

Diagrams modeled using IQUATE are therefore intended to solemnly focus on one actual situation. Thus focus the attention on the analysis of one isolated cause or effect avoiding information overload. For each actual situation a separate diagram should exist. To generate comprehensive reports, model elements must be reused between diagrams. In this the interaction and relationship between actual situations is articulated.

Cause Model

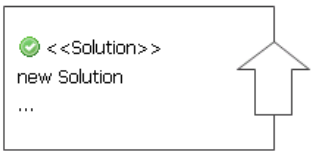
The *cause model* is used to model causes of low information quality. The cause model allows the information quality related analysis of the relationship between actions and information.

Often business processes or activities are used to describe the causes for low information quality. As business processes and activities may be highly complex their relationship to information quality is of insufficient precision. We therefore use actions as modeling element that infer low information quality.

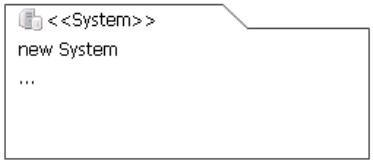


Actions on data are important reasons for low information quality. Actions as collection, update, maintenance, or deletes may result in lowered information quality. Likewise the absence of these actions may result in decay of data, thus in lowered information quality. Therefore we design an action as a central element in the cause model.

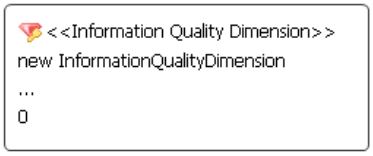
Actions cause changes in information quality, yet they are themselves not without reason. These reasons may be of various natures, such as inadequately lived or designed processes, misleading incentives, unfitting responsibilities or organizational structures, insufficient technical resources, inappropriate user interfaces, and much more. To achieve a parsimonious model we generalize these as *causes*. Causes can be connected to actions.



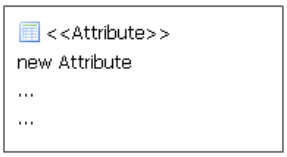
Not only the root causes for lowered information quality should be analyzed and modeled, but also which solution would provide reasonable countermeasures. We therefore design *solutions* as parsimonious model elements that may be connected to actions. Additionally solutions can be connected to causes, to model a solution to a particular cause. Solutions are equally to causes meant as verbatim descriptions.



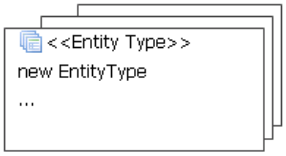
Actions are executed in information systems. For comprehensibility we designed them as *systems*.



By actions only specific information in certain information quality dimensions is affected. Furthermore, an action may have different severities, depending on the information quality dimension. We therefore design the relationship between actions and systems as a tertiary relation, that is directly associated with its impact on a specific *information quality dimension*. For parsimony reasons we do not propose a certain set of IQ dimensions nor metrics for their measurements, but only design a verbatim model element with a severity score associated to it.



Actions usually intervene with specific attributes accessible through a system. We therefore design *attributes* and *entities*. Here entities are connected to systems through attributes, no direct connection is designed.



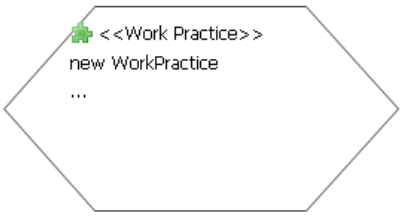
By now we designed the effects of actions on information. Causes of low information quality may also occur due to low schema quality [20]. Therefore, we allow the cause and solution elements to be directly connected to entities. This can be used to model shortcomings in schemata.

Impact Model

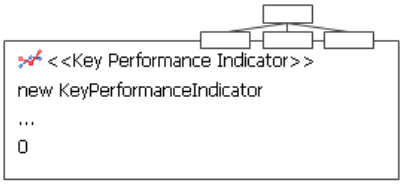
The impact model is intended to model the impacts of (low) information quality. In this the relation between information associated with a particular value of information quality and business outcomes is established. Business outcomes are measured by key performance indicators (KPI) that are influenced by work practices and information products.

The impact model shares four model elements with the cause model: Causes, Solutions, Entity Types, and Information Quality Dimensions. Therefore their definitions are omitted here.

