Journal of Industrial Engineering and Management

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953

Print ISSN: 2013-8423

The costs of poor data quality

Anders Haug, Frederik Zachariassen, Dennis van Liempd

University of Southern Denmark (DENMARK)

adg@sam.sdu.dk; frz@sam.sdu.dk; dvl@sam.sdu.dk

Received August 2010
Accepted January 2011

Abstract:

Purpose: The technological developments have implied that companies store increasingly more data. However, data quality maintenance work is often neglected, and poor quality business data constitute a significant cost factor for many companies. This paper argues that perfect data quality should not be the goal, but instead the data quality should be improved to only a certain level. The paper focuses on how to identify the optimal data quality level.

Design/methodology/approach: The paper starts with a review of data quality literature. On this basis, the paper proposes a definition of the optimal data maintenance effort and a classification of costs inflicted by poor quality data. These propositions are investigated by a case study.

Findings: The paper proposes: (1) a definition of the optimal data maintenance effort and (2) a classification of costs inflicted by poor quality data. A case study illustrates the usefulness of these propositions.

Research limitations/implications: The paper provides definitions in relation to the costs of poor quality data and the data quality maintenance effort. Future research may build on these definitions. To further develop the contributions of the paper, more studies are needed.

Practical implications: As illustrated by the case study, the definitions provided by this paper can be used for determining the right data maintenance effort and

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

costs inflicted by poor quality data. In many companies, such insights may lead to significant savings.

Originality/value: The paper provides a clarification of what are the costs of poor quality data and defines the relation to data quality maintenance effort. This represents an original contribution of value to future research and practice.

Keywords: data quality, master data management, data quality costs

1 Introduction

Data are used in almost all the activities of companies and constitute the basis for decisions on operational and strategic levels. Poor quality data can, therefore, have significantly negative impacts on the efficiency of an organization, while high quality data are often crucial to a company's success (Madnick et al., 2004; Haug et al., 2009; Batini et al., 2009; Even & Shankaranarayanan, 2009). However, several industry expert surveys indicate that data quality is an area, to which many companies seem not to give sufficient attention or know how to deal with efficiently (Marsh, 2005; Piprani & Ernst, 2008; Jing-hua et al., 2009).

Vayghan et al. (2007) classify the data that most enterprises deal with in three categories: master data, transactional data, and historical data. Master data are defined as the basic characteristics of business entities, i.e. customers, products, employees, suppliers, etc. Thus, typically, master data are created once, used many times and do not change frequently (Knolmayer & Röthlin, 2006). Transaction data describe the relevant events in a company, i.e. orders, invoices, payments, deliveries, storage records etc. Since transactions are based on master data, erroneous master data can have significant costs, e.g. an incorrect priced item may imply that money is lost. In this context Knolmayer and Röthlin (2006) argue that capturing and processing master data are error-prone activities where inappropriate information system architectures, insufficient coordination with business processes, inadequate software implementations or inattentive user behaviour may lead to disparate master data.

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

In spite of the importance of having correct and adequate data in a company, there seems to be a general agreement in literature that poor quality data is a problem in many companies. In fact, much academic literature claims that poor quality business data constitute a significant cost factor for many companies, which is supported by findings from several surveys from industrial experts (Marsh, 2005). On the other hand, Eppler and Helfert (2004) argue that although there is much literature that claims that the costs of poor data quality are significant in many companies, only very few studies demonstrate how to identify, categorize and measure such costs (i.e. how to establish the causal links between poor data quality and monetary effects). This is supported by Kim and Choi (2003) who state: "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." and "In practice, low quality data can bring monetary damages to an organization in a variety of ways". According to Kim (2002), the types of damage that low quality data can cause depend on the nature of data, the nature of the use of data, the types of responses (by the customers or citizens) to the damages, etc.

As such, companies typically incur costs from two sides when speaking of master data quality. Firstly, companies incur costs when cleaning and ensuring high master data quality. Secondly, companies also incur costs for data that are not cleaned as poor master data quality might lead to faulty managerial decision-making. The purpose of this paper is to provide a better understanding of the relationship between such costs. To help determine the optimal data quality maintenance efforts, the paper provides: (1) a definition of the optimal data maintenance effort; and (2) a classification of costs inflicted by poor quality data. In this context the paper argues that there is a clear trade-off relationship between these two cost types and that the task facing the companies in turn is to balance this trade-off.

The remainder of the paper is organized as follows: First, literature on data quality is discussed in Section 2. Next, Section 3 proposes a model to determine the optimal data maintenance effort and a classification of different types of costs inflicted by poor quality data. Section 4 presents a case study to illustrate the application of the proposition. The paper ends with a conclusion in section 5.

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

2 Data quality literature

Firstly, this section makes a clarification of the term 'data quality' and then provides a fundamental understanding of the impacts of poor quality data. Finally, the section discusses existing models of the relationship between data maintenance effort and costs inflicted by poor quality data.

2.1 Data quality

To understand the concept of ‘data quality’, to begin with a distinction between data, information and knowledge may be appropriate. Popular definitions of such terms have been made by Davenport and Prusak (1998), who define data as “discrete, objective facts about events” and information as data transformed by the value-adding processes of contextualization, categorization, calculation, correction and condensation. Similar definitions are provided by Newell et al. (2002), who define data as “providing a record of signs and observations collected from various sources” and information as when “data are presented in a particular way in relation to a particular context of action”. In contrast to ‘data’ and ‘information’, the meaning of ‘knowledge’ is much more debatable, which is a discussion often relating to whether knowledge is perceived as being of an impersonal and static nature or being personal and related to action (Newell et al., 2002). However, a deeper discussion about the meaning of the meaning of ‘knowledge’ is beyond the scope of this paper, which, as mentioned, focuses on data quality.

