A CLASSIFICATION AND ANALYSIS OF DATA QUALITY COSTS

(Research in Progress)

Martin J. Eppler
University of Lugano (USI), School of Communication Sciences, Switzerland
Martin.Eppler@lu.unisi.ch

Markus Helfert
Dublin City University, School of Computing, Ireland
Markus.Helfert@computing.dcu.ie

Abstract: Many information quality initiatives and projects need to demonstrate the potential benefits of their IQ-related activities already in their planning stage. In doing so, practitioners rely on cost estimates based on current non-quality data effects (that are then compared to data quality improvement costs). In producing such estimates on costs caused by low quality data, it is difficult to identify all potential negative monetary effects that are the result of low quality data (as well as all possible costs associated with assuring high quality data and their progression). Consequently, this article reviews and categorizes the potential costs associated with low quality data and examines their progression. This analysis can help practitioners to identify cost saving potentials and argue a more convincing business case of their data quality imitative. For researchers, the proposed classification framework and the cost progression analyses can be helpful to develop quantifiable measures of data quality costs and to prepare – subsequently – benchmarking studies, comparing different cost levels in different organizations. Thus, the paper contributes elements of a future cost-benefit analysis method for data quality investments.

Key Words: Data Quality, Data Quality Costs, Data Quality Benefits, Cost Measurement, Optimal Data Quality Costs, Data Quality Cost Benefit Analysis

INTRODUCTION

What kind of return on investment can an organization expect from its data quality initiative? Many companies find this question difficult or even impossible to answer. The reason for this is the arduous task of quantifying the current costs of low quality data that will eventually be reduced because of deployed data quality activities (which again cause costs that need to be accounted for in an ROI calculation). Calculating the current costs caused by insufficient levels of data quality is particularly difficult because many of these costs are indirect costs, that is to say costs where there is no immediate link between inadequate data quality and negative monetary effects. Consequently, it is difficult to identify these often hidden, indirect costs, let alone quantify them.

To facilitate the task of identifying and characterizing such costs, this article reviews existing cost types resulting from low quality data and structures them with the help of classifications. This can enable practitioners to better argue their business case by more easily identifying current costs of low quality data. For researchers, the presented classification framework can be helpful to develop quantifiable measures of data quality costs and to prepare – subsequently – benchmarking studies, comparing different cost levels in different organizations in a coherent manner and based on consistent cost distinctions. Thus, this paper contributes first elements of a cost-benefit analysis method for data quality investments. Such a

future cost-benefit analysis should not only rely on static information about data quality costs. It must also take into account the subsequent progression or evolution of these costs (as investments in data quality rise). Consequently, different possible cost progressions are analyzed in the discussion section of this paper.

BACKGROUND, RATIONALE, AND PURPOSE

Although there is a plethora of literature that claims that the costs of missing data quality are substantial in many companies (e.g. [6], [8], [13], [17], [22]), there are still very few studies that actually demonstrate how to identify, categorize and measure such costs (and how to establish the causal links between data quality defects and monetary effects). [31] confirm this assessment in their analysis of data quality costs:

> "There have been limited efforts to systematically understand the effects of low quality data. The efforts have been directed to investigating the effects of data errors on computer-based models such as neural networks, linear regression models, rule-based systems, etc. [Kauffman et al 1993]. In practice, low quality data can bring monetary damages to an organization in a variety of ways. As observed earlier in [Kim 2002], the types of damage it can cause depend on the nature of data, nature of the uses of the data, the types of responses (by the customers or citizens) to the damages, etc."

This lack of insight regarding the monetary effects of low quality data, however, is not only an open research problem, but also a pressing practitioner issue. Humbert and Elisabeth Lesca, two information management consultants and university professors, conclude similarly that it is very rare that a company analyses the costs resulting from non-quality information [17, p. 116]. Although we agree with the fact that data quality costs are context-dependent [6], that is to say that the types of damage caused by low quality data depend on the nature of the managed data, its uses and responses, we believe that proven approaches from other cost domains (such as accounting or manufacturing) can be fruitfully applied to the data quality field. Specifically, cost classifications based on various criteria can be applied to the data quality field in order to make its business impact more visible. Generic classifications of data quality costs can offer various advantages, ranging from clearer terminology, changes in perspectives, to more consistent measurement metrics. A classification, according to [1] is the ordering of entities into groups or classes on the basis of their similarity. Classifications minimize within-group variance, and maximize between-group variance, thus facilitating analysis, organization and assessment (if the goal of within group homogeneity & between-group heterogeneity is met). Classification (according to [5]) can also be described as a spatial-temporal segmentation of the world (or one aspect of it). It exhibits the following properties: there are consistent, unique classificatory principles in operation; the categories are mutually exclusive; the system is complete. Taxonomies, as a special kind of classification, are tied to a *purpose*, in the context of this paper making sure that all relevant data-quality related costs are taken into account when performing a cost-benefit analysis for a data quality program. [1] points out the crucial difference between taxonomy and typology: Whereas a typology is conceptual, deductive, and based on reasoning (e.g., a two by two matrix classification, see Figure 1), a taxonomy is empirical, inductive, and based on large sets that are examined and grouped, e.g., through cluster analysis [1, p. v]. Figure 1 shows a simple data quality cost typology based on the criterion of direct or indirect relation to data quality taken from [16, p. 210] who has adapted it from guidelines of the US Department of Defense. While this typology already provides valuable insights into information quality investment costs and low information quality effects, it does not relate or specify them and it does not provide a comprehensive IQ cost overview. In addition it mixes – within the same category – costs of assuring high quality data and costs resulting from low quality data (e.g., training, correction, and error costs).

