BIDQI: The Business Impacts of Data Quality Interdependencies model

Business impact	*Data quality characteristic:*	Accessibility	Conciseness	Timeliness	Completeness	*Ranking*
Lost revenue		3	2	7	6	4.50
Operational inefficiencies		5		4	7	4.00
Lost business opportunities		5		6	2	3.25
Poor decision making		1		3	3	1.75
…						

Marco Spruit
Vincent van der Linden

Technical Report UU-CS-2019-001
February 2019

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

BIDQI: The Business Impacts of Data Quality Interdependencies model

Marco Spruit & Vincent van der Linden

Department of Information and Computing Sciences, Utrecht University, The Netherlands
corresponding author: m.r.spruit@uu.nl

As the data volumes within enterprises grow, the number of errors in stored data and the organizational impact of these errors is likely to increase. CIOs and business executives must be able to justify the expense of the initiative and convey the value proposition effectively to senior management. In order to do this, data quality needs to be expressed terms of costs and organizational consequences, to be able to convey the value of improving data quality correctly. By creating the Business Impacts of Data Quality Interdependencies (BIDQI) model in which data quality characteristics are linked to business impacts arising from data quality issues, this research aims to provide a high-level method to discover the consequences and costs of poor data quality within organizations. The model will be able to assist researchers and practitioners in determining the actual costs of a data quality problem within an organization by giving them a tool to identify partially hidden costs which are caused by poor data quality. The constructs of the model are based on an extensive literature review and expert interviews were conducted to establish the interdependencies.

1. Introducing the impacts of poor data quality in organisations

Data volumes within enterprises grow at a very high pace and enterprises are increasingly dependent on the timely availability of high quality data. In fact, many organisations' basis for competition has changed from tangible products to intangible information. Relatively new concepts such as cloud computing have steadily increased the importance of data quality and information security management (*e.g.* Baars & Spruit, 2012). The current trend of Big Data poses significant additional data quality issues on top of that, all the way from technical to information governance challenges as outlined in Alves de Freitas *et al.* (2013), among others.

Poor quality information can have significant social and business impacts (Strong et al., 1997) and there is strong evidence that data quality problems are becoming increasingly prevalent in practice (Redman, 1998; Wang and Wang, 1996). Most organisations have experienced negative effects of decisions based on information of inferior quality (Huang et al., 1999). Information quality issues have become important for organisations that want to perform well and obtain competitive advantage. Especially when enterprises need to reorganise their IT function, they often face many data quality challenges—both syntactic and semantic—which is testified by the emergence of master data management as a separate research domain, as overviewed in Spruit & Pietzka (in press). A comprehensive overview of current data quality research topics in both research and practice is documented in Sadiq (2013).

The DataWarehousing Institute (TDWI) estimates that poor quality customer data alone cost U.S. businesses over $600 billion a year. However, these data quality issues are often either not seen or ignored by most executives. According to TDWI's Data Quality Survey (Eckerson, 2002), almost half of all companies have no plan for managing data quality. The survey drew responses from 647 individuals in a range of positions, industries, and countries.

In the Netherlands the lack of quality in relational data alone causes €400 million extra costs every year as shown by a survey held among 20,000 Dutch organizations employing ten or more people, by Nyenrode Business University. A press release by Human Inference, a developer of data quality solutions and one of the initiators of the survey, states that the total amount of €400 million consists of costs that are calculated based on directly quantifiable aspects, such as wrongly addressed invoices and product deliveries which do not arrive at the right addresses. If indirect and 'opportunity' costs made as a result of badly maintained prospect databases were counted in "the figure would most likely become much higher" (HumanInference, 2006).

As the data volumes within enterprises grow, the number of errors in stored data and the organizational impact of these errors is likely to increase (Klein, 2002). More and more organizations do believe that quality information is critical to their success (Wang, 1998). However, not many of them have turned this belief into effective action. Enterprises seem reluctant to address, solve and prevent data quality issues until it is too late, making it a reactive process. The reason for this seems to be twofold: Management either accepts the status quo of their data environment as normal and acceptable, or they are unaware of the actual costs of poor quality data (English, 1999). Identifying the costs of poor data quality currently is indeed a cumbersome task. Creating a business case for fixing an organization's data environment has proven to be difficult.