Data quality is often defined as 'fitness for use', i.e. an evaluation of to which extent some data serve the purposes of the user (e.g. Lederman et al., 2003; Tayi & Ballou, 1998; Watts & Shankaranarayanan, 2009). Another way to understand the concept of data quality is by dividing it into subcategories and dimensions. An often cited definition is provided by Ballou and Pazer (1985), who divide data quality into four dimensions: accuracy, timeliness, completeness, and consistency. They argue that the accuracy dimension is the easiest to evaluate as it is merely a matter of analysing the difference between the correct value and the actual value used. They also argue that the evaluation of timeliness can be carried out in a similar unproblematic manner. As for the evaluation of the completeness of some data, this can also be done relatively straight forward, as long as the focus is on whether the data are complete or not in contrast to defining the level of completeness, e.g. the percentage of data completeness. On the other hand, an

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

evaluation of consistency is a little more complex, since this requires two or more representation schemes in order to be able to make a comparison.

Another data quality classification is provided by Wand and Wang (1996). They limit their focus to intrinsic data qualities, of which they define four intrinsic dimensions: completeness, unambiguousness, meaningfulness and correctness. Wand and Wang (1996) take as their basis a paper, which features a review of cited data quality dimensions, i.e. the comprehensive literature review of Wang et al. (1995). Wang et al. (1995) summarize the most often cited data quality dimensions as shown in Table 1.

Accuracy	25	Flexibility	5	Sufficiency	3	Informativeness	2
Reliability	22	Precision	5	Usableness	3	Level of detail	2
Timeliness	19	Format	4	Usefulness	3	Quantitativeness	2
Relevance	16	Interpretability	4	Clarity	2	Scope	2
Completeness	15	Content	3	Comparability	2	Understandability	2
Currency	9	Efficiency	3	Conciseness	2		
Consistency	8	Importance	3	Freedom from bias	2		

Table 1. “Cited data quality dimensions”. Source: Wang et al. (1995).

Wang and Strong (1996) propose a data quality classification which divides data quality into four categories: intrinsic, contextual, representational, and accessibility. For each category they define a set of dimensions, 18 in all. The definition by Wang and Strong (1996) is discussed by Haug et al. (2009) who argue that 'representational data quality' can be perceived as a form of 'accessibility data quality' instead of a category of its own. Thus, Haug et al. (2009) define three data quality categories: intrinsic, accessibility and usefulness. Levitin and Redman (1998) provide another perspective by arguing that since processes to produce data have many similarities to processes that produce physical products, data producing processes could be viewed as producing data products for data consumers. With a basis in this view of data as resources, Levitin and Redman discuss how thirteen basic properties of organizational resources may be translated into properties for data.

2.2 Impacts of poor quality data

The development of information technology during the last decades has enabled organizations to collect and store enormous amounts of data. However, as the data volumes increase, so does the complexity of managing them. Since larger and

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

more complex information resources are being collected and managed in organizations today, this means that the risk of poor data quality increases (Watts & Shankaranarayanan, 2009). Another often mentioned data related problem is that companies often manage data at a local level (e.g. department or location). This implies the creation of 'information silos' in which data are redundantly stored, managed and processed (Lee et al., 2006; Smith, 2008; Vayghan et al., 2007). In this vein, Lee et al. (2006) argue that data silos imply that many companies face a multitude of inconsistencies in data definitions, data formats and data values, which makes it almost impossible to understand and use key data. From a solution perspective, ERP systems have been promoted as a panacea for dealing with the lack of data integration by replacing inadequately coordinated legacy systems (Davenport, 1998; Knolmayer & Röthlin, 2006). However, it has been suggested that data problems may get intensified when using ERP systems since the ERP modules are intricately linked to each other, which is the reason why poor quality data input in one module can affect the functioning of other modules negatively (Park & Kusiak, 2005).

Poor quality data can imply a multitude of negative consequences in a company. To start with, poor quality data that is not identified and corrected can have significantly negative economic and social impacts on an organization (Ballou et al., 2004; Wang & Strong, 1996). The implications of poor quality data carry negative effects to business users through: less customer satisfaction, increased running costs, inefficient decision-making processes, lower performance and lowered employee job satisfaction (Kahn et al., 2003; Leo et al., 2002; Redman, 1998). Poor data quality also increases operational costs since time and other resources are spent detecting and correcting errors. Since data are created and used in all daily operations, data are critical inputs to almost all decisions and data implicitly define common terms in an enterprise, data constitute a significant contributor to organizational culture. Thus, poor data quality can have negative effects on the organizational culture (Levitin & Redman, 1998; Ryu et al., 2006). Poor data quality also means that it becomes difficult to build trust in the company data, which may imply a lack of user acceptance of any initiatives based on such data.

When focusing on clarifying the effects of poor quality data, it is clear that many companies experience significant costs as a result of poor quality data, although the exact extent of such costs is difficult to estimate. According to Redman (1998),

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

studies to produce estimates of the total cost of poor data quality have proven difficult to perform. Additionally, data quality research has not yet advanced to the point of having standard measurement methods for any of these issues. On the other hand, Redman (1998) claims that many case studies feature accuracy measures, but he does not provide references or mentions if these are academic studies. According to Redman (1998), measured at the field level, the reported error rates are in the interval of 0.5–30%. Furthermore, Redman (1998) claims that at least three proprietary studies have yielded estimates in the 8-12% of revenue range, but informally 40-60% of the expense of the service organization may be consumed as a result of poor data. Much indicates that the economic effect of even small data inaccuracies can be very significant. Häkkinen and Hilmola (2008) argue that marginal data inaccuracies (e.g. 1-5%) may not necessarily represent a major problem in manufacturing, but that such inaccuracies will have direct effects in terms of lost sales and operational disruptions in the after-sales organizations.

