PAY-AS-YOU-GO DATA QUALITY IMPROVEMENT FOR MEDICAL CENTERS

Endler G[1], Baumgärtel P[1], Lenz R[1]

Abstract

Medical practitioners often link their databases to support new use cases of the medical sector, e.g. in economic planning or treatment coordination. Data quality requirements for these use cases differ from the original requirements on the databases. We argue that any system seeking to support data quality in this scenario requires significant evolutionary power. We suggest an approach to continuously improve data quality which scales with arising requirements in a pay-as-you-go manner.

Keywords** – **Data Quality Management, Demand-Driven Improvement

1. Introduction

It is common for resident medical practitioners to affiliate into groups to increase their power to compete [9]. Examples of such groups are Networked Practices or Medical Supply Centers (MSC in the following). Apart from benefits for the center's patients, for example interdisciplinary treatment under a single roof, there are also organizational and financial benefits for the practitioners. To achieve the latter, the new role of *practice manager* is required. These financial officers are in charge of a center's enterprise resource planning. For this function, a consolidated view over all participating practitioners' processes and data is necessary. The data for most arising use cases is already available in the practitioners' patient management systems. However, these databases are largely insular and display a high degree of heterogeneity. Additionally, necessary data may be distributed over several or all of the local databases.

1. 1. Current Situation

Although there are several standards like xDT or DICOM for data exchange, no overarching standard has yet been adopted. Neither can a center require newly joining practitioners to change their patient management systems - such a requirement would prevent many practitioners from joining in the first place. It follows that the existing databases need to be integrated into a central knowledge base. The heterogeneities in such an integration scenario are a core reason for low data quality. Another potentially negative impact on data quality is the fact that these data were collected for a different purpose. While they may be fit for use [17] for the original function, they may not be suitable for the new use cases of practice managers since data quality is context dependent [1].

[1] Chair for Computer Science 6 (Data Management), Dept. of Computer Science, University of Erlangen-Nuremberg

Another challenge is the highly volatile nature of the healthcare system, meaning that technology and legislation may frequently change, requiring new types of data or data with higher quality. Currentness of a center's data is a consideration as well: It is estimated that "2% of records in a customer file become obsolete in a month" [4].

1. 2. Objectives

As it is impossible to foresee all future requirements on data and data quality, it is necessary to adopt an evolutionary approach that is able to scale with newly arising requirements. This will enable a practice manager to adapt the center's data quality standards in a demand-driven manner.

2. Methods

2. 1. Requirements Analysis

Through interviews with practice managers and practitioners, we identified the main new use cases. While most of these deal with financial controlling and planning [5], they nevertheless require extensive data from the practitioners' local systems. As an example, resident practitioners operate on a health insurance mandated budget. Any benefits they provide to their patients beyond this budget are only fractionally remunerated. To prevent this, an MSC manager needs a complete view over all practitioners' data - if the extent of benefits provided is unknown, the manager cannot know if and where to countersteer. Based on these data needs we developed a core database schema to serve as a central point of information for a center. Since the practitioners will be loath to relinquish control of their databases, we cannot directly influence their data. Thus, all quality considerations concern the central database. To further gauge the data quality needs in MSCs, we conducted a survey among practice managers1.

2. 2. Data Quality Management

Data quality, generically defined as the "fitness for use" of data [17], is an important concern in all application areas of databases [8]. It is a multi-dimensional concept [18,16], with varying definitions of dimensions in the literature. The three most commonly mentioned are correctness, completeness, and currentness (sometimes called "currency"). Finding ways to measure data quality in a specific context is regarded as non-trivial [3]. For a given project, however, it is not enough to assemble the necessary dimensions. All data quality considerations must be regularly evaluated and, if necessary, improved. The generic approach for this is given by Wang [17, 18] with the Total Data Quality Management (TDQM) Cycle (see *Figure 1*).