Direct IQ Costs	**Indirect IQ Costs**
1. Controllable Costs - Prevention Costs - Appraisal Costs - Correction Costs	1. Customer Incurred Costs
2. Resultant costs - Internal Error Costs - External Error Costs	2. Customer Dissatisfaction Costs
3. Equipment and Training Costs	3. Creditability Lost Costs

Figure 1: A sample data quality cost typology [16]

Another important definition – besides that of classification – regards the concept of cost. We believe that in the data quality field, a narrow definition of the cost concept is counterproductive, excluding many important negative effects of data quality. Consequently, we define the term cost in this context as *a resource sacrificed or forgone to achieve a specific objective* or as the *monetary effects of certain actions* or a lack thereof. The specific objective mentioned in the previous definition of cost is, in our case, a certain level of data quality. The data quality costs are thus the actual negative monetary effects that result from not reaching a desired data quality level. How such costs can be compiled is described in the next section.

METHODS

In order to develop a systematic classification of data quality costs that can be used for future cost-benefit analyses within companies, we have proceeded in the following, exploratory sequence of steps:

1. First, we have identified specific cost examples from data quality literature (such as journal articles and MITIQ proceedings). They are listed in tables 1 and 2 of the results section.
2. Second, we have clustered these examples into cost groups based on shared criteria (e.g., where the costs originate, who bears the costs, how the costs can be measured, which IQ attributes they affect, etc)
3. Third, we have reduced the cost groups into major cost categories that are, as far as possible, mutually exclusive and collectively exhaustive.
4. As a next step, we have related the various cost categories to one another in an instructive way (e.g., through an information lifecycle perspective), in order to convert the findings into a useful management framework.
5. The developed classifications are then evaluated for different application purposes and their advantages and disadvantages are discussed.
6. As a last step, we have analyzed central cost categories in terms of their mutual influence and progression in order to move from a static taxonomy to a more dynamic understanding of data quality costs (see the discussion section).

Through this process of refinement, the concept of data quality cost is iteratively sharpened, but also viewed from various perspectives.

RESULTS

The process outlined in the previous section has resulted in several types of results that are presented in this section. First, it has produced a low data quality cost example directory, as listed in table 1. Table one lists specific costs that have been mentioned in the IQ literature. These costs, as one would imagine, are not based on the same classification principle or abstraction level. However, the list of 23 examples in Table 1 illustrates the scope of costs associated with (low) information quality.

1. higher maintenance costs [3, 24, 9]
2. excess labor costs [15, p. 32]
3. higher search costs [27]
4. assessment costs [20]
5. data re-input costs [23, p. 87]
6. time costs of viewing irrelevant information [21]
7. loss of revenue [31]
8. costs of losing current customer [17]
9. costs of losing potential new customer. [ibid., p.205]
10. 'loss of orders' costs [ibid., p. 207]
11. higher retrieval costs [4, p. 316]
12. higher data administration costs [9]
13. general waste of money [31]
14. costs in terms of lost opportunity [ibid.]
15. costs due to tarnished image (or loss of goodwill) [ibid.]
16. costs related to invasion of privacy and civil liberties [ibid.]
17. costs in terms of personal injury and death of people [ibid.]
18. costs because of lawsuits [ibid.]
19. Process failure costs [8]
20. information scrap and rework costs [ibid.]
21. lost and missed opportunity costs [ibid.]
22. costs due to increased time of delivery [17]
23. costs of acceptance testing [8]

Table 1: Costs resulting from low quality data: examples at various levels of abstraction

As a second preliminary result, we have generated a list of direct costs associated with assuring data quality. Ten examples of such costs associated with raising information quality are listed in Table 2. In a cost-benefit analysis these costs would have to be compared with the sum of the costs listed in Table 1.

The examples in Table 2 illustrate that improving information quality results in different types of costs (such as training, infrastructure, or administrative costs) that occur at different stages of information quality management (e.g., at the prevention phase, assessment phase, the detection phase, the repairing phase, or the improvement phase).

1.	Information quality assessment or inspection costs [8]
2.	Information quality process improvement and defect prevention costs [ibid.]
3.	Preventing low quality data [31]
4.	Detecting low quality data [ibid.]
5.	Repairing low quality data [ibid.]
6.	Costs of improving data format [9]
7.	Investment costs of improving data infrastructures [ibid.]
8.	Investment costs of improving data processes [ibid.]
9.	Training costs of improving data quality know-how [ibid.]
10.	Management and administrative costs associated with ensuring data quality [ibid.]

Table 2: Cost examples of assuring data quality

As a third result, we have examined both the cost examples and the underlying classification criteria in these examples in order to combine the results in a listing of possible data quality cost classifications. The various data quality cost classifications are listed in Table 3. Table 3 illustrates the fact that data quality costs can be categorized in terms of several informative dimensions, each highlighting other facets and requiring different measurement methods (such as measuring at the source of the problem or where its effects are visible).

A. DATA QUALITY (DQ) COSTS BY ORIGIN / DATA QUALITY LIFE CYCLE COSTS
Costs are categorized in terms of their origination (where the costs are caused along the information life cycle):

- Costs due to incorrect capture or entry (time & effort to identify incorrect entries, repairing wrong entries, informing about capture modifications)
- Costs due to incorrect processing
- Costs due to incorrect distribution or communication
- Costs due to incorrect re-capture/re-entries
- Costs due to inadequate aggregation (e.g., inconsistent aggregations)
- Costs due to inadequate deletion (e.g., data loss)
- Etc.

