

A DATA QUALITY FRAMEWORK FOR SMALL BUSINESSES

(Research in Progress)

Peggy Leonowich-Graham
Colorado Technical University¹
pjleonowich@excite.com

Mary Jane Willshire
University of Portland²
willshir@up.edu

Abstract: Data quality, while important, is often not a priority because it must compete for limited corporate resources. However, it is essential, especially for small businesses, to have a low cost method to address data quality issues. The focus of this research is to use process improvement to extend an existing theoretical model, Total Data Quality Management, TDQM, to a new field of practice that could be used by small businesses. This study modified TDQM to create a framework that is less resource intensive and more cost effective. The goal of this research is to make data quality attainable for small companies.

Key Words: Data Quality, Information Quality, TDQM, Information Product

INTRODUCTION

Every day around the nation, company management makes decisions that are based on the data in their systems. However, what method is used to keep the data quality to an acceptable standard? Decisions made based on poor data quality can be costly to an organization [4]. Information technology enables companies to deliver quality information. Until companies treat information as a product and manage it accordingly, quality information will not be delivered to consumers consistently and reliably. The results will be lower margins, missed opportunities and tarnished images [3].

Small companies are faced with an even greater obstacle in achieving data quality. Small companies have limited personnel and financial resources. They cannot afford to dedicate people full time to a data quality process. They do not have disposable income to commit to a process that cannot guarantee a return on investment. This research specifically targets the problems faced by a small company that would like to have a cost effective method for improving its data quality. The remainder of this paper discusses the existing data quality model, TDQM, with our proposed modified, lower cost model, the Small Business Quality Framework, SBQF. Each model is presented and then a cost comparison is shown.

HOW MUCH QUALITY

¹ Peggy Leonowich-Graham, Colorado Technical University, 4435 N. Chestnut Street, Colorado Springs, CO 80907.

² Mary Jane Willshire, School of Engineering, University of Portland, 5000 N Willamette Blvd., Portland, OR 97203.

In determining how much quality is needed there needs to be an agreement to the definition of data quality. Ken Orr defined data quality as the measure of agreement between the data views presented by an information system and the same data in the real world [5]. A system's data quality of 100% would mean that the data views were completely in agreement with the real world. It is unlikely that a system would have 100% compliance. According to Orr the real concern with data quality is to ensure not that the data quality is perfect but that the quality of the data in the information system is accurate enough, timely enough and consistent enough for the organization to survive and make reasonable decisions [5].

One approach is that the data quality process must take into account the subjective perceptions of the data holders and the objective measurements based on the data set in question [6]. Pipino details a three step process to assess data quality [6]. The first step is to perform a subjective and objective data quality assessment. Next compare the results of the assessments, identify discrepancies and determine the cause of the discrepancies. Lastly, identify the plan of action for correction and improvement.

Redman indicates that the quality management principles that work in a non data domain can be applied to the data environment [7]. He indicates that there are five fundamental principles [7]. First, data customers are the ultimate arbiters of their quality. Second, measurement of performance against customer needs is essential. Third, techniques that find and eliminate root data errors are better than identifying and fixing existing data errors. Fourth, control helps managers establish a basis for future performance. Fifth, senior management plays a critical role in a data quality program.

One of the research questions was what is the meaning of sufficient quality and how do you know that you have achieved it. How good is good enough? This research focused on whose opinion counted, the purpose of the quality review and a 4 step process to determine sufficient data quality had been achieved. The determination of sufficient data quality becomes a choice. The following 4 step process was used to decide that sufficient quality had been achieved:

1. It has sufficient benefits
2. It has no critical problems
3. The benefits sufficiently outweigh the problems
4. In the present situation, and all things considered, further improvement would be more harmful than helpful [1].

TDQM

Total Data Quality Management, TDQM, is a method to deliver high quality information products, IP, to information consumers [9]. The field of product manufacturing has an extensive body of Total Quality Management, TQM, literature with principles, guidelines and techniques for product quality [3]. In TQM literature, the widely practiced Deming cycle for quality enhancement consists of: Plan, Do, Check, and Act [2]. Adapting the Deming cycle, the TDQM approach has the components Define, Measure, Analyze and Improve. Thus, in order to apply the TDQM an organization must perform the following steps: [9]

1. Clearly articulate the IP in business terms;
2. Establish an IP team consisting of a senior executive as the TDQM champion, an IP engineer who is familiar with the TDQM methodology and members who are information suppliers, manufacturers, consumers and IP managers;
3. Teach IQ assessment and IQ management skills to all the IP constituencies;
4. Institutionalize continuous IP improvement.

The definition component of the TDQM cycle identifies Information Quality, IQ, dimensions. The measurement component produces IQ metrics. The analysis component identifies root causes for IQ problems and calculates the impacts of poor data quality information. Lastly the improvement component provides techniques for improving IQ. They are applied along IQ dimensions according to requirements specified by the consumer [3].