2. 3. System Evolution: The Pay-As-You-Go Approach

Taking into consideration all feasible ways of measuring data quality from the start will lead to very high upfront effort. Some of the work expended may even be needless, and the attempt to foresee every problem may lead programmers to ignore the principle of "design for change" [14]. Additionally, changes in data format, technology or healthcare legislation are frequent and may effectively invalidate previous solutions. By contrast, a pay-as-you-go approach [11] allows a new system to be imperfect at the beginning, and to be improved by the users in a demand driven

[1] questionnaire available at *www6.informatik.uni-erlangen.de/people/greg/DQSurvey.pdf*

manner. By deferring part of the system design to runtime in this way, we gain continuous adaptability [12]. In fact, in a scenario like ours, the pay-as-you-go stance may even be the only viable approach [13].

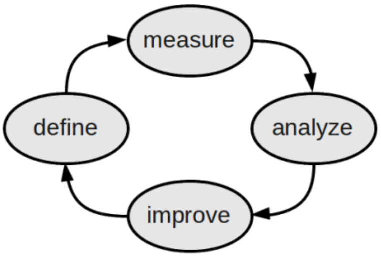

Figure 1: Total Data Quality Management (TDQM) cycle

3. Results

3. 1. Data Quality Dimensions and Metadata

Of the standard data quality dimensions mentioned in section 2.2, interviews with domain experts showed population completeness1 of the medical benefits provided in the center to have the most direct impact on revenue (see the example in section 2.1). Additionally, there is no reference data for these, since the number of benefits provided at a practice varies. This means that population completeness is non-trivial to measure and must be estimated instead [6]. In some other cases, counting NULL-values of a table may deliver a realistic measure of completeness.

Table 1: Metadata

Type	Name	Granularity
Provenance (tracing)	Practice	Entry
	System	Entry
	Type of system	Entry
Timestamps (currentness, completeness)	Created at source	Entry
	Loaded from source	Entry
	Changed at source	Entry
	Changed at central database	Entry
Values (correctness)	Unit	Attribute
	Range	Attribute
NULL count (completeness)	Rows	Attribute
	Columns	Tuple

The dimension deemed second most important is correctness of data, which can be measured using plausibility rules. Currentness of data was considered least important by domain experts. Based on literature [16], our interviews and domain context, we identified several metadata to support data quality monitoring (see *Table 1*). *Provenance* is used to trace the origin of possibly dirty data items. *Timestamps*, *Values*, and *NULL count* help measure quality along the dimensions mentioned above.

[1] Population completeness: percentage of real-world entities that have a corresponding entry in a database [16].

3. 2. TDQM in Medical Supply Centers

The metadata in *Table 1* serve as a starting point for data quality measurements. However, we cannot guarantee that all future quality requirements can be supported by these metadata (see section 2.3). Therefore, we intend a rule system similar to Blechinger's work [2] as the core of our data quality management approach. It will enable MSC managers to define new quality requirements in an intuitive way. Through the definition of additional rules and associated metrics, the managers can create indicators for different dimensions of data quality relevant for the MSC. This introduces evolutionary capabilities, and accounts for the requirements of an ever-changing health system.
While inherently supporting the pay-as-you-go approach as ways of measuring data quality are evaluated and possibly changed in every iteration of the cycle, TDQM is highly generic. To make it applicable for our domain, we adapted TDQM for use in MSCs, and extended it with an initial definition phase. *Figure 2* shows this concretization. In the initial definition phase, the standard dimensions and metrics and the rule system are implemented. These serve as a starting point for data quality monitoring. Afterwards, MSC managers can customize the system to their needs.

The monitoring system's purpose is twofold: For one, it calculates the implemented metrics and based on these, maintains a list of potential problems. Secondly, it needs to estimate the utility and cost of resolving these problems, and order the list accordingly. These estimations are still work in progress. Jeffery et al suggest that a sensible measure for estimating utility for tasks like this is the value of perfect information (VPI) [11]. A definition of VPI is given by Russel and Norvig [15]. The cost estimate may incorporate the amount of data affected, the number of sites involved and the general complexity of the task presumably necessary for correction of the error. Additionally, a simple urgency measure is available in that warnings are displayed when imperfect data is accessed.