While the cause model took a technical perspective, the impact model necessarily needs to convey a business related perspective. As the cause model evolved around actions, the impact model is designed around Information Products. Consequently, the model does not focus on attributes and systems, but on *Information Products* (IP). This draws the attention to value associated with high quality information and the return on investment it generates likewise to physical goods. Like physical goods, information products may cause expenses and generate revenues. Yet, while this relationship exists, it is often tedious to analyze. Thus we refrain from explicitly modeling expenses and revenues, but model the impact of information products on work practices and KPIs moderated by information quality dimensions.



Work Practices are defined by Alter [1] as structured or unstructured activities eventually organized in business processes. They do not refer to ideal processes, but lived practices. This is important, as the modeled impact on the business outcome should not depend on ideal, but on lived practices.



Key Performance Indicators (KPIs) are used to model the impacts on the business outcome. KPIs can be organized in a hierarchical structure. Thus specialized KPIs with directly analyzable relations to Information Products or Work Practices may be modeled that than aggregate to higher order constructs. For parsimony no specific set of KPIs is proposed by the modeling technique.

IMPLEMENTATION & APPLICATION

IQUATE was implemented as an Eclipse Plugin using the Graphical Modeling Framework (GMF) [8] which is based on the Graphical Editing Framework (GEF) [10] and the Eclipse Modeling Framework (EMF) [9]. The reports are generated as a model to text transformation using the Modeling Workflow Engine (MWE) and XPand.

The application of IQUATE is demonstrated on an example of the direct order industry. In this we explain a methodology that can be applied using IQUATE. In step 1 the static situation of Information Products is modeled. Step 2 than focuses on causes for low information quality and finally in step 3 the impacts of low information quality are modeled.

Step 1: Analysis of Information Products

The approach is to throw spotlights on actual situations, thus analyzing in one diagram only one actual situation. To achieve meaningful results, connections between these isolated diagrams need to be established. This is achieved by the reuse of model elements. This allows the generation of reports covering multiple actual situations. The most important model element that may create such an umbrella spanning all cause and impact diagrams is the information product. In the first step therefore a static view of information products is modeled. Figure 2 depicts how the Information Product "Packing List" is assembled out of two entities: "Shipping Address" and "Order Position".

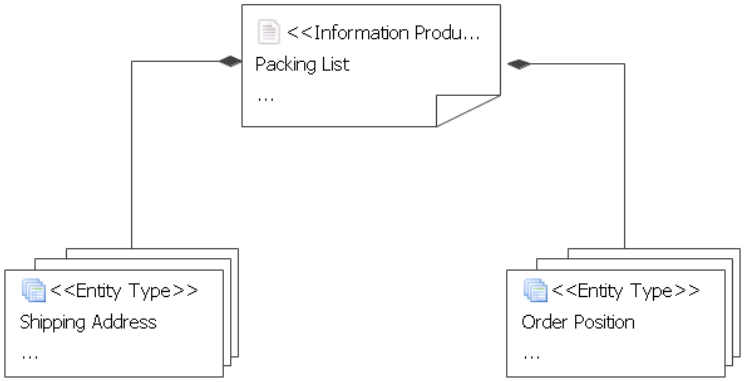


Figure 2: Assembly of Information Products

IQUATE was designed under the premise, that information is reused over the organization in different information systems. Most organizations do not possess a well-defined global data schema, yet information is present in all systems. IQUATE understands a local data schema of a system as a view on a (eventually unknown or undefined) global schema. Therefore systems are directly related to the finer granularity of attributes, which themselves assemble entities (Figure 3). This is a feasible concept, as IQUATE is neither intended to replace ER or UML models nor to create organizational data schemata.