In contrast to the apparent lack of large studies of data quality in academic journal papers (Eppler & Helfert, 2004; Kim & Choi, 2003), many industry experts provide such studies. These industry experts include Gartner Group, Price Waterhouse Coopers and The Data Warehousing Institute, which claim to identify a crisis in data quality management and a reluctance among senior decision-makers to do enough about it (Marsh, 2005). Marsh (2005) summarizes the findings from such surveys into the following bullet-points (quoted from: Marsh, 2005):

- "88 per cent of all data integration projects either fail completely or significantly over-run their budgets"
- "75 per cent of organisations have identified costs stemming from dirty data"
- "33 per cent of organisations have delayed or cancelled new IT systems because of poor data"
- "$611bn per year is lost in the US in poorly targeted mailings and staff overheads alone"

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

- "According to Gartner, bad data is the number one cause of CRM system failure"
- "Less than 50 per cent of companies claim to be very confident in the quality of their data"
- "Business intelligence (BI) projects often fail due to dirty data, so it is imperative that BI-based business decisions are based on clean data"
- "Only 15 per cent of companies are very confident in the quality of external data supplied to them
- "Customer data typically degenerates at 2 per cent per month or 25 per cent annually"
- "Organisations typically overestimate the quality of their data and underestimate the cost of errors"
- "Business processes, customer expectations, source systems and compliance rules are constantly changing. Data quality management systems must reflect this"
- "Vast amounts of time and money are spent on custom coding and traditional methods - usually fire-fighting to dampen an immediate crisis rather than dealing with the long-term problem"

2.3 Data maintenance effort and costs inflicted by poor quality data

As mentioned in the introduction, although there seems to be agreement in literature that the costs of poor data quality are significant in many companies, only very few studies demonstrate how to identify, categorize and measure such costs (Eppler & Helfert, 2004; Kim & Choi, 2003). In practice, low quality data can bring monetary damages to an organization in a variety of ways.

Raman (2000) argues that evidence from previous studies shows that the quality of point-of-sale data is often poor and that even at well-run retailers it cannot be taken for granted. Raman offers a taxonomy of retail-data quality, quantifies these costs to the extent possible, highlights the impact of data quality on Internet retailing, and offers guidelines to managers for improving quality. The focus of the

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

paper, however, is limited to: (1) the direct costs of scanning the wrong price of items, (2) costs associated with inventory-data inaccuracy; and (3) cost of phantom stock-outs. For the first-mentioned the consequence of inaccurate data is simply a subtraction of the sum of overpriced items from the sum of underpriced items. On the costs of inventory-data inaccuracy and phantom stock-out, Raman only offers some estimates related to very specific contexts. Raman recommends two steps to improve data quality, which in headlines can be formulated as: (1) "companies should make greater use of the data that they have stored"; and (2) "that companies start measuring data quality to the extent possible".

Ge and Helfert (2007) analyse three major aspects of information quality research: (1) information quality assessment, (2) information quality management, and (3) contextual information quality. In relation to information quality assessment, among others, Ge and Helfert classify typical information quality problems which are identified by previous research, as shown in Table 2.

	Data Perspective	User Perspective
Context-independent	Spelling error Missing data Duplicate data Incorrect value Inconsistent data format Outdated data Incomplete data format Syntax violation Unique value violation Violation of integrity constraints Text formatting	The information is inaccessible The information is insecure The information is hardly retrievable The information is difficult to aggregate Errors in the information transformation
Context-dependent	Violation of domain constraint Violation of organization's business rules Violation of company and government regulations Violation of constraints provided by the database administrator	The information is not based on fact The information is of doubtful credibility The information presents an impartial view The information is irrelevant to the work The information consists of inconsistent meanings The information is incomplete The information is compactly represented The information is hard to manipulate The information is hard to understand

Table 2. "Classification of information quality problems identified in literature". Source: Ge and Helfert (2007).

On the issue of information quality management, Ge and Helfert (2007) argue that this is an intersection between the fields of quality management, information management and knowledge management. Finally, on the issue of contextual information quality they provide an overview of which publications that relate to different data application contexts, which include: database, information

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

manufacture system, accounting, marketing, data warehouse, decision-making, healthcare, enterprise resource planning, customer relationship management, finance, e-business, World Wide Web and supply chain management.

Data quality costs	Costs caused by low data quality	Direct costs	Verification costs
			Re-entry costs
			Compensation costs
		Indirect costs	Costs based on lower reputation
			Costs based on wrong decisions or actions
			Sunk investment costs
	Costs of improving or assuring data quality	Prevention costs	Training costs
			Monitoring costs
			Standard development and deployment costs
		Detection costs	Analysis costs
			Reporting costs
		Repair costs	Repair planning costs
			Repair implementation costs

Table 3. “A data quality cost taxonomy”. Source: Eppler and Helfert (2004).