B. DQ COSTS BY EFFECT
Costs categorized in terms of their effects (where the costs are actually covered):

- Costs of lost customers for marketing
- Costs of scrap and re-work in production
- Costs of identifying bad data in operations
- Costs of re-entry at data capture point
- Costs of screening at data use points
- Costs of tracking mistakes
- Costs of processing customer data complaints
- Etc.

C. DQ COSTS BY INFORMATION QUALITY ATTRIBUTE
Criterion: what missing IQ attributes are driving costs.

- Costs due to untimely arrival of information (e.g., missed opportunity)
- Costs due to inaccurate information (correction costs)
- Costs due to inaccessible information (higher information gathering costs)
- Costs due to inconsistent information (higher checking and comparing costs)
- Costs due to unreliable information (checking costs)
- Etc.

D DQ COSTS BY EVOLUTION
Criterion: the evolution of the cost as data quality improvement spendings are increased.

- Decreasing data costs
- Marginal data costs
- Increasing data costs
- Fixed data costs
- Variable data costs
- Exponential data costs

Other possible DQ cost categorizations include those based on the following criteria:

- Direct costs vs. indirect costs
- Avoidable costs vs. unavoidable costs
- By impact: major, minor, neglectable, etc.
- One-time vs. ongoing costs
- Variable costs vs. fixed costs
- Visible costs vs. invisible costs
- Occurring (actual) costs vs. dormant (latent) costs
- Proportional costs vs. non-proportional costs
- Short term costs vs. long run costs, current vs. past costs (out of pocket costs, historic costs, etc.)
- Monetary costs vs. opportunity costs
- Controllable vs. incontrollable costs
- Quantifiable vs. non quantifiable costs

Table 3: Possible classifications of data quality related costs

What the previous three tables have made clear, is that data quality costs consist of two major types, namely improvement costs and costs due to low data quality. Whereas improvement costs can be categorized along the information quality process (from prevention, detection, to repair), costs due to low quality data can be categorized in terms of their measurability or impact, e.g., direct versus indirect costs. Combining these two insights, we can devise a simple classification of data quality costs that is more comprehensive than the one presented in Figure 1.

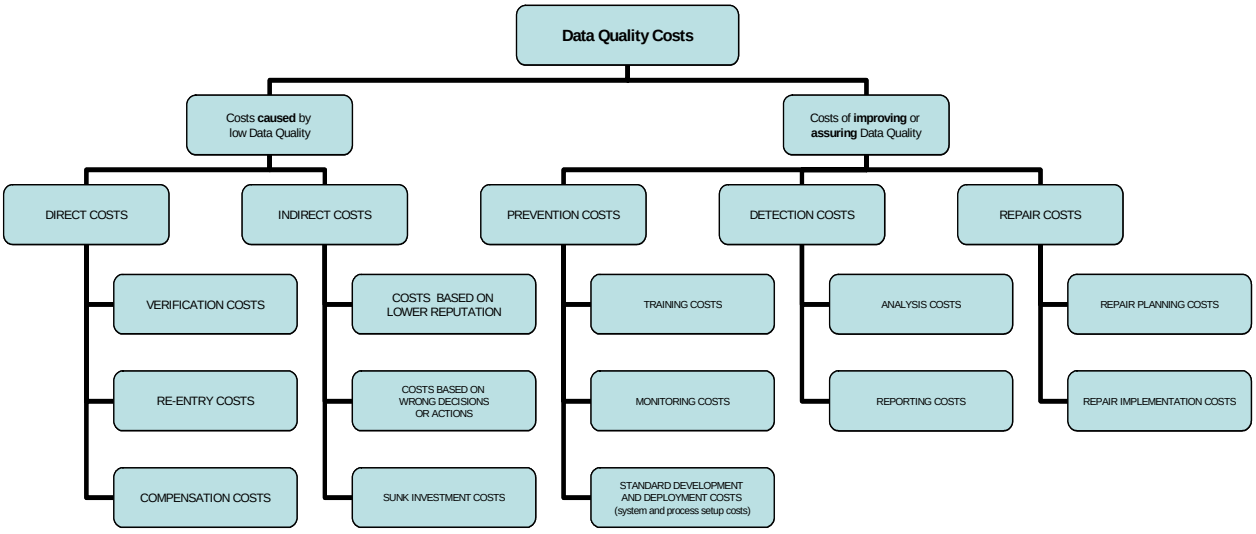

Figure 2: A data quality cost taxonomy

The Tables 1 through 3 have been used to generate the initial data quality taxonomy that is based on the core distinction of costs due to low data quality versus costs caused by data quality assurance measures (Figure 2). In the low quality data cost section, the key distinction is, as stated, the one among direct costs and indirect costs. Direct costs are those negative monetary effects that arise immediately out of low data quality, namely the costs of verifying data because it is of questionable credibility, the costs of re-entering data because it is wrong or incomplete, and the costs of compensation for damages to others based on bad data. Indirect costs are those negative monetary effects that arise, through intermediate effects, from low quality data. Such indirect costs are loss of a price premium because of a deteriorating reputation, the costs incurred by sub-optimal decisions based on bad data, or the investment costs that have to be written off because of low quality data. In terms of costs that arise in order to improve data quality (that is to say to lower the costs of low data quality), we distinguish among prevention, detection, and (one-time) repair costs. While this classification is informative, it cannot yet be used for the pro-active cost management of information quality and the cost-benefit analysis of information quality programs. The following conceptual framework should be a step in this direction.