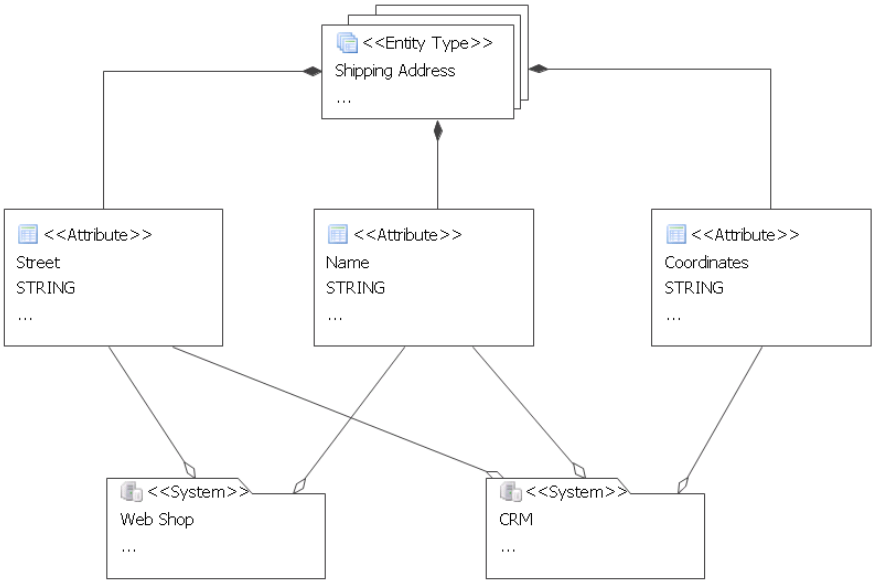


Figure 3: System view on Information

Step 2: Causes of low information quality

In contrast to IP-MAP IQUATE is not used to model quality gates but to model the impact of actions on information. Figure 4 shows an exemplary action "Call and keying in misspelled name" as an action that might occur in a call center where the caller's name is not fully understood. In this example misspelling lowers the data quality dimension "free of error" by 15% of the Attribute "Name" in the system "CRM". Several causes are modeled: "malfunction in headset", "bad connection", "hardly understandable pronunciation", and "missing check with caller". A solution to the action of misspelled name could be to check the caller automatically against the customer base to resolve uncertainties by fuzzy search algorithms. Additionally cause specific solutions may be provided. The cause "Malfunction in headset" could be fixed by "Regular replacements of headset" or by the solution "Only use high quality headsets".