4. Conclusion

In a volatile domain like the healthcare system, data quality improvement needs to be continuous and sustainable. To contribute to a solution, we extended and adapted the generic TDQM methodology, and are currently building tools for data quality monitoring to support this adaptation. We want to achieve this in a pay-as-you-go manner, improving data quality monitoring by demand while almost seamlessly integrated into the regular work processes of an MSC. We established basic artifacts to serve as a starting point for data quality monitoring. We examined the practicality of the generic pay-as-you-go approach for our domain of application, and assembled the necessary use cases as well as their data needs.

In contrast to [10], our concretization of TDQM does not differentiate between "ex-ante" and "ex-post" improvement of data quality. Rather, it embeds itself in the day-to-day routine of the involved parties and iterates continuously. We cannot rely on standards like HL7 or IHE, since most of these are prevalent mainly in the hospital sector and not in local practices. In addition, these standards provide domain specific declarative and functional specifications, but no standards for data quality processes. Adherence to standards may alleviate individual data quality problems, but they neither offer guarantees to this nor make any assertions on problems like measuring population completeness. Also, while reliance on standards clearly is important for clinical studies, it may in fact prove a crutch to evolutionary capabilities of systems used by medical professionals in their own practice due to the rapidly changing conditions in healthcare legislation [7].

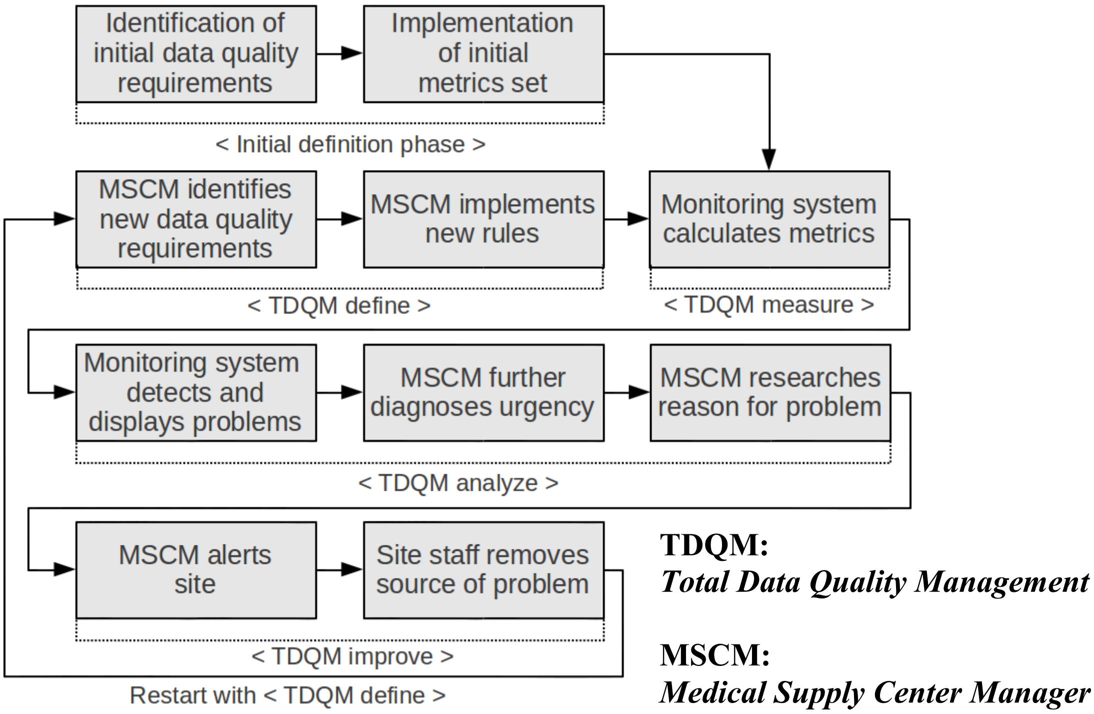
Identification of initial data quality requirements
Implementation of initial metrics set
< Initial definition phase >
MSCM identifies new data quality requirements
MSCM implements new rules
Monitoring system calculates metrics
< TDQM define >
< TDQM measure >
Monitoring system detects and displays problems
MSCM further diagnoses urgency
MSCM researches reason for problem
< TDQM analyze >
MSCM alerts site
Site staff removes source of problem
< TDQM improve >
Restart with < TDQM define >
TDQM:
Total Data Quality Management
MSCM:
Medical Supply Center Manager

Figure 2: TDQM cycle for Medical Supply Centers

5. Acknowledgements

The project is supported by the German Federal Ministry of Education and Research (BMBF), project grant No. 13EX1013D.