The derived conceptual framework depicted in Figure 3 uses a life cycle approach to distinguish between high DQ costs and low DQ costs that must be continually compared in an information quality program assessment. At the data entry level, preventive and corrective costs must be compared over time. It must be analyzed, whether investments in preventive costs lead to a reduction in corrective costs. In the data processing phase, the framework distinguishes between process improvement costs and re-processing costs. Again costs in process improvements should lead to cost reductions in re-processing. For the final phase, data use, we distinguish among two cost and benefit types that must be measured, quantified and compared: direct and indirect low IQ costs versus direct and indirect high IQ benefits. In a cost-benefit analysis, the upper part of the framework must thus be continually compared with the lower one to determine the overall cost impact of an information quality program.

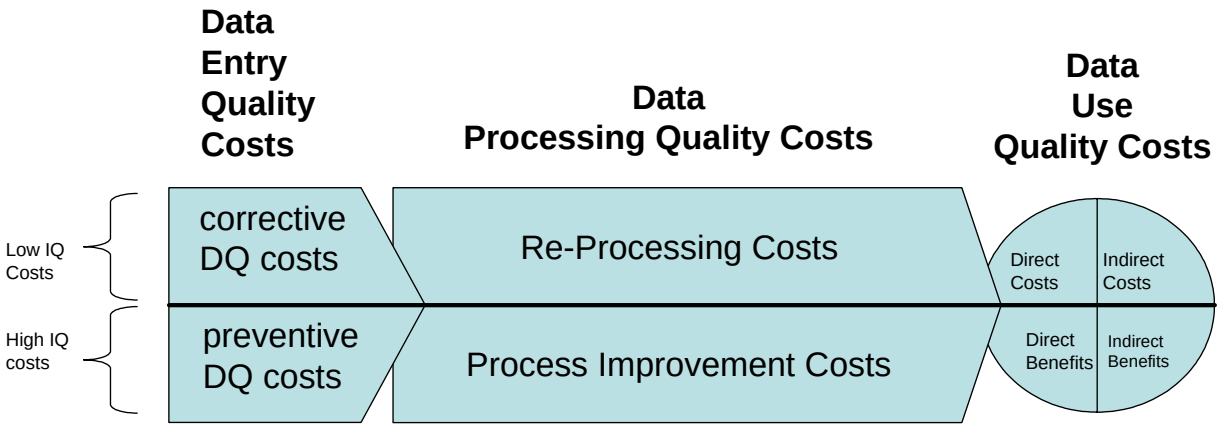

Figure 3: A framework for a macro classification of data quality costs

To conduct these cost comparisons, we must next look at different application scenarios and at the non-trivial relationship between low and high DQ costs.

DISCUSSION

The framework presented in the previous section can be used in a variety of business and research situations. Such scenarios are explored in this section. Specifically, we suggest four scenarios in which a data quality classification can be useful.

The first scenario is *Data Quality Risk Assessment.* Before investing in a data quality project or initiative (even before putting together a business case), a company may want to examine the potential risks associated with low quality data in order to better position the issue within its corporate context. Instead of an undirected, heuristic search for possible data quality mine fields (e.g., based on past experiences and events), the presented taxonomy and framework outlines examples of what to look for. The direct and indirect data quality costs can be examined in terms of their likelihood and effect, thus contributing to an overall risk assessment of low data quality in a company.

The second application scenario for the framework is, as stated initially, the *Data Quality Business Case*. New IT initiatives typically have to prove their feasibility by outlining how the invested money will yield benefits for a company in terms of time-optimization, higher quality levels, or lower costs. An IT analyst or prospective data quality project manager can use the framework to list such potential costs that are going to be reduced because of the data quality project.

A third application possibility for the data quality cost classification is *Data Quality Program Assessment.* Whereas business cases are ex-ante estimates of the cost benefits of a project, assessments are after action reviews that show where and how costs have been lowered because of an initiative. In this context, the framework can be used to outline all possible cost reduction effects that haven taken place as a direct or indirect result of a data quality initiative.

A last important application scenario for a data quality cost classification is *Benchmarking*. Whether in research or in practice, comparing data quality cost levels among organizations is an important objective. Based on benchmarking figures, companies can set more realistic (and competitive) goals for their data quality levels. Based on consistent benchmarking information, researchers can find correlations and causalities that show what the drivers for data quality costs really are. For both target groups, however, a consistent taxonomy and terminology is essential.

ANALYSING THE COST OF DATA QUALITY

Even if this article does not aim to provide a comprehensive data quality cost theory, we intend to illustrate the application of our proposed taxonomy in one scenario and contribute first elements of a data quality cost model, which provides the conceptual starting point for data quality cost benefit considerations. However, as there are many claims regarding the importance of data quality costs in general, and cost benefit analysis in particular, there is currently no validated economic theory of data quality costs that could be used as a basis for data quality cost analysis. So far there are only a few approaches analyzing the cost and benefit structure from an economic perspective, as for instance in [6,18]. Nevertheless, the concept of quality costing is not new. There are many approaches of cost of quality. Most of these approaches are at present solely used in the context of manufacturing. For this reason, and in order to build a data quality cost model, we first briefly review major quality cost approaches in the context of manufacturing. This provides the basis for linking our cost classifications to current quality cost theory and consequently helps to develop first elements for a data quality cost model.

One of the first quality cost models was developed by [10]. Its basis is the classification of quality costs into three main categories: prevention, appraisal and failure costs. [7] developed the concept of a process

oriented quality cost model, which determines the total cost of quality as the sum of costs of achieving quality (costs of conformance) and the costs due to a lack of quality (costs of non-conformance). To illustrate the contribution of each of the quality cost elements, costs of quality models are used as a conceptual basis. Costs of conformance and costs of non-conformance have typically an inverse relationship to one another: As investments in achieving quality increases, the costs due to lack of quality decreases. These effects are normally shown as curves in relation to a quality level (expressed as a value of a quality metric). In the traditional cost of quality model, the total cost of quality has a point of diminishing returns, i.e. a minimum prior to achieving the maximum quality. Figure 4 illustrates this scenario of a cost of quality model, which has a total cost minimum prior to achieving 100% of quality. However, manufacturing experience has shown that increased attention to defect prevention leads to large reduction in appraisal costs. This experience is reflected in modern total quality management philosophies with a focus on continuous improvement and the aim of highest quality. The corresponding cost of quality model typically corresponds to the curves in the gray area in Figure 4 [30]. As illustrated, the idealistic model in the gray area concludes that the optimum total costs can be shifted toward 100% of the quality metric [11, 12]. The observation of these two seemingly conflicting propositions led to study of dynamics in the cost model in [30], in which time was considered as third dimension.