Part of that difficulty is that data quality efforts are competing with other initiatives for IT budget dollars and staffing. CIOs and business executives must be able to justify the expense of the initiative and convey the value proposition effectively to senior management. In order to do this, data quality needs to be expressed terms of costs and organizational consequences, to be able to convey the value of improving data quality correctly. There is currently no clear view on how data quality affects an organization as a whole, which makes expressing the added value of data quality improvement initiatives such a hard task.

2. Research objectives and methodology

Quantifying data quality improvement is a way to convince companies that steps should be taken to improve data quality throughout their business. In order to quantify data quality improvements, a thorough understanding of data quality itself is needed. This research further clarifies the term data quality by investigating its characteristics and its impact on organizations.

This work presents the Business Impacts of Data Quality Interdependencies (BIDQI) model in which data quality characteristics are linked to business impacts arising from data quality issues. , this research aims to provide a high-level method to discover the consequences and costs of poor data quality within organizations. The model will be able to assist researchers and practitioners in determining the actual costs of a data quality problem within an organization by giving them a tool to identify partially hidden costs which are caused by poor data quality.

The research questions (RQ) that were formulated to achieve the creation of this model were as follows: RQ1: How can data quality be defined and quantified, and which characteristics of data quality are the most relevant in the context of this research? RQ2: What are the business impacts of poor data quality and how can these be categorized? RQ3: What data quality characteristics are the most relevant characteristics per business impact? RQ4: What are the interdependencies between data quality characteristics and their business impacts, and how can these interdependencies be used to identify hidden costs? The following two sections present the results of the literature research on data quality definitions.

3. Finding appropriate data quality characteristics

Before investigating the impacts of data quality, the term itself must first be defined. The coming sections elaborate on defining data, information, knowledge, wisdom and quality of information and data and explain the relations between these concepts. Figure 1 shows the relation of the terms data, information, knowledge and wisdom. The first mention of this hierarchy can be found in (Ackoff, 1989).

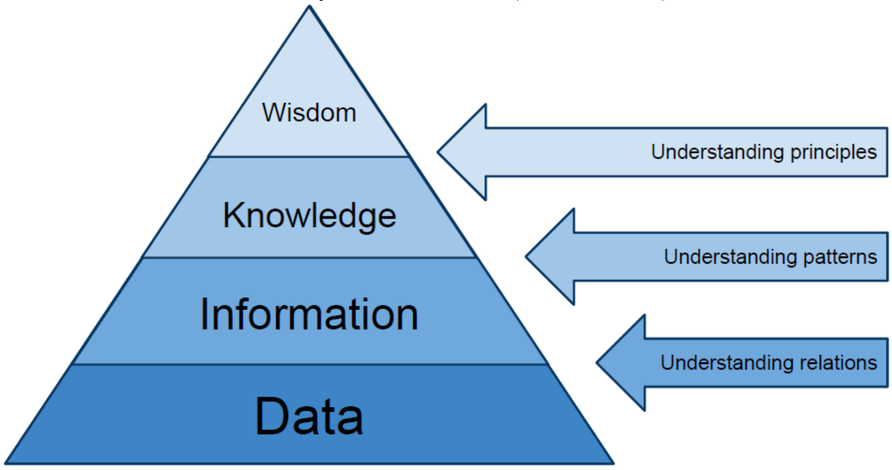

Fig. 1. Commonly used Knowledge Pyramid, first described in Ackoff (1989).

Data: Data refers to an elementary description of things, events, activities, and transactions that are recorded, classified, and stored, but not organized to convey any specific meaning (Turban et al., 1997). Hence, data has no real meaning out of context and it requires an association with something else (Jashapara, 2004). Data can be seen as the representation of a real world entity, a fact. Without a descriptive definition data has no meaning or true value.