Eppler and Helfert (2004) review and categorize the potential costs associated with low quality data. They propose a classification framework and a cost progression analysis to support the development of quantifiable measures of data quality costs for researchers. To address the lack of literature on poor data quality versus costs, according to Eppler and Helfert, “cost classifications based on various criteria can be applied to the data quality field in order to make its business impact more visible”. Based on a literature review, Eppler and Helfert identify 23 examples of costs resulting from poor quality data, which amongst others are: higher maintenance costs, excess labour costs, assessment costs, data re-input costs, loss of revenue, costs of losing current customers, higher retrieval costs, higher data administration costs, process failure costs, information scrap and rework costs and costs due to increased time of delivery. Additionally, Eppler and Helfert identify 10 cost examples of assuring data quality, which are 1) information quality assessment or inspection costs, 2) information quality process improvement and defect prevention costs, 3) preventing low quality data, 4) detecting low quality data, 5) repairing low quality data, 6) costs of improving data format, 7) investment costs of improving data infrastructures, 8) investment costs of improving data processes, 9) training costs of improving data quality know-how and lastly 10) management and administrative costs associated with ensuring data quality. Finally, Eppler and Helfert (2004) argue that data quality costs consist of

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

two major types: improvement costs and costs due to low data quality. Based on this, they devise a simple classification of data quality costs, as shown in Table 3.

3 Proposition

This paper extends the literature on data quality costs, especially the work of Eppler and Helfert (2004), by proposing:

(1) A definition of the optimal data maintenance effort

(2) A classification of costs inflicted by poor quality data

The two propositions are defined and discussed in the following subsections.

3.1 Defining the optimal data maintenance effort

The first proposition of this paper is shown in Figure 1. The vertical axis indicates the incurred, aggregated costs of dealing with poor quality data. The second and horizontal axis deals with the quality of data. The two curves in the figure represent costs inflicted by poor quality data and the costs of maintaining high data quality, respectively. The costs inflicted by poor quality data are for example faulty decisions based on poor data quality, whether this is of operational or strategic character. The costs of ensuring and maintaining high data quality simply refer to the work of assurance or improving data quality. The total costs associated with data quality are the aggregated cost of the two explained curves. There are two basic assumptions associated with Figure 1. Firstly, during data maintenance the focus is on the most critical data (i.e. the ones with the highest payoff per resources spent) before moving on to less critical ones. This implies that the first work of assuring data quality would have the greatest effect, i.e. the costs inflicted by poor quality data decreases exponentially. The second assumption is that the costs of the efforts to ensure high data quality are not causally related to the their importance, i.e. focusing on a set of poor quality data with great impact on costs is not necessarily cheaper than focusing on data with little impact on costs. Thus, the costs of assuring data quality is a linear relationship between data quality and assurance costs.

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

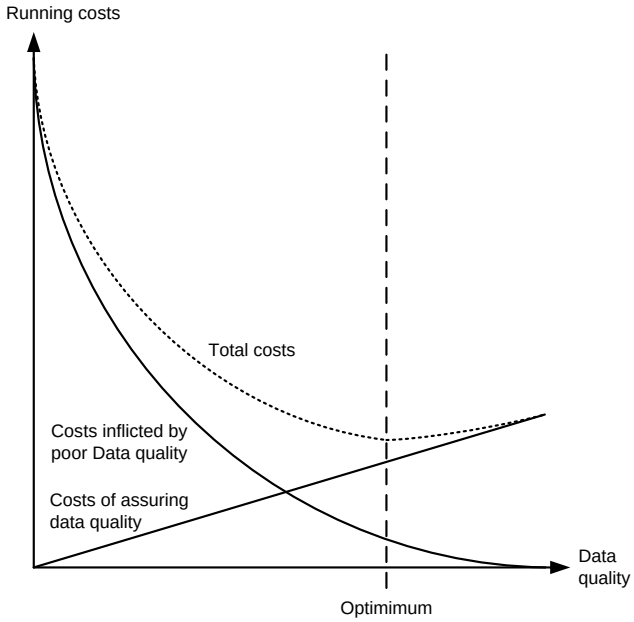

Figure 1. “Total costs incurred by data quality on the company”.

What can be derived from Figure 1 is that the connection between costs inflicted by poor quality data and costs of ensuring high data quality can be logically categorized as a trade-off, which is a situation involving the loss of one quality in return for gaining another quality. The central thesis here is that extensively cleaning data, thereby ensuring high quality of the data, becomes less profitable at some point. This is illustrated by the dotted line termed “total costs”.

Although Figure 1 seems to provide a very logical perspective on the estimation of the optimal data quality maintenance efforts, there is still some way to go. To apply the figure on an area of a company, the two types of costs needs to be evaluated, i.e. the costs of maintaining data and the costs inflicted by poor quality data. The first (costs of assuring data quality) is relatively easy to evaluate, since this is simply a question of registering resources used on this work, i.e. internal hours spent, consultant fees, software, etc. On the other hand, estimating the costs inflicted by poor quality data is much more difficult because of the many indirect and intangible effects associated with it. To support the task of estimating the costs inflicted by poor quality data, the next section looks closer at the nature of such costs.

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

3.2 Estimating costs inflicted by poor quality data

To simplify the estimation of costs inflicted by poor quality data, costs are broken down into two dichotomies. The first dichotomy relates to how visible the costs are, namely direct versus hidden costs. This dichotomy is used in a great deal of management accounting literature (Joshi at al., 2001; Srinidhi, 1992) as well as data quality literature (Kengpol, 2001). Hidden costs are sometimes referred to as expenses that are not normally included in the purchase price of equipment or a machine, such as maintenance, supplies, training and upgrades. For this reason terms such as strategic activity based costing (Kaplan & Cooper, 1998), total cost of ownership (Ellram & Siferd, 1993) and cost-to-serve (Braithwaithe & Samakh, 1998) have been invented and invested in to include all costs associated with a given action taken by a company or department. Although this definition of hidden costs can be claimed to be a valid one, this paper will define hidden costs as costs that the company is incurring but which management is not aware of. An example of such a cost could be the faulty decisions stemming from not knowing the profitability of products. Contrarily, direct costs can be defined as costs that are immediately present and visible to management. This could for example be faulty delivery addresses for registered customers, resulting in wrong deliveries.