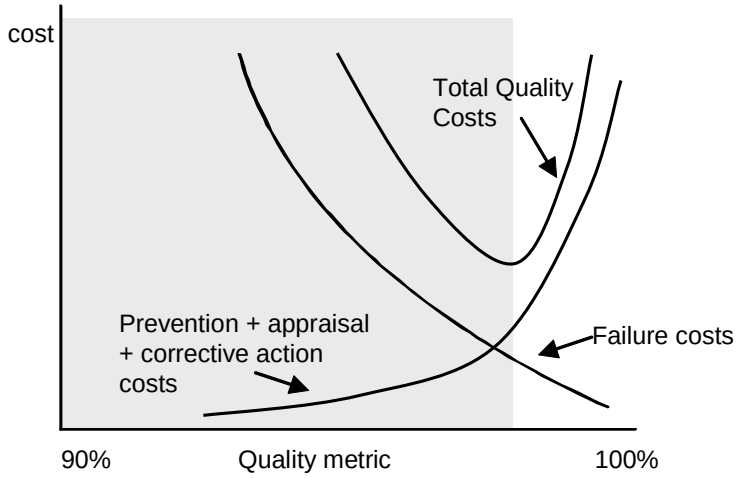

Figure 4: Conceptual relation between quality costs and quality [30]

At present most of these quality cost models are solely used in the context of manufacturing and service industries, which are – compared to data quality – well understood and analyzed. For these models, there are many assumptions and realistic cost estimations. However, the key question of quality cost models in the context of data quality still remains: Are the assumptions and the cost curve progressions observed in manufacturing similar for data quality management? In particular, does the concept of diminishing returns also apply to data quality?

First, let us independently characterize the (static) progression in relation to the level of data quality (e.g. from low data quality to high data quality) of the data quality cost categories with omitting initially prevention measures. Unfortunately at present there is not much credible data regarding data quality costs available, which makes a cost characterization difficult. In addition, the actual cost curve progression is highly application dependent and in practice extremely complicated to estimate. Much depends on the implemented information system and subjective estimation of data consumers. But from experiences in various data quality projects, the following observations seem reasonable:

- Costs caused by low data quality: These costs depend highly on the subjective estimation of data consumers and its context. But realistically the costs caused by low data quality should be

expected to decrease with increasing quality (monotonically decreasing). It might be realistic for many contexts to assume that this cost curve is convex. For instance, in a customer database an increase in accuracy of customer records from 85% to 90% might contribute to the company's reputation more than an increase from 90% to 95% respectively (thus monotonically decreasing gradient). However this might not be true for other contexts like healthcare, where due to the severe consequences, a linear (monotonically decreasing!) relationship between costs caused by low quality and quality level could be possible. For instance a reduction of the number of incorrect prescriptions could directly reduce patients' health risk caused by wrong prescriptions.

- Repair costs: These costs are zero at minimum or no quality and rise to its maximum at maximum quality. In manufacturing normally a convex curve is assumed. Similar can be stated in the context of data quality, in which due to the rather limited capabilities of automatic data cleansing methods humans play an important role in the data creation process. For example by applying business rules (data constraints) some, but not all data input errors can be automatically detected and corrected. Manual methods are required [31], which are typically increasingly expensive towards higher levels of quality, and consequently result in a convex curve for repair costs (at least for a relatively high data quality level).
- Detection costs: For these costs similar consideration as for repair costs can be assumed. However it seems reasonable that at a relatively high data quality level lower detection costs than repair costs arise. This is due to the reason that tool-based detection of data defects is further possible even if data defect corrections have to be carried out manually. Consistency rules, for instance, can detect inconsistencies in databases. Actual reconciliation, however, has to be carried out manually by humans.

Total data quality costs are determined from the summarization of all involved costs. In a first step, let us now illustrate a situation, in which prevention measures are not implemented. Figure 5 illustrates the total costs as sum of costs caused by low data quality, detection costs and repair costs. As shown, depending on the particular cost characteristics an optimum of total data quality is reached at a point before maximum quality. This is the point costs caused by low data quality and costs for assuring data quality (without prevention!) are balanced. From an economic perspective this quality level should be achieved.