Information: Information is data in context, in other words: usable data. Information can be considered as systematically organized data, which has meaning and value to the recipient. For example, the number 0031302533708 is correct data, a representation of a fact. It can not be considered information before we know the meaning or context of this value. When we have defined the context of the value, in this case the context of the value is 'telephone number', the data is transformed into information.

Data and information are often used interchangeably (Wang, 1998). In practice the distinction between these two terms can be made intuitively by users. When investigating the quality of data, context and meaning will prove to be invaluable. The terms data and information with respect to quality will be used synonymously in this work, as is general practice in the field of data quality research.

Knowledge: Knowledge can be described as information in context and implies understanding the significance of information. In knowledge management the definition of knowledge has its roots in the ideas of logical behaviorism based on work by Gilbert Ryle and Michael Polanyi (Ryle, 1949; Polanyi, 1967). They distinct the terms tacit and explicit knowledge; tacit knowledge meaning knowing how (intelligence) and explicit knowledge meaning knowing that (possessing knowledge). Polanyi (1967) gives the example of riding a bike. Staying upright and engaged in the activity of riding is tacit knowledge (knowing how). Articulating what exactly keeps a person upright (knowing that), is part of explicit knowledge.

In the context of this research, knowledge should be interpreted as the value added to information by people who have the experience to use it to its full potential.

Wisdom: Wisdom is described as applied knowledge. Ackoff (1989) asserts that it is the difference between efficiency and effectiveness that differentiates Wisdom from the lower levels in the understanding hierarchy. The lower levels data, information and knowledge contribute to efficiency, while wisdom is required to assure effectiveness. Data, information, and knowledge have value in that they facilitate the process of pursuing goals, objectives and desired outcomes. Wisdom is used to choose the right things to pursue and thus the effectiveness of the choice takes into account the value of the outcome (Ackoff, 1996). Wisdom is viewed as reapplied knowledge, which is tested and reconfigured through lessons learned in the past. The terms knowledge and wisdom will not be used in this research. The focus lies on data and information quality. Data and information constitute the foundation of the knowledge pyramid, and their quality will define and support the higher concepts of knowledge and wisdom.

4. Quality of data and information

Data quality (DQ) can be considered as the quality of data values, or in other words the accuracy of those values. This has long been the view on data quality in practice (Levitin and Redman, 1995). An investigation of data quality literature reveals many other characteristics of data quality (or information quality) than the mere accuracy of data values.

Definitions of quality found in literature and practice can, in general, be described as coming from either product-based or service-based perspectives. The product-based approach, commonly called data quality, focuses on the design and internal information systems view, and defines quality as the degree data satisfies initial specified requirements or the degree to which the data corresponds with real-world entities and facts. Typical criteria to measure the quality include completeness and accuracy of data. The issue with this approach is that there can still be deficiencies with respect to the initial specification of requirements of the data and the actual use of the data. This in turn has lead to a service-based approach to quality, commonly called information quality, which focuses on the information consumer and the consumer's use of the data. Using the term information instead of data implies that the delivery and use of data must be considered when one judges quality.

Information quality has been defined differently by several authors, but examining these definitions reveals a consensus about what information quality is. Huang et al. (1999) define information quality as information that is fit for use by information consumers. Kahn et al. (2002) define information quality as the characteristic of information to meet or exceed customer expectations, and as information that meets 'specifications' or 'requirements'. Other authors also describe information quality as information that is most useful to the information customer. Fitness of use seems to be the most appropriate way to describe information quality and coincides with Juran's widely accepted definition of quality (Juran et al., 1974).

In order to evaluate information quality, many researchers have formulated key characteristics, often described as dimensions of information quality. These dimensions can be used to make the information quality concept more concrete and measurable. Several studies have confirmed information quality is a multi-dimensional concept (Ballou and Pazer, 1985; Redman, 1996; Wang and Strong, 1996; Wang, 1998; Ballou et al., 1998; Huang et al., 1999). A review of information quality literature reveals a multitude of frameworks which were created in order to investigate information quality within information systems. The most norable of these frameworks is the framework created by Wang and Strong. They formulated fourteen information quality dimensions and grouped them within four information quality categories (intrinsic, contextual, representational, accessible). A graphical representation of their model is shown in figure 2. Wang and Strong (1996)'s use of dimensions has been adopted by many other researchers, who have refined or finetuned the model to their own research context.