The second dichotomy relates to the level on which the costs are inflicted. More specifically, the second dichotomy refers to the fact that data can be viewed on both an operational and a strategic level. On an operational level, data are used as a basis for carrying out tasks and making decisions, which normally have a relatively short time span. An example of operational data can be delivery addresses, pricing of products and other order processing related data. Shipping products to the customer in the right quantity, at the right address and at the right time can be considered as an operation, in which it is paramount that the company can rely on the data being of the right quality. On a more strategic level, data can be seen as a basis for making decisions in companies, where the decisions can be regarded as having a relatively longer time span when compared to operational data. For example, these data can be cost allocations in a company, which is used to determine the pricing of products. If the company is not able to track and locate both its variable and its fixed costs, it will not be possible for the company to determine a given price on a given product. Another example of strategic data could be cost-benefit analyses pertaining to product profitability. If a company

Journal of Industrial Engineering and Management
doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

presently produces three products, it is vital to know which products can be deemed profitable and which can be deemed non-profitable.

Here, it should be made clear that operational and strategic data can be one and the same. That is, in themselves data are not operational or strategic, but data can only be recognized as operational or strategic because management in a given company perceives them that way. As a result, while some data can be seen as strategic in one company, other companies might regard them as operational. It should also be noted that besides the operational and strategic levels, a tactical level also exists in between. This level has not been included, as the purpose of this paper is to provide an initial and better understanding of the relationship between such costs. Future research should investigate what happens when this dichotomy is changed to include three levels.

In Figure 2, the two dichotomies are combined to provide some general categories of costs of poor quality data. The four categories generated by the two dichotomies will be subsequently discussed.

Hidden costs	E.g. long lead times, data being registered multiple times, employee dissatisfaction, etc.	E.g. focus on wrong customer segments, poor overall production planning, poor price policies, etc.
Direct costs	E.g. manufacturing errors, wrong deliveries, payment errors, etc.	E.g. few sales, low efficiency, problems in keeping delivery times, etc.
	Effects of poor quality data on operational tasks	**Effects of poor quality data on strategic decisions**

Figure 2. “Four types of costs incurred by poor quality data”.

In Figure 2, it is highlighted that depending on the two dimensions of direct costs versus hidden costs and operational data versus strategic data, four types of costs incurred by poor data quality can be operationalized. In the figure, examples of each type of these costs are given. When the cost can be classified as a direct cost with an operational view on data, costs can for example be associated with poor order processing data. Shipping the wrong product in the wrong quantity at the wrong time to the wrong customer at the wrong price are examples of mistakes that will eventually incur costs for the company. Another classical example is the

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

direct cost associated with poor production quality, where it is obvious that faulty data produces products that are not assembled properly, for example. Contrarily, when the cost can be categorized as a hidden cost, but still with an operational use of data, the company will incur costs on a day-to-day level of which they are in fact aware. Costs associated with this are for example long lead times. A company that has been producing products with the same lead time for a long period of time runs the risk of taking this for granted, not realizing that the lead times could actually be shorter if the data were corrected. Such data pertain to for example poor data input to Material Requirements Planning (MPR) systems.

When costs are direct but are instead considered from a strategic data perspective, costs incurred stem from operations, which the company knows are inefficient and have a big impact on the strategic direction in which the company is currently heading. An example of this could be the awareness of having lost sales in recent periods due to decision-making based on unreliable data. Not running the newly placed strategic inventory location properly could be an example of costs incurred due to data not being sufficiently cleaned and organized. Lastly, when costs are not visible to management and data are regarded as being strategic, management knows that some data are faulty, but does not realize that this has consequences for the company's overall profit potential. In this case, an example would be a wrong allocation of costs (typically fixed costs) regarding calculating individual product profitability. Not tracking and allocating costs properly would lead to wrong decision-making such as pricing policies and a focus on the wrong customer segments due to products appearing profitable while others appear unprofitable, even though they might in fact be profitable.

3.3 Application of the contributions

To utilize the contributions of this paper, it is important to define a clear delimitation of the data in focus. The focus when using the two proposed models could for example be on item data, sales order data, production planning data, etc. The narrower the scope, the easier it is to estimate costs associated with poor quality data. On the other hand, if using a scope that is too narrow, important data may be neglected. Thus, the use of the proposed models may be the investigation of a series of datasets separately, followed by placing these investigations in a

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

common context. In practice, the focus of such data investigations may even be placed on particular database tables.

4 Case study

In this section, a case study illustrates the way in which a company attempted to improve the quality of their data. The company in question is a manufacturer, developer and supplier of a wide range of automotive spare parts for vans, cars, and trucks with total revenue of around 130.000 Euro per year. The focal company primarily targets the automotive spare parts market, although the company applies these products for a variety of different uses all over the world. In a normal situation, a new car is equipped with parts produced by the same auto manufacturer. For example, a radiator installed in the car is typically the same brand as the car itself. Some parts in cars, vans and trucks are, however, more prone to breaking, compared to others. These are for example the heating and cooling systems of the car. Typical causes for the breakdown of these car parts are normal wear and tear, but also (head-on) collisions with other cars. The original car manufacturers actually produce these spare parts as well, but have not specialized in the cheap production of these. As this is a costly endeavour for the original car manufacturer, an after sales market for car, van and truck spare parts exists. The case company currently employs workers in countries all over the world and has 18 subsidiaries. Before turning to the empirical data, a short section denoting the methodological choices taken is given.