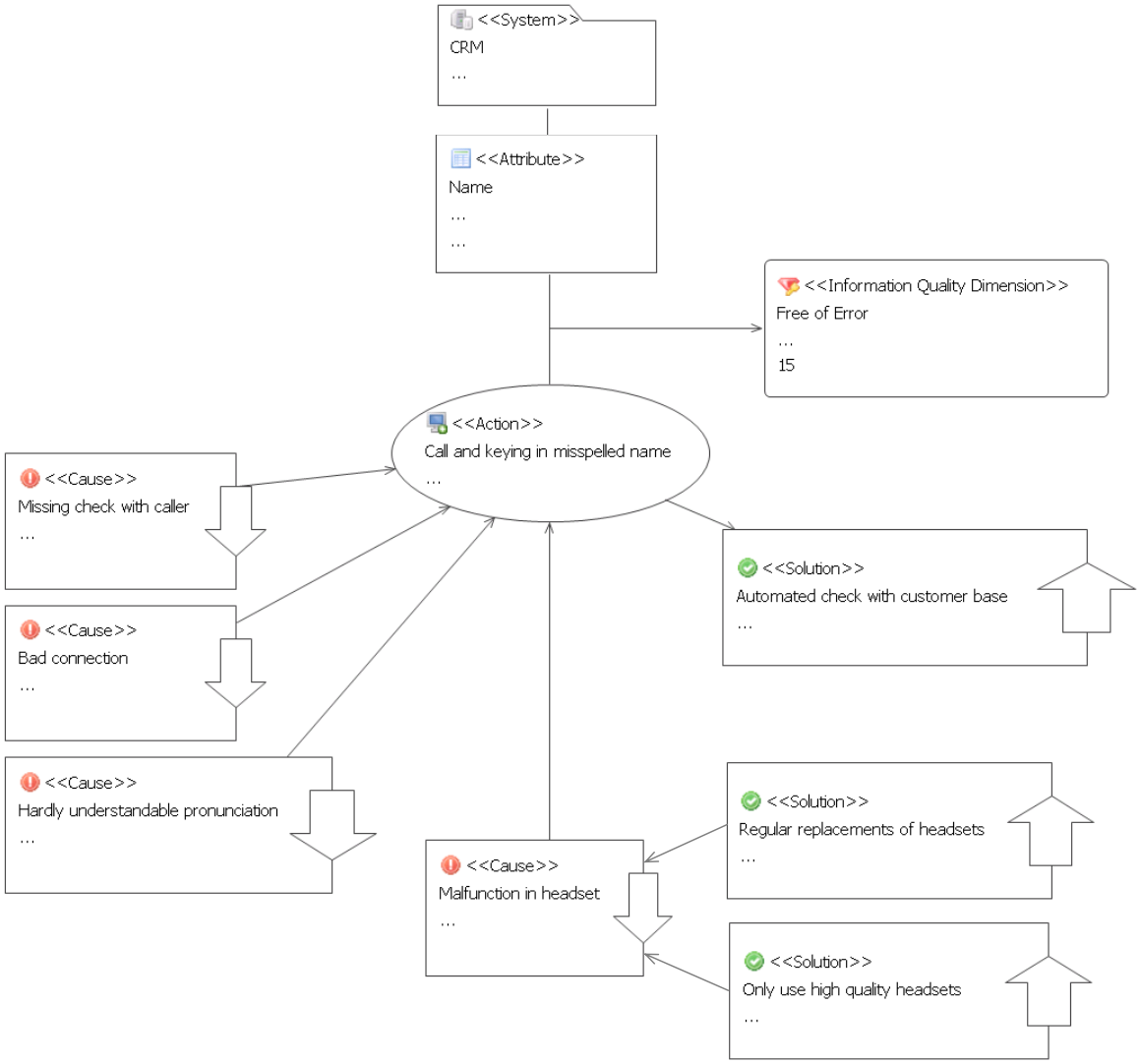


Figure 4: Cause of low information quality

Step 3: Impact of low information quality

Low information quality may impact an organization's outcome in several aspects. Batini et al. [3] provide a comprehensive classification of costs of poor data quality. For parsimony IQUATE is designed to model the impact of information products with certain information quality dimensions on hierarchically structured KPIs.

Figure 5 analyzes the shipping costs. On the first level it consists of postage and packaging as well as the return rate. The return rate is influenced by canceled orders, no satisfaction with the product and errors in the address. The last is an information quality problem that is due to the fact that the information product Packing List's "Free of Error" dimension is lowered by 15%.

Figure 6 is an example that analyzes as a particular aspect of customer dissatisfaction wrong salutations. Here, due to "distributed maintenance of salutations" without clear responsibilities the "consistent representation" of salutations is lowered by 5%. This impacts the "customer dissatisfaction due to incorrect salutations. A cause to this work practice could be the "Heterogeneous IS landscape". The "integration of customer master data" is a solution to this cause [14].

The implementation of IQUATE generates a report where to each model element its relations in terms of causes and impact to other model elements is explained. This provides a reliable overview over the importance of certain elements and an assessment of the severity of actions or work practices.