6. References

[1] Bertossi L, Rizzolo F, Jiang L. Data Quality Is Context Dependent. In: Castellanos M, Dayal U, Markl V, Aalst W, Mylopoulos J, Rosemann M, et al., editors. Enabling Real-Time Business Intelligence. vol. 84 of Lecture Notes in Business Information Processing. Springer Berlin Heidelberg; 2011. p. 52–67.

[2] Blechinger J, Lauterwald F, Lenz R. Supporting the Production of High-Quality Data in Concurrent Plant Engineering Using a MetaDataRepository. In: AMCIS 2010 Proceedings; 2010.

[3] Dustdar S, Pichler R, Savenkov V, Truong HL. Quality-aware service-oriented data integration: requirements, state of the art and open challenges. SIGMOD Rec. 2012 Apr;41(1):11–19.

[4] Eckerson WW. Data quality and the bottom line: Achieving business success through a commitment to high quality data. The Data Warehousing Institute. 2002;p. 1–36.

[5] Endler G, Langer M, Purucker J, Lenz R. An Evolutionary Approach to IT Support for Medical Supply Centers. In: Proceedings der 41. Jahrestagung der Gesellschaft für Informatik e.V. (GI); 2011.

[6] Endler G, Baumgärtel P, Held J, Lenz R. Data quality for managers of medical supply centers. In: und Epidemiologie e V (GMDS) Deutschen Gesellschaft für Medizinische Informatik B, editor. GMDS 2012; 2012.

[7] Endler G. Data quality and integration in collaborative environments. In: Proceedings of the on SIGMOD/PODS 2012 PhD Symposium. PhD '12. New York, NY, USA: ACM; 2012. p. 21–26.

[8] Fan W, Geerts F. Foundations of Data Quality Management. Özsu MT, editor. Morgan & Claypool Publishers; 2012.

[9] Hellmann W, Eble S. Gesundheitsnetzwerke managen - Kooperation erfolgreich steuern. Medizinisch Wissenschaftliche Verlagsgesellschaft; 2009.

[10] Holzer K, Dorda W, Duftschmid G, Nachbagauer A, Strasser N, Wrba T, et al. Ein Vorgehensmodell zur Erhöhung der Datenqualität klinischer Routinedaten im Kontext der Sekundärnutzung. In: Proceedings of eHealth 2012; 2012. p. 205–210.

[11] Jeffery SR, Franklin MJ, Halevy AY. Pay-as-you-go user feedback for dataspace systems. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. SIGMOD ’08. New York, NY, USA: ACM; 2008. p. 847–860.

[12] Lenz R. Information Systems in Healthcare - State and Steps towards Sustainability. IMIA Yearbook 2009. 2009;1:63–70.

[13] Madhavan J, Jeffery SR, Cohen S, Dong X, Ko D, Yu C, et al. Web-scale Data Integration: You Can Only Afford to Pay As You Go. In: In Proc. of CIDR-07; 2007.

[14] Parnas DL. Software aging. In: Proceedings of the 16th international conference on Software engineering. ICSE ’94. Los Alamitos, CA, USA: IEEE Computer Society Press; 1994. p. 279–287.

[15] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc.; 2003.

[16] Scannapieco M, Missier P, Batini C. Data Quality at a Glance. Datenbank-Spektrum. 2005;14:6–14.

[17] Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998 February;41:58–65.

[18] Wang RY, Ziad M, Lee YW. Data Quality. Springer US; 2002.

Corresponding Author

Gregor Endler
Chair for Computer Science 6 (Data Management)
Dept. of Computer Science, University of Erlangen-Nuremberg
Martensstraße 3, D-91058 Erlangen
Email: gregor.endler@fau.de