4.1 Methodology

A qualitative and exploratory research design was undertaken in order to investigate the level of master data quality by the focal company (Stake, 2000). The research method consisted of ethnographic observations and semi-structured interviews, because the investigated data are relatively unstructured and analysis of them involves explicit interpretation (Silverman, 2005). Using semi-structured interview protocols gave the interviewer the flexibility to focus on what the company believed was the most important problems as regards their current level of data quality. In terms of data coding, within case analysis was used as a means to structure, reduce and make sense of the data collected (Miles & Hubermann, 1994).The single case study can be reported as being a holistic, representative

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

case design with a single unit of analysis (the case company) (Yin, 2009). The case is representative because the case company is typical of many other major manufacturing companies as the company has had problems in managing their data quality, which is also the main sampling criterion. As this type of case study methodology pertains to a single case, it is only possible to generate an analytical generalization. A statistical generalization is, therefore, not achievable, as this type of research can be regarded as exploratory research. This is a limitation of the paper when seen from a statistical viewpoint.

One researcher spent significant time in the actual focal company, participating for 6 months both at official meetings as an observer and in unofficial, unstructured interviews with the company's chief operating officer (COO), business intelligence managers, supply chain manager and several sales managers. It should be mentioned here that a confidentiality agreement was signed with the company leaving all information anonymized. As one of the researchers participated in the meetings, the researcher runs the risk of blurring his role as a researcher with his role in the company. In order to minimize bias as much as possible, triangulation in the form of a combination of interviews, direct observation, documentation and participant observation was carried out (Yin, 2009). With respect to qualitative validity criteria, credibility was ensured by checking the authenticity of the case description with the case company, after which any discrepancies were changed. Recognizing that two social contexts are never identical, transferability can only be ensured by applying the results to other cases in future research. Even though only one of the authors spent time at the case company, dependability was sought to be ensured by comparing all three authors' interpretations of the results, and working out any disagreements on interpretation. Finally, confirmability can only be ensured through the blind peer-review process.

4.2 Analysis

Because a wide range of cars, vans and trucks exists, many different types of spare parts have to be produced by the focal company. In fact, the company currently has a stock-keeping unit (SKU) count of approximately 8,500. Combined with the many countries to which the focal company is selling products, customers exceed 10,000. This creates a complex situation for the organization, in which data to be managed are abundant with pricing of products being a particularly time-

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

constraining activity. The company is currently employing two full-time business intelligence managers whose sole task is to clean data and price products. Products are priced on a range of factors with a benchmarking towards prices of customers being the main one. As the market for the company's products can be as seen as a commodity market, precise pricing of the app. 8,500 products is particularly important. During recent years, Chinese competitors have entered the market, which has had the consequence that the focal company has been put under pressure in terms of maintaining profitability. The company, therefore, decided to improve and subsequently maintain various data elements in the organization, thereby hopefully ensuring less costs associated with bad data quality.

The two business intelligence managers knew that the company incurred quite heavy costs due to costs inflicted by relatively simple operational tasks. Such tasks pertained to for example shipping products to the right delivery address or bar coding the products with correct ID tags. Additionally, many of the customers of the company had had individual pricing agreements with the company but these agreements were not systematized, which meant that the sales people of the company used a lot of time on retrieving and processing individual and unique customer data. Besides costs that were readily visible, the business intelligence managers also knew that the company was incurring hidden costs associated with operational tasks. For example, both managers would spend a lot of their time recording, retrieving, systematizing and updating pricing information gathered from the company's nearest competitors. These data were important for the company since this allowed them to price their products according to the current market situation. This updating of prices involved, however, many countries with many individual pricing lists being gathered from many different competitors. This often meant that the two managers together with other marketing personnel were carrying out uncoordinated, duplicative work. That is, data were at times registered twice. Considering the quite time-consuming work load for this data storage activity, the company would incur many hidden costs pertaining to this operational task. The COO of the company estimated that the costs of these unnecessary activities were the equivalent of payroll costs for two full-time marketing employees. In an attempt to improve data quality on this operational level, the company attempted together with one of the authors of this paper to develop a pricing model, in which data for pricing products would happen automatically through a computer programme instead of having several employees trying to

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

update prices manually. It was a clear goal for the organization to improve data quality pertaining to pricing to a certain level. That is, intelligence managers, the COO and several marketing related employees expressed that it would never be possible nor expedient to obtain 100% correct prices. Instead, the data quality improvement initiative should be seen as a way to get better, but not perfect, prices. Trying to obtain perfect prices would mean a far too time-consuming data discipline, in which the company was not interested. This aim of data quality improvement goes well with the statement earlier in Figure 2, in which improvement of data quality is only applicable to a certain level. Trying to maintain data quality over a certain threshold will result in costs pertaining to data discipline inexpediently exceeding costs saved by better decision-making due to better data quality.