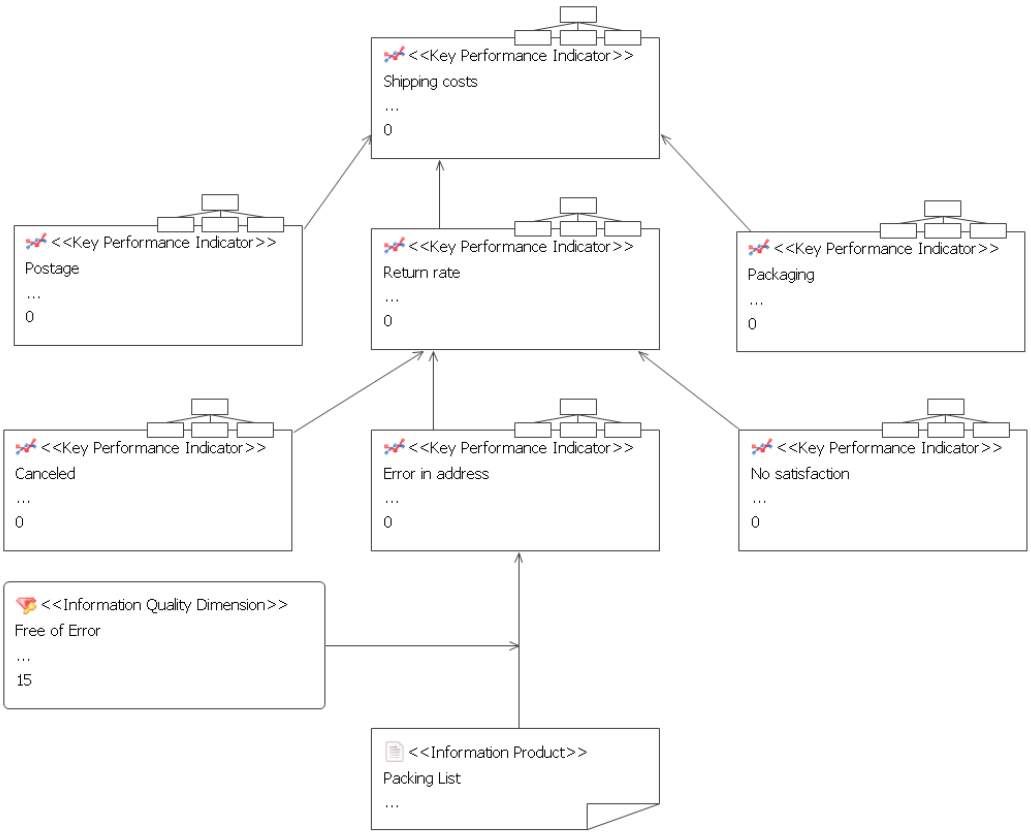


Figure 5: Impact of low information quality on return rate and shipping costs

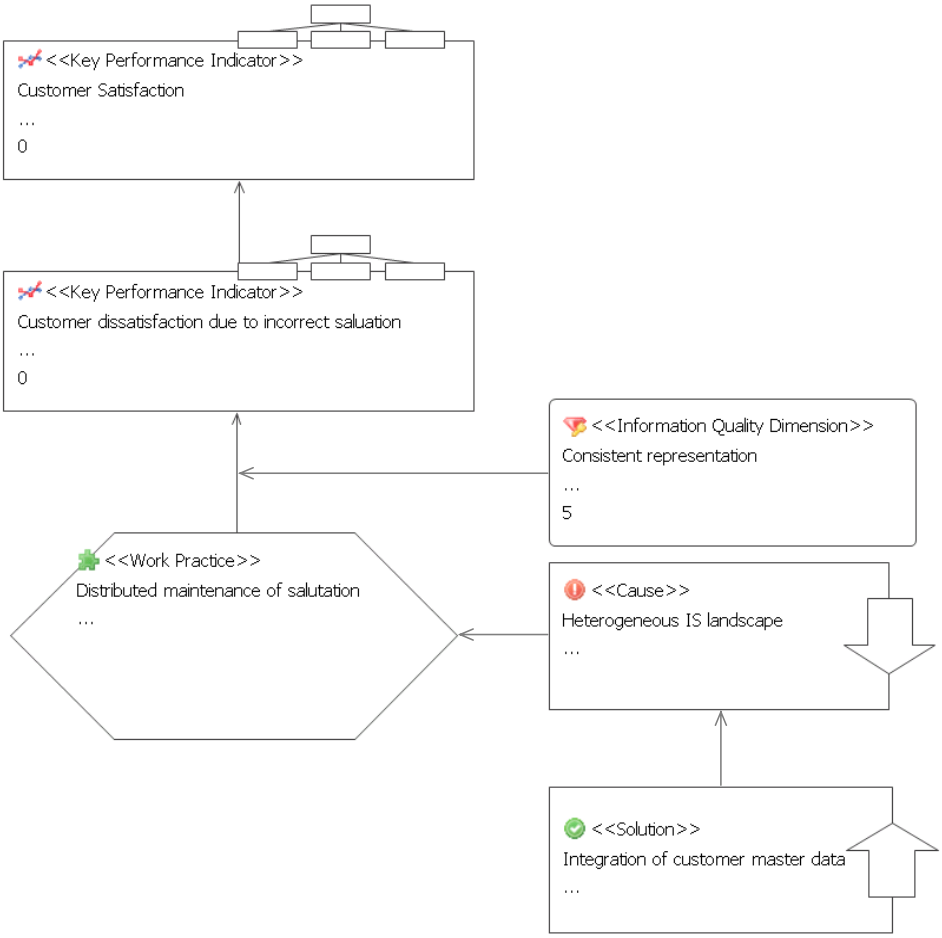


Figure 6: Impact of low information quality on customer satisfaction

CONCLUSION

Data quality has severe impacts on the performance of an organization. Nevertheless, no modeling technique existed, that allowed the structured analysis of these problems and their impacts on the organizational performance. In this contribution we presented the modeling technique IQUATE that fills this gap. As data quality problem arise in the internal and external dependencies of the technical and social subsystems of an organization, we based IQUATE on the Work System methodology. A reuse of existent modeling techniques like UML or ER for the technical subsystem or BPMN for the social subsystem was not considered feasible, as these techniques are specialized to their respective subsystems and would add unnecessary complexity. Nevertheless, an integration of IQUATE in these techniques (or IP-MAP) could be done as future work.

We demonstrated the feasibility of IQUATE by a three step analysis of an information quality problem common in the direct order industry. In that we showed that the approach fills a gap not addressed by existing modeling techniques and is an optimization for the analyses of data quality related problems. According to [12] this is a feasible evaluation of the designed artifact.