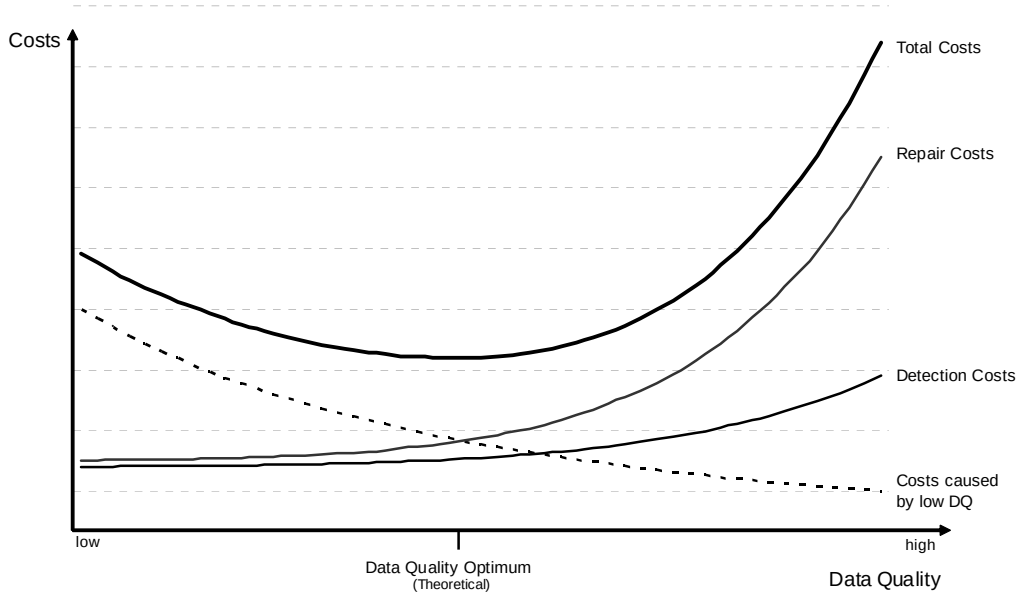

Figure 5: Model for optimum data quality (initial situation without prevention)

Now, is it possible to improve the situation above and get beyond the initial data quality optimum? What effects have prevention measures? In manufacturing, it is argued that preventing quality defects results in significant reduced repair and detection costs (e.g. [12, 26]). This is due the assumption, that the sooner a defect is detected or prevented, the more savings can be made during subsequent processes. Experiences in software engineering observed similar effects. For instance [19] could show that the cost of defect fixing grow dramatically when the defects are found in a later development stage, i.e. when the defects have already influenced the next phase. Therefore they concluded that a substantial amount of costs will be saved by introducing systematic defect detection early in a software project. Similar effects are also observed in data quality improvement projects, in which causes of data quality problems are systematically eliminated. Examples are for instance described in [8] or [22]. Unfortunately, considering the limited research in data quality costs, currently it is (still) unfeasible to quantify the effects of data quality prevention measures on repair and detection costs. But as the experience in manufacturing, software engineering and data quality projects suggest, we assume that increased prevention costs result in significant savings of repair and detection costs. In addition, this can be further justified considering the increasing complexity of informational environments with numerous data duplications and interlinked data chains.

A second effect can be observed by the difference between proactive prevention measures and the rather different reactive repair actions. To prevent data defect repetition, prevention measures need the implementation of permanent system changes (improved processes, software systems, organizational structures, learning). Typically these measures require substantial long-term investments in the quality system with the effect of an overall data quality improvement. In our model we can represent this with a very low increasing prevention cost curve (e.g. gradient close to zero for a relatively high data quality level). In addition, and similar to prevention measures, many of the repair and detection measures are associated with permanent costs savings due to the permanent reduction in data quality defects. Such effects can be observed for instance with the implementation of data cleansing routines in the ETL-process in Data Warehouse systems. Consequently, these costs are prevention costs, which should be represented in the prevention cost curve. This observation directs us to the second effect of introducing prevention measures. Once the quality improvement through prevention measures is sustained, expenses to maintain the system (e.g. repair and detection costs) should be reduced in following periods [30]. As illustrated in Figure 6, these cost savings of introducing prevention measures are reflected in a reduced gradient (observed as shift to the right) of both curves over time.

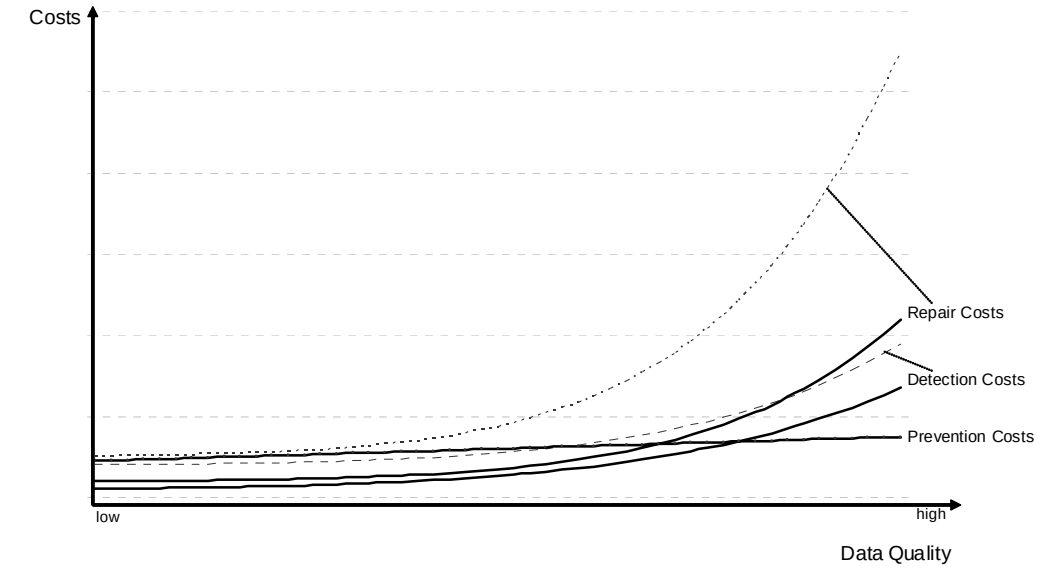

Figure 6: The effect of prevention on repair and detection costs

The overall picture is illustrated in Figure 7, which includes the costs caused by low data quality and summarizes the effect of prevention costs on the total costs. Due to the typical cost progression of prevention costs (low gradient) and its effect on repair and detection costs (shift to right with reduced gradient) the total costs curves shifts to the right (thus results in a reduced gradient). Consequently the optimum data quality moves to the right, resulting in a higher to achieving data quality. Clearly this effect depends on the particular cost structure and dependency of prevention, detection and repair costs, but the dynamic model (originally developed for manufacturing by [30]) illustrates that the optimum data quality level could move in the direction towards improvement over time.