The company also incurred both direct and hidden costs on a strategic level. Direct costs were mainly associated with supply chain or logistical operations. In all, the company had 18 subsidiaries, each with their own assigned inventory location. Besides these, minor inventory locations were located in the different countries to which the company was supplying. At the time of the empirical investigation, the company had engaged in a long debate concerning the centralization versus decentralization of inventories. These arguments were, however, difficult to reach an agreement upon since cost data pertaining to the use of the inventory locations were either missing or faulty. Due to this the company knew that unnecessary costs occurred when they moved goods to and from different inventories. This meant costs regarding unnecessary transportation of goods, not meeting delivery deadlines and that either stock-outs or limited capacity at inventory settings were incurred by the company. Lastly, the company also incurred costs at a hidden, strategic level. That is, the company essentially had no calculations of customer profitability, but only had rough guidelines such as the volume sold and contribution margin. This meant that the sales staff would spend time on servicing customers with many time-consuming demands and a relatively small profit gain. Not knowing the costs of having products produced, the sales staff were also quite often not capable of determining the optimal price that the customer should pay for the product. The company estimated carefully that such costs contributed with 5-7 % of total fixed costs of the company. In order to improve data quality at this strategic level, the company set out trying to gather information on costs related to inventory capacity and transportation costs. This resulted in a decision to centralize

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

inventories, thereby removing several smaller inventories located especially in Europe. The COO and business intelligence managers and the supply chain department all expressed satisfaction with this decision considering the quality of the data that were used in order to make the final decision. There were, however, also doubts as to whether this decision actually would be the best for the company logistically as data collected sometimes were not sufficiently reliable. This doubt stemmed from the fact that cost data from certain inventory locations were either missing or were obviously wrong. It was, however, judged by the company that it would be a too great a data exercise to gather precise information on all inventory locations. Instead, the costs concerning an adequate level of information versus a not too expensive data collection process were sought balanced.

In the case described, the proposed matrix (Figure 2) provided a perspective on costs of poor data quality, which contributed to a better understanding of this issue. More specifically, the matrix helped dividing data costs into cost types of different concreteness, which helped in evaluating the accuracy of the optimal data assurance effort.

5 Conclusions

This paper proposed a model for determining the optimal level of data maintenance efforts from a cost perspective. More specifically, the optimum is found by adding the costs of data maintenance work to the costs inflicted by poor quality data (such as errors in sales orders, delivery addresses, etc.). As the model shows, the optimal level of data maintenance is not to achieve perfect data, but only a level where the costs of the maintenance work do not exceed savings from the costs inflicted by poor quality data. This data maintenance effort is dependent on the characteristics of the particular company. Different industries have different characteristics, i.e. the relation between costs of poor data quality and costs of assuring data quality. For example, for airplane manufacturers the costs inflicted by poor quality data may be very high compared to the costs of increasing the data quality, while for a manufacturer of simple components the opposite may be the case.

While the first dimension (i.e. costs of data maintenance) is rather straight forward to calculate, the costs inflicted by poor quality data are much more difficult to

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

define. To provide a better understanding of such costs, the paper proposed four categories of such costs. The four categories were created by defining two dimensions: Hidden versus direct costs and operational versus strategic consequences of poor quality data. These four categories provide a better understanding of how to estimate data quality related costs. Examples of effects of poor data quality on operational tasks where costs can be considered as hidden are long lead times and employee dissatisfaction. These types of costs are difficult to track and the company might not notice that it is in fact incurring these costs. When speaking of costs associated with manufacturing errors and wrong deliveries, it was determined that what the company here is dealing with are direct costs due to poor data quality. Contrarily, examples of effects of poor data quality on strategic decisions on costs that are hidden are a focus on wrong customer segments and poor price policies. Finally, direct costs associated with effects of poor data quality on strategic decisions are for instance few sales and problems in meeting delivery deadlines. However, estimates of costs related to poor quality data would still be associated with great uncertainties. But, the more exact estimates, the more the company will profit from such work. This was also empirically illustrated by the use of single case study.

Having defined the optimal effort for data maintenance and having provided some clarification of how to understand the costs inflicted by poor quality data, the next step is to make the model even more operational. This means that more detailed methods for evaluating the different types of costs inflicted by poor quality data need to be defined. The propositions presented in this paper represent the initial ideas of a research project, currently ongoing at the University of Southern Denmark. The focus of this research project is to understand how data quality is related to the expenses of a company. To achieve such insights, the activities to be carried out during 2010 include conducting a number of case studies, which is to end up in a large questionnaire survey. The ideas presented in this paper represent the initial foundation for this work.

To sum up, this paper has produced a better understanding of how to define the optimal data maintenance effort and of the nature of costs inflicted by poor quality data. Although these contributions are to be further elaborated on in future research, in their present form they provide a better understanding of the topic which hopefully aids companies in their data quality work.

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

References

Ballou, D. P., Madnick, S., & Wang, R. (2004). Assuring information quality. *Journal of Management Information Systems*, 20, 9–11.

Ballou, D. P., & Pazer, H. (1985). Modeling data and process quality in multi-input multi-output information systems. *Management Science*, 31(2), 150-162. doi:10.1287/mnsc.31.2.150

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41(3), Article 16.

Braithwaite, A., & Samakh, E. (1998). The cost-to-serve method. *International Journal of Logistics Management*, 9(1), 64-88.

Davenport, T.H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review*, 76(4), 121-131.

Davenport, T.H., & Prusak, L. (1998). Working Knowledge: How Organizations Manage What They Know. *Harvard Business School Press*, Cambridge, MA.

Ellram, L. M., & Siferd, S. P. (1993). Purchasing: The cornerstone of the Total Cost of Ownership concept. *Journal of Business Logistics*, 14(1), 163-184.

Eppler, M., & Helfert, M. (2004). *A classification and analysis of data quality costs*. MIT International Conference on Information Quality, November 5-6, 2004, Boston.

Even, A., & Shankaranarayanan, G. (2009). Utility cost perspectives in data quality management. *Journal of Computer Information Systems*, 50(2), 127-135.