TDQM is extremely resource intensive and there is no guarantee that there will be a significant return on the

investment. Like commissioning an internal audit, TDQM only adds value if it finds repairable problems with the data quality. If no problems are found, the money expended can only be justified as providing 'peace of mind'. A Fortune 500 company may have little problem expending resources to attain 'peace of mind', but what about the small business? Everyday, small businesses are faced with crucial decisions such as deciding between meeting the payroll or purchasing additional office equipment. Commissioning an internal audit of company data quality seems a luxury.

Unfortunately, as we enter the age of information, that small businesses data quality becomes a necessity, not a luxury. While a business may be small in employee size or annual revenue, that does not mean their information structure is comparable. In many cases the opposite is true; the small company's information structure and database size are just as large as that of any Fortune 500 company. The small company is equally at data quality risk as that of the Fortune 500 company.

SMALL BUSINESS QUALITY FRAMEWORK, SBQF

TDQM is a resource intensive commitment and well outside the resources available in a small organization. Further, TDQM assumes that the organization will change their Information Product or Information Production Method as a result of the TDQM process. Therefore, a significant portion of the business sector (small business with governmental oversight), do not find TDQM a feasible path in supporting their Information Product. It is this type of organization that is the target audience for our research.

The purpose of this research is to use process improvement to extend TDQM by developing a limited-resource, low commitment TDQM process that provides Information Quality Data at an acceptable quality level. This approach involves improving the process by customizing the sixteen information quality dimensions to the existing information product and developing a customized survey.

Three specific targets of this work were:

- Provide a TDQM methodology whose cost is within the available resources of small companies. To do this, the estimated expenditures by a small company using the Small Business TDQM must be cut by 80% in comparison with the expenditures using a full TDQM methodology.
- Provide a TDQM methodology that specifically reduces the initial commitment of organization personnel by 60% or more.
- Provide a TDQM methodology that does not result in a new Information Production System.

The proposed solution seeks to revise only the above stated areas of TDQM.

The small business quality framework, SBQF, is the proposed result of our research effort. It is an extension and significant modification of TDQM. It still treats data quality from the information product perspective; but, the framework strives to reduce the resources required to perform data quality in a small business environment. In applying SBQF, an organization must perform the following steps:

1. Define IP;
2. Train IP Expert Team;
3. Modify IQ Dimensions to IP;
4. Conduct Survey on Sample of Population;
5. Conduct Expert Analysis;
6. Develop IP Plan.

To perform SBQF an organization must first determine the information product to examine. Next a small team of product experts is formed. The team must be thoroughly knowledgeable on the information product chosen. The team is then trained on the basic concepts of quality and the sixteen IQ dimensions [8]. The sixteen IQ dimensions are accuracy, objectivity, believability, reputation, access, security, relevancy, value added, timeliness, completeness amount of data, interpretability, ease of understanding, concise representation, consistent representation and ease of manipulation. Next the expert team defines the sixteen IQ dimensions as they apply to the specific domain and to

the information product. The expert team develops a survey using these definitions. A two part survey is distributed to a sample of the product's population. The first part of the survey uses the refined definitions and the second part of the survey examines the perceived importance of the sixteen IQ dimensions. The expert team then analyzes the survey results. Lastly, the expert team develops an IP plan to correct the data quality weaknesses.

The proposed framework differs from TDQM in that it provides a low cost quality review thru product customization. SBQF utilizes existing personnel and knowledge in the organization and as such is able to minimize costs. The framework differs from TDQM in that it customizes the IQ dimensions to the information product. The IQ dimensions are defined in the product specific terminology. This step allows for the elimination of quality training to the population. The IQ survey is also written in the product specific terminology. The survey is only administered to a sample of the population. This further reduces the number of individuals in the quality review process. Limiting the number of personnel involved in the review process is important to small businesses due to resource limitations.

Application of SBQF in a Small Company

The Small Business Quality Framework was applied to a non-profit Mental Health Agency in Colorado. The organization employs 300 people and treats mental health and substance abuse problems in Colorado. As a Mental Health facility the organization must comply with file keeping procedures, data security requirements and State reporting procedures. As a result the databases are immense. The organization has devoted significant resources to internally develop software to gather, store, secure and cross reference the collected data.

The information product selected in the first step of SBQF for quality review was the Colorado Client Assessment Record (CCAR). The CCAR is a tool used by clinicians to assess the level of functioning of patients with mental illnesses. It is a 4-page form with over 40 questions. The CCAR is completed on a patient's first visit; then updated on an annual basis and upon discharge. The CCAR is used by the State to measure how successful a mental health agency is in treating patients and is used by the State to determine funding.