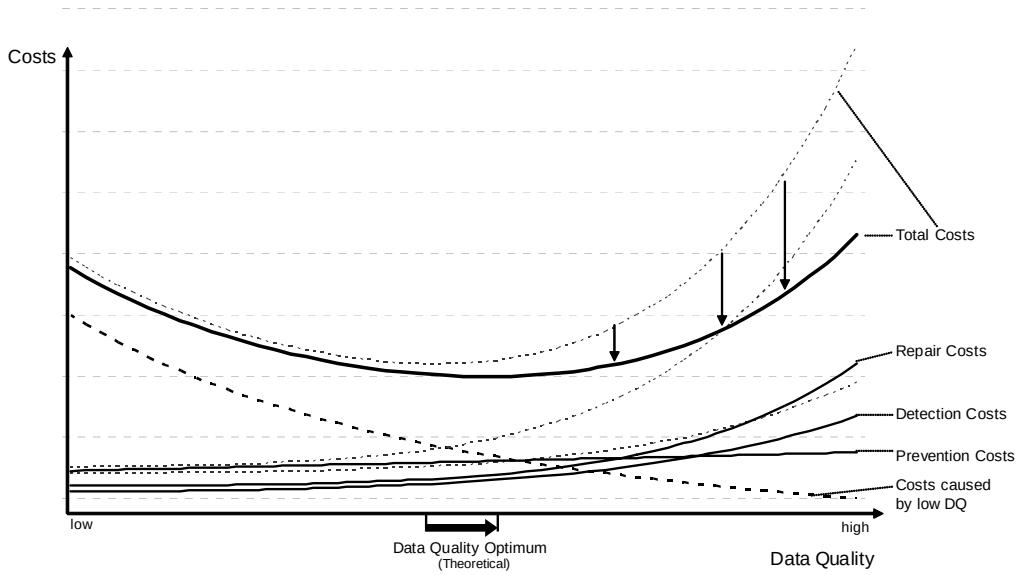

Figure 7: A model for optimum data quality (considering prevention costs)

In summary, this first element of a data quality cost model illustrates, similar to models in manufacturing, the dynamics of different data quality cost categories. As we have discussed earlier, the predominant role of employees in the data creation process and the limited capability of data quality tools result in exponentially increasing costs for improving and assuring data quality. This results in a situation in which data quality is at a point below its maximum (achievable) quality. By introducing prevention measures, the initial situation can be significantly improved and the quality level can be dynamically increased towards a maximum of data quality. Nevertheless, the assumptions of the model should be validated in additional studies by further detailing and mapping the cost curves to the proposed data quality taxonomy. However, in order to achieve an economic data quality optimum, the model illustrates the importance of continually comparing corrective and preventive data quality costs in an information quality program.

LIMITATIONS

The framework and data quality cost analyses presented in the previous sections has several advantages, but also different drawbacks; both are described in this section. In terms of advantages, our discussion of cost increases the scope for data quality costs that have not always been considered when constructing a data quality business case. It also allows us to switch perspectives, examining costs through different categorization criteria. However, the presented framework is not a testable model. Clearly, the present distinctions are only prolegomena for the development of indicators, scorecards and the like. Major barriers towards the creation of such measurement systems are the missing causal links that are still a subject for further research.

Our first elements of a data quality cost model (the analysis section) illustrated how data quality can be dynamically improved by implementing prevention measures. However as our first model at present is based on experiences made in data quality projects, further empirical studies are necessary to evaluate and possibly quantify the effects of introducing prevention measures on detection and repair costs. Carrying out this further research an interesting question about the “optimal” mix between reactive repair and proactive prevention measures could be addressed. Our conceptual model of data quality costs also provides the basis for explaining other effects, like for instance that underestimated costs caused by low data quality result in a data quality optimum at a lower level (than actually from an economical point expected). This is often observed in data quality projects in practice. So far, our first data quality cost model omitted that quality expectations from data quality consumers often dynamically increase, which would led to a further increase in data quality. Further research should integrate this effect in the data quality cost model.

CONCLUSION

If the data quality field is to make significant progress in terms of its acceptance in the business world, the costs associated with low data quality must be made more explicit, prominent, and measurable. They most be compared to the cost of assuring data quality, so that an optimal investment point for data quality can be approximated. A systematic method for data quality *cost benefit analysis* can help companies to determine such an optimal level of investment in data quality. Today, however, we are very far from such a methodology to calculate the optimal level of data quality. One reason for this is the lack of overview of all relevant data quality costs, either the costs of assuring data quality or the costs of low quality data. By classifying data quality costs, we can open a diagnostic perspective that is both systematic and informative. This paper has made a first step in this direction by providing an overview on such possible cost classifications and by analyzing their mutual influence and their progression. Future research should strive to further consolidate these classifications and validate our cost progression models, ideally through real-life observations in the data quality field.

REFERENCES

[1] Bailey, K.D., Typologies and Taxonomies: An Introduction to Classification Techniques, Thousand Oakes: Sage, 1994.

[2] Ballou, D.P., Wang R. Y., Pazer, H. and Tayi, G.K., Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.

[3] Barnes, D.K., Managing Networks containing cumb devices with intelligent agents, in: Klein, B.D., Rossin, D.F. (Eds.) *Proceedings of the 2000 Conference on Information Quality*, 2000, pp. 308-317.