Ge, M., & Helfert, M. (2007). *A Review of Information Quality Research - Develop a Research Agenda*. International Conference on Information Quality, November 9-11, 2007, Cambridge, Massachusetts, USA.

Haug, A., Pedersen, A., & Arlbjørn, J.S. (2009). A classification model of ERP system data quality. *Industrial Management & Data Systems*, 109(8), 1053-1068. doi:10.1108/02635570910991292

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

Häkkinen, L., & Hilmola, O-P. (2008). ERP evaluation during the shakedown phase: Lessons from an after-sales division. *Information Systems Journal*, 18(1), 73-100.

Joshi, S., Krishnan, R., & Lave, L. (2001). Estimating the hidden costs of environmental regulation. *The Accounting Review*, 76(2), 171-198. doi:10.2308/accr.2001.76.2.171

Jing-hua, X., Kang, X., & Xiao-wei, W. (2009). *Factors influencing enterprise to improve data quality in information systems application —An empirical research on 185 enterprises through field study*. 16th International Conference on Management Science & Engineering, September 14-16, 2009, Moscow, Russia.

Kahn, B., Strong, D., & Wang, R. (2003). Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45, 184-192. doi:10.1145/505248.506007

Kaplan, R. S., & Cooper, R. (1998). *Cost and effect: Using integrated cost systems to drive profitability and performance*. Boston: Harvard Business School Press.

Kim, W. (2002). On Three Major Holes in Data Warehousing Today. *Journal of Object Technology*, 1(4), 39-47. doi:10.5381/jot.2002.1.4.c3

Kim, W., & Choi, B. (2003). Towards Quantifying Data Quality Costs. *Journal of Object Technology*, 2(4), 69-76. doi:10.5381/jot.2003.2.4.c6

Kengpol, A. (2001). 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. & R. Katz-Haas (Eds.): *Proceedings of the Sixth MIT Information Quality Conference*, pp. 206-216, Boston.

Knolmayer, G., & Röthlin, M. (2006). Quality of material master data and its effect on the usefulness of distributed ERP systems. *Lecture Notes in Computer Science*, 4231, 362-371. doi:10.1007/11908883_43

Lederman, R., Shanks, G., & Gibbs, M.R. (2003, June). *Meeting privacy obligations: the implications for information systems development*. Proceedings of the 11th

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

European Conference on Information Systems. Paper presented at ECIS: Naples, Italy. Retrieved June 29th, 2009, from: http://is2.lse.ac.uk/asp/aspecis/20030081.pdf

Lee, Y., Pipino, L., Funk, J., & Wang, R. Y. (2006). *Journey to data quality*. Cambridge, Mass: The MIT Press.

Leo, L., Pipino, L. Yang, W. L., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218.

Levitin, A. V., & Redman, T. C. (1998). Data as a resource: Properties, implications, and prescriptions. *Sloan Management Review*, 40(1), 89-101.

Madnick, S., Wang, R., & Xian, X. (2004). The design and implementation of a corporate householding knowledge processor to improve data quality. *Journal of Management Information Systems*, 20(1), 41-49.

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, 12(2), 105–112. doi:10.1057/palgrave.dbm.3240247

Miles, M.B., & Huberman, M.A. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, California, CA.

Newell, S., Robertson, M., Scarbrough, H., & Swan, J. (2002). *Managing Knowledge Work*. Basingstoke: Palgrave-Macmillan.

Park, K., & Kusiak, A. (2005). Enterprise resource planning (ERP) operations support system for maintaining process integration. *International Journal of Production Research*, 43(19), 3959-3982. doi:10.1080/00207540500140799

Piprani, B., & Ernst, D. (2008). A Model for Data Quality Assessment. *Lecture Notes in Computer Science*, 5333, 750-759. doi:10.1007/978-3-540-88875-8_99

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

Raman, A. (2000). Retail-data quality: evidence, causes, costs, and fixes. *Technology in Society*, 22, 97–109.
doi:10.1016/S0160-791X(99)00037-8

Redman, T.C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82.
doi:10.1145/269012.269025

Ryu, K.-S., Park J.-S., & Park, J.-H. (2006). A data quality management maturity model. *ETRI Journal*, 28(2), 191-204.
doi:10.4218/etrij.06.0105.0026

Silverman, D. (2005). *Doing qualitative research*. London: Sage Publications.

Smith, H. A., & McKeen, J. D. (2008). Master data management: Salvation or snake oil? Export find similar. *Communications of the Association for Information Systems*, 23(4), 63-72.

Srinidhi, B. (1992). The hidden costs of specialty products. *Journal of Management Accounting Research*, 4, 198-208.

Stake, R.E. (2000). Case studies, in Denzin, N.K. and Lincoln, Y.S. (Eds.), *The handbook of qualitative research* (pp. 435-454). California: Sage Publications.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, 41(2), 54-57.
doi:10.1145/269012.269021

Vayghan, J. A., Garfinkle, S. M., Walenta, C., Healy, D.C., & Valentin, Z. (2007). The internal information transformation of IBM. *IBM Systems Journal*, 46(4), 669-684.
doi:10.1147/sj.464.0669

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.
doi:10.1145/240455.240479

Wang, R. Y., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-34.

doi:10.3926/jiem.2011.v4n2.p168-193 JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953
Print ISSN: 2013-8423

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4), 623–640.
doi:10.1109/69.404034

Watts, S. G., & Shankaranarayanan, A. E. (2009). Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48, 202–211.
doi:10.1016/j.dss.2009.07.012

Yin, R.K. (2009). *Case Study Research: Design and Methods*. Los Angeles, LA: Sage Publications.