Therefore, numerous different dimensions have been created and different models have been constructed. To get a complete view of information quality definitions and definitions of information quality dimensions, eleven other frameworks were inspected to find the most prevalent dimensions of information quality. These dimensions will be used to construct the data quality interdepency model. The dimensions which will be used in the framework will be elaborated on in section 5.

Table 1 shows the results of the literature research on information quality dimensions. The author, name of the model and dimensions used are summarized. A selection of the frameworks which were investigated are described.

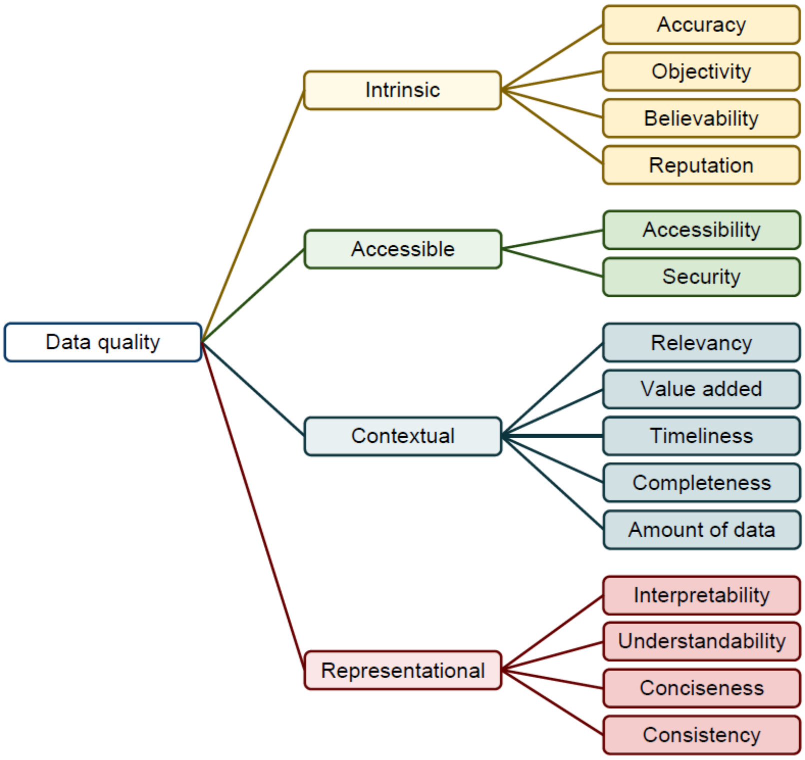

Fig. 2. Data quality dimensional overview, based on (Wang and Strong, 1996).

Zeist and Hendriks (1996) identified the information quality characteristics categories functionality, reliability, efficiency, usability, maintainability and portability. The category functionality includes the characteristics suitability, accuracy, interoperability, compliance, security and traceability. Reliability covers the characteristics maturity, recoverability, availability, degradability and fault tolerance. The category efficiency contains the time and resource behaviour. Usability includes the understandability, learnability, operability, luxury, clarity, helpfulness, explicitness, customisability and user-friendliness characteristics of information. Maintainability pertains to the characteristics analysability, changeability, stability, testability, manageability and the reusability. Finally, the category portability contains the characteristics adaptability, conformance, replace-ability and installability.

Alexander and Tate (1999) suggest a quality framework for the web and it includes criteria such as authority, accuracy, objectivity, currency, orientation and navigation.

Shanks and Corbitt (1999) described a semiotic-based framework for the quality of data and it consists of four semiotic levels. Syntactic information quality covers the characteristic consistency. Semantic information quality includes the characteristics accuracy and completeness. The information must be comprehensive, unambiguous, meaningful and correct. Pragmatics information quality include the characteristics usability and usefulness. Furthermore, they list the characteristcs timeliness, conciseness, accessibility and reputation.