[4] Bovee, M., Srivastava, R.P., Mak, B., A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality, in: Pierce, E., Katz-Haas, R. (Eds.): *Proceedings of the Sixth MIT Information Quality Conference*, 2001, Boston, pp. 311-324.

[5] Bowker, G.C. Leigh Star, S., *Sorting Things Out, Classification And Its Consequences.* The MIT Press. Cambridge, Massachusetts, 1999.

[6] Cappiello, C., Francalanci, C., *DL4.B Considerations about Costs Deriving from a Poor Data Quality*, DaQuinCIS Project Report, December 2002, Available online at http://www.dis.uniroma1.it/~dq/docs/Deliverables/DL4B.pdf.

[7] Crosby, P.B., *Quality is Free*, New York, McGraw-Hill, 1979.

[8] English, L., *Improving Data Warehouse and Business Information* Quality. New York: Wiley & Sons: , 1999.

[9] Eppler, M., *Managing Information Quality*, New York/ Berlin: Springer, 2003.

[10] Feigenbaum, A.V., Total Quality control, *Harvard Business Review*, 1956, 34 (6), pp. 93-101.

[11] Gryna, F.M., Quality Costs, in Juran, J.M., Gryna, F.M. (Eds.), *Juran's Quality Control Handbook*, 4th Ed., New York: McGraw-Hill, 1998, page 4.2.

[12] Gryna, F.M., Quality and Costs, in Juran, J.M., Godfrey, A.B (Eds.), *Juran's Quality Handbook*, 5th Ed. New York: McGraw-Hill, 1998, page 8.1

[13] Huang, K., Lee, Y. and Wang R.Y., *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.

[14] Irani, Z., Love, P.E.D., The Propagation of Technology Management Taxonomies for evaluating investements in information systems, in: *Journal of Management Information Systems*, Winter 2000/2001, Vol. 17, Issue 3.$

[15] Kahn, B. K., Katz-Haas, R. and Strong, D.M., How to get an Information Quality Program Started: The Ingenix Approach, in: Klein, B.D., Rossin, D.F. (Eds.) *Proceedings of the 2000 Conference on Information Quality*, 2000, pp. 28-35.

[16] Kengpol, A., The Implementation of Information Quality for the Automated Information Systems in the TDQM Process: A Case Study in Textile and Garment Company in Thailand in: Pierce, E., Katz-Haas, R. (Eds.): *Proceedings of the Sixth MIT Information Quality Conference*, 2001, Boston, pp. 206-216.

[17] Lesca, H., Lesca, E., *Gestion de l'information, qualité de l'information et performances de l'entreprise*, 1995, Paris: Litec.

[18] Mandke, V.V., Nayar, M.K. Cost Benefit Analysis of Information Integrity, in Fisher, C., Davidson, B. N. (Eds.) *Proceedings of the Seventh International Conference on Information Quality, 2002*, Boston, pp. 119-131.

[19] Megen, van, R., Meyerhoff, D.B, Costs and benefits of early defect detection: experiences from developing client server and host application, in *Software Quality Journal*, (4) 4, 1995, pp. 247-256.

[20] Naumann, F., Rolker, C., Assessment Methods for Information Quality Criteria, in: Klein, B.D., Rossin, D.F. (Eds.) *Proceedings of the 2000 Conference on Information Quality*, 2000, pp. 148-162.

[21] Neus, A., Managing Information Quality in Virtual Communities of Practice: Lessons learned from a decade's experience with exploding Internet communication, in: Pierce, E., Katz-Haas, R. (Eds.): *Proceedings of the Sixth MIT Information Quality Conference*, 2001, Boston, pp. 119-131.

[22] Redman, T. C., *Data quality for the information age*, Boston, MA: Artech House, 1996.

[23] Segev, A., Wang, R., Data Quality Challenges in Enabling eBusiness Transformation, in: Pierce, E., Katz-Haas, R. (Eds.): *Proceedings of the Sixth MIT Information Quality Conference*, 2001, Boston, pp. 83-91.

[24] Strong, D. M., Lee, Y. W. and Wang, R. Y., 10 Potholes in the Road to Information Quality, in: *Computer IEEE*, 30 (8), 1997, pp. 38-46.

[25] Strong, D. M., Lee, Y.W. and Wang, R.Y., Data Quality in Context, in: *Communications of the ACM*, 40 (5), 1997, pp. 103-110.

[26] Teboul, J., *Managing Quality dynamics*, Upper Saddle River: N.J.: Prentice Hall., 1991.

[27] Verykios, V.S., Elfeky, M.G., Elmagarmid, A.K., Cochinwala, M., Dalal, S., On the Accuracy and Completeness of the Record Matching Process, in: Klein, B.D., Rossin, D.F. (Eds.) *Proceedings of the 2000 Conference on Information Quality*, 2000, pp. 54-

[28] Wang, R. Lee, Y.W., Pipino, Strong, D., Manage Your Information as a Product, in: *Sloan Management* Review, 39 (4), 1998, pp. 95-105.

[29] Wang, R. Y.; Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers, in: *Journal of Management Information Systems*, 12 (4), Spring 1996, pp. 5 – 33.

[30] Wasserman, G.S, Lindland, J.L., A Case Study illustrating the existing of dynamics in traditional cost-of-quality Models, in: *Quality Engineering* 9 (1), 1996, pp. 119-128.

[31] Won, K., Choi, B., Towards Quantifying Data Quality Costs", in *Journal of Object Technology*, 2 (4), July-August 2003, pp. 69-76 http://www.jot.fm/issues/issue_2003_07/column6