As the next step, an expert team of four willing professionals was identified and recruited to participate in the research effort. The expert team formed had seventy five years of experience in mental health and forty years of experience with the CCAR. The team began their work by redefining the sixteen IQ dimensions as they would pertain to the CCAR. The team then developed a two-part survey to examine the data quality of the CCAR and the perceived importance of the sixteen IQ dimensions. The survey was distributed to a sample of the CCAR user population.

The purpose of the first part of the survey was to determine if the CCAR correctly measured each of the quality dimensions. Areas of concern appeared to be objectivity, believability, access, completeness, amount of data, interpretability and concise representation. However, upon applying a confidence interval approach, 95% of the population was within an acceptable range in all areas of concern except for the dimensions of interpretability and concise representation. The CCAR tool did not capture these dimensions well.

The second part of the survey examined whether each of the quality dimensions was important. The results of the survey indicated that all sixteen of the quality dimensions were important to the information product population. The least important of the sixteen dimensions were access, value added, and ease of manipulation.

Information Product Plan

Analysis of the survey results shows that the CCAR tool had deficiencies in the area of interpretability and concise representation. The expert team formed a plan with recommendations on ways to correct the lack of data quality in these two dimensions. The recommendation on how to improve interpretability was to incorporate CCAR training into standard clinical meetings. The source of the problem was how to eliminate the rating differences between clinicians. This plan would use video clinical scenarios as a training tool. The video would be shown and attendees would independently complete a CCAR while watching the video. Afterwards, the patient scenario would be discussed and the appropriate CCAR rating would be explained by the senior clinician.

The second dimension that required an improvement plan was concise representation. The expert team determined that the CCAR was not a good measure of a person’s psychiatric well-being. The CCAR is a measure of a point in time of a person’s psychiatric state. Additional tools were needed for the treatment process. The CCAR is already being supplemented with the Global Assessment which is a tool that examines all significant events in a person’s life. To effectively treat a patient a clinician uses the CCAR, Global Assessment and Progress Notes in the development of a treatment plan. The CCAR should not be used for treatment without supplemental treatment tools.

Activity	Time expended Per person	Cost
Customize the definitions for the Data Quality dimensions	2 hours	\$247.74
Develop the survey	1 hour	\$123.87
Conduct survey (assume 43 people, each taking about 30 minutes)	0.5 hour	\$292.77
Analyze the survey results	1 hours	\$123.87
Total costs		\$788.25

Table 1. Actual Costs

Cost of SBQF

As can be seen in Table 1, the SBQF method was a low cost approach to examining quality. The method cost the organization under \$800 to examine the quality of the CCAR and formulate an action plan to correct the deficiencies found. No new personnel positions were created or hired for the process. We expect that other small organizations would see a similar sort of expenses associated with the SBQF approach.

Activity	Time expended Per person	Cost
TDQM Expert (either new hire or dedicate existing person)		\$75,000
Cost to train all personnel (322 people)	1 hour	\$4,000
Conduct survey (assume 322 people, each taking about 30 minutes)	0.5 hour	\$2000
Analyze the survey results (cost incorporated into expert’s salary)		\$0
Total costs		\$81,000

Table 2. Estimated Costs

Estimated Cost of TDQM

As can be seen in Table 2, the estimated cost of TDQM was significantly higher. The anticipated cost for this approach was \$81,000. A TDQM champion trains and surveys the entire population. It is a comprehensive examination of quality.

CONCLUSION

The SBQF approach provided a means for a small organization to examine data quality. The framework allowed the small Mental Health facility to identify two problem dimensions in their information product, the CCAR. The small team of experts were able to make recommendations to address the data quality issues. We feel that SBQF is a low cost method that could be used to identify and correct data quality in a small organization. Data quality

becomes much more attainable for small organizations through the use of SBQF. It is a versatile framework that uses expertise which already exists within the organization, and does not require that the IP be replaced. The next step in our research project is to put the recommendations into place and measure the changes in the data quality of the CCAR data as a result of the SBQF.

REFERENCES

- [1] Bach, James, "Good Enough Quality: Beyond the Buzzword," *IEE Computer Society*, August 1997, pp. 96-97.
- [2] Deming, E.W., *Out of the Crisis*, Center for Advanced Engineering Study, MIT, Cambridge, MA, 1986.
- [3] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.
- [4] Knight, Bob, "The Data Pollution Problem," *Computerworld*, September 28, 1992, pp. 81-82.
- [5] Orr, Ken, "Data Quality and Systems," *Communications of the ACM*, February 1998, vol. 41, issue 2, pp. 66-71.
- [6] Pipino, Leo L., Yang W. Lee, and Richard Wang, "Data Quality Assessment," *Communications of the ACM*, April 2002, pp. 211-218.
- [7] Redman, Thomas C., "What Those Who Know Quality Need to Understand," *Proceedings of the 55th Annual Quality Conference*, May 2001, pp. 534-538.
- [8] Wang, R. and D. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, Spring 1996, pp. 5-34.
- [9] Wang, R., "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, February 1998, pp. 58-65.