REFERENCES

[1] Alter, S. "The work system method for understanding information systems and information system research", *Communications of the Association for Information Systems*, 9, 2002, pp. 90-104

[2] Ballou, D., Wang, R. Y., Pazer, H., Tayi, G. K. "Modeling information manufacturing systems to determine information product quality", *Management Science*, 44, 4, 1998, pp. 462-484

[3] Batini, C., Barone, D., Mastrella, M., Maurino, A., Ruffini, C. "A Framework and a Methodology for Data Quality Assessment and Monitoring", *Proceedings of the 12th International Conference on Information Quality (ICIQ)*, Cambridge, USA, November 9-11, 2007

[4] Batini, C., Cappiello, C., Francalanci, C., Maurino, A. "Methodologies for data quality assessment and improvement", *ACM Computing Surveys*, 41, 3, New York, NY, USA, July, 2009, pp. 16:1-16:52

[5] Becker, D., McMullen, W., Hetherington-Young, K. "A Flexible And Generic Data Quality Metamodel", *Proceedings of the 12th International Conference on Information Quality (ICIQ)*, Cambridge, USA, November 9-11, 2007

[6] Bostrom, R., Heinen, S. "MIS Problems and Failures: A Socio-Technical Perspective PART II: The Application of Socio-Technical Theory", *MIS Quarterly*, 1, 1, 1977, pp. 11-28

[7] Caro, A., Calero, C., Piattini, M. "A Portal Data Quality Model For Users And Developers", *Proceedings of the 12th International Conference on Information Quality (ICIQ)*, Cambridge, USA, November 9-11, 2007

[8] Eclipse Foundation (2010) Graphical Modeling Project, http://www.eclipse.org/modeling/gmp/

[9] Eclipse Foundation (2011) Eclipse Modeling Framework Project, http://www.eclipse.org/modeling/emf/

[10] Eclipse Foundation (2011) Graphical Editing Framework, http://www.eclipse.org/gef/

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

[12] Hevner, A. R., March, S. T., Park, J., Ram, S. "Design Science in Information Systems Research", *MIS Quarterly*, 28, 1, 2004, pp. 75-105

[13] Kokemüller, J. "An empirical investigation of factors influencing data quality improvement success", *Proceedings of the 17th Americas Conference on Information Systems (AMCIS)*, Detroit, Michigan, USA, August 5-8, 2011

[14] Kokemüller, J. "Optimistic Integration of Enterprise Information Systems", *Proceedings of the 5th International Conference on Research and Practical Issues of Enterprise Information Systems*, Aalborg, DK, October, 16-18, 2011

[15] Kokemüller, J., Weisbecker, A. "Master Data Management: Products and Research", *Proceedings of the 14th International Conference on Information Quality (ICIQ)*, Potsdam, Germany, November 7-8, 2009, pp. 8-18

[16] Land, F. "Evaluation in a socio-technical context", *Proceedings of the International Conference on Home Oriented Informatics and Telematic*, Deventer, Netherlands, 2000, pp. 115-126

[17] Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. , *Journey to data quality*, MIT Press, Cambridge, USA, 2006

[18] Lyytinen, K. "Different Perspectives on Information Systems: Problems and Solutions", *ACM Comput. Surv.*, 19, 1, 1987, pp. 5-46 ee = http://doi.acm.org/10.1145/28865.28867

[19] Madnick, S. E., Wang, R. Y., Lee, Y. W., Zhu, H. "Overview and Framework for Data and Information Quality Research", *Journal of Data and Information Quality*, 1, 1, New York, NY, USA, June, 2009, pp. 1-22

[20] Moody, D. L., Shanks, G. G. "Improving the quality of data models: empirical validation of a quality management framework", *Information Systems*, 28, 2003, pp. 619 - 650

[21] Nunamaker Jr, J. F., Chen, M., Purdin, T. D. M. "Systems development in information systems research", *Journal of Management Information Systems*, 7, 3, 1991, pp. 89-106

[22] OMG (2010) Unified Modeling Language (UML), http://www.omg.org/spec/UML/

[23] Scannapieco, M., Pernici, B., Pierce, E. "IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework", *Proceedings of the 7th International Conference on Information Quality (ICIQ)*, Cambridge, USA, 2002, pp. 279-291
[24] Stvilia, B. "A Model For Information Quality Change", *Proceedings of the 12th International Conference on Information Quality (ICIQ)*, Cambridge, USA, November 9-11, 2007
[25] Wang, R. Y., Strong, D. M. "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems*, 12, 4, 1996, pp. 5-33