Information quality criteria as mentioned by authors Naumann and Rolker (2000) include subject, object and process criteria. Subject criteria cover believability, concise representation and understability of information, interpretability and relevancy of information and added value. Objective criteria include completeness, security, objectiveness, timeliness and verifiability. Process criteria ensure that information should be accurate, have proper linkage to other information, be available, and concise.

Table 1. Data quality dimensions in existing frameworks.

#	Author(s)	Data quality model	Components
01	Wang and Strong (1996)	A Conceptual Framework for Data quality	4 categories 16 dimensions
02	Zeist and Hendriks (1996)	Extended ISO Model	6 quality characteristics 32 sub-characteristics
03	Alexander and Tate (1999)	Applying a Quality Framework to Web Environment	6 criteria
04	Katerattanakul and Siau (1999)	IQ of Individual Web Site	4 categories
05	Shanks and Corbitt (1999)	Semiotic-based Framework for Data Quality	4 semiotic descriptions 4 goals of information quality 11 dimensions
06	Dedeke (2000)	Conceptual Framework for measuring IS Quality	5 quality categories 28 dimensions
07	Naumann and Rolker (2000)	Classification of Information Metadata Criteria	3 assessment classes 22 information quality criteria
08	Zhu and Gauch (2000)	Quality metrics for information retrieval on the WWW	6 quality metrics
09	Leung (2001)	Adapted ISO Model for Intranets	6 characteristics 28 dimensions
10	Eppler and Muenzenmayer (2002)	Conceptual Framework for IQ in the website Context	2 ‘manifestations’ 4 quality categories 16 quality dimensions
11	Klein (2002)	<none>	5 information quality dimensions
12	Kahn, Strong and Wang (2002)	Mapping IQ Dimensions into the PSP/IQ model	2 quality types 4 information quality classifications 16 information quality dimensions

It is apparent that there are similarities between the different frameworks, and that there are some characteristics that have been renamed by certain researchers, but may cover the same subject as previously defined characteristics. The characteristics mentioned in previous research were collected and compared in order to find the characteristics most important to this research. Even though more recent researches (*e.g.* Poeppelmann & Schultewolter, 2012) have been conducted to uncover and structure the spectrum of data quality dimensions—or perhaps because of that—we conclude here that the selection of standard works in Table 2 can be considered complete, and that current research now should focus on how to map these data quality dimensions to help answer business needs.

5. Selecting data quality characteristics

The twelve frameworks were compared, and the most common characteristics were abstracted. Table 2 shows which data quality characteristics are present in the frameworks.

The characteristics were selected due to the fact at least half of the investigated frameworks noted these characteristics as an important part of data quality. The author of this thesis feels believability and reputation should be perceived as results of data quality, not characteristics of data quality; these characteristics are therefore omitted from the new model. The rightmost column in Table 2 shows the number of times a data quality characteristic was mentioned.

Table 2. Data quality characteristics within the twelve frameworks under investigation.

Characteristic	01	02	03	04	05	06	07	08	09	10	11	12	Occurrences
Accuracy	x	x	x	x	x	x	x	x	x	x	x		11
Timeliness	x	x	x		x	x	x	x	x	x	x	x	11
Accessibility	x	x	x	x	x	x	x	x	x	x		x	11
Relevancy	x	x	x	x		x	x	x	x		x		9
Completeness	x			x	x	x	x			x	x	x	8
Objectivity	x		x		x		x	x		x	x		7
Understandability	x	x				x	x		x	x		x	7
Conciseness	x		x	x			x			x		x	6
Consistency	x		x		x	x				x		x	6
Security	x	x					x		x	x			5

The descriptions of the characteristics used in the frameworks were used to judge if a characteristic was equal to a characteristic as described by Wang and Strong (1996), and included in the numbering. This selection method has led to a revised data quality model in which the most relevant characteristics to this research are noted, depicted in figure 3.

A brief description of the characteristics can be seen in table 3. These definitions are taken from Wang and Strong (1996)'s article on their multidimensional model for data quality. A more thorough description of the characteristics can be found in the following paragraphs. With the most relevant data quality characteristics selected, the next step is to investigate the business impacts of poor data quality.

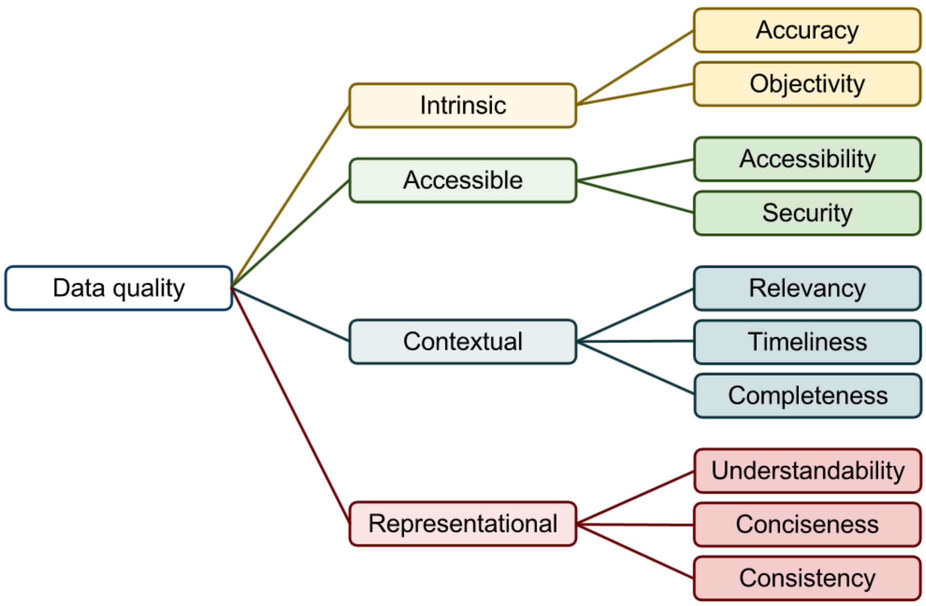

Fig. 3. Selected data quality dimensions in our BIDQI model.

Table 3. Data quality characteristics overview.

Characteristic	Description
Accuracy	Extent to which data are correct, reliable and certified free of error
Timeliness	Extent to which the data are sufficiently up-to-date for the task at hand
Accessibility	Extent to which data are available, or easily and quickly retrievable
Relevancy	Extent to which data are applicable and helpful for the task at hand
Completeness	Extent to which data are not missing and are of sufficient breadth and depth for the task at hand
Objectivity	Extent to which data are unbiased, unprejudiced and impartial
Understandability	Extent to which data are clear without ambiguity and easily comprehended
Conciseness	Extent to which data are compactly represented without being overwhelming
Consistency	Extent to which data are presented in the same format and compatible with previous data
Security	Extent to which access to data are restricted appropriately to maintain their security

6. Business impacts of poor data quality

The goal of the BIDQI model is to give a high-level overview of data quality and its impacts on organizations. Constructing a list of impacts which arise from poor data quality is therefore essential for the creation of the model. The aim was to create a complete list of cost-related impacts. To discover the full array of impacts occurring due to data quality problems, we identified a multitude of cases in which data quality was an important driver or root cause of a larger problem. These cases were collected from a large consulting firm's Knowledge Exchange -a large database which holds a repository of past projects -, as well as from case descriptions from several data quality tool vendors, and results from a survey done by The Data Warehousing Institute.

The data set used comprised of thirteen business cases in which data quality was an important driver. Several industries were examined, namely: three banks, two telecom companies, two software manufacturers, two manufacturing companies, two insurance companies and two IT companies. The TDWI survey and case descriptions from data quality tool vendors were used to gain insight into which impacts were generally identified as data quality related impacts.

By comparing the impacts regularly mentioned in literature and the survey to the business cases within the knowledge repository of the consultancy firm, eleven distinct impacts were identified. Table 4 shows in which industry a certain impact occured.

The eleven impacts that have been derived were shown to two experts in order to validate the completeness of this impact range, and they concurred it was complete. Employee morale and system credibility were not identified in the cases of Accenture, but will be included in the list of business impacts. These impacts are less likely to be identified by business, and if they are identified, they are often not communicated outside of the company. However, they could be of interest to this study; an assumption was therefore made that these impacts were relevant enough to include in the list.

Table 4. Business impact occurrences in thirteen data quality-driven business cases.

#	Business impacts per sector (number of cases)	Bank (3)	Telecom (2)	Software (2)	Manufac-turing (2)	Insurance (2)	IT (2)
01	Lost sales opportunities		x	x			x
02	Customer service costs	x	x	x			x
03	Customer dissatisfaction	x	x	x			x
04	Lost revenue	x	x				
05	Operational deficiencies	x	x	x	x	x	
06	Delays in system/project deployment		x			x	
07	Regulatory compliance	x					
08	Poor decision making	x	x	x	x		x
09	Lost business opportunities				x		
10	Employee morale						
11	System credibility						

7. The costs of poor data quality

This section explains the business impacts arising from poor data quality and provides examples of cost calculations for each impact.

7.1 Lost sales opportunities

This impact concerns lost sales opportunities due to data quality issues. This includes failing to cross-sell products, inability to keep up with trends, and inability to properly identify customer needs. For instance, the lack of accurate profile data may cause possible new customers omitted from a marketing campaign, leading to missed customer lifetime value as explained in the section on customer dissatisfaction. Incorrect market trend or economic data may lead to the failure of marketing campaigns. If poor data quality contributed to these campaigns failing, the cost of the failed campaigns and the corresponding missed customer value can be calculated and attributed to poor data quality.

7.2 Customer service costs

Increased customer service costs are the costs made by having employees correct data either on a daily basis or periodically. These costs may vary, but having employees correct data is time consuming, and therefore creates a cost that can be measured in those employees' wages. During the interviews it was also mentioned that the employees that had to correct data were often not hired for this specific task, but were burdened with the task because it simply needed to be done. This will naturally have an impact on employee morale as well. Both the employees working on correcting the data in the databases (IT) and employees who are working on finding the correct values for the data are considered the basis of the costs of this impact. If customer service has to call customers in order to get the correct data, like an address, the time spent on this task is also included in this impact.

Calculating the costs of this business impact is quite straightforward:

$$\text{FTE(employeesCorrectingDataQuality)} * \text{wages}$$

To calculate the costs of this impact, one should categorize the number of employees working on scrap and rework because of data quality issues. The amount spent on repairing faulty data can be expressed as the FTEs(full time equivalent) of the employees working on repairing faulty data, times their wages.

However, if a company uses software to analyze its data quality, the costs of using this software should be added to the total costs of this business impact:

$$\text{FTE(employeesCorrectingDataQuality)} * \text{wages} + \text{softwareCosts}$$

Additional time spent on work caused by data quality issues will be a recurring cost calculation method.

7.3 Customer dissatisfaction

Costs that arise due to customer dissatisfaction can be hard to measure. Losing a dissatisfied customer will lead to a direct cost: the customer lifetime value of that customer, which will be explained below. But there are more, harder to identify consequences. The dissatisfied customer might advice friends, colleagues, and family not to do business with the company they had a bad experience with. In terms of data quality, customer satisfaction can have several causes. Errors in names, addresses, billings, or product information can lead to a dissatisfied customer. Miscommunication due to data quality errors is another example of a cause for customer dissatisfaction.

The most viable way to calculate the costs associated with this business impact is by calculating the customer lifetime value. Customer lifetime value indicates the net profit or loss to the firm from a customer over the entire life of transactions of that customer with the organization (Jain and Singh, 2002). Detailed customer lifetime value calculation is outside of the scope of this research, but an brief overview of the method follows: One can calculate the expected lifetime value of a customer by extracting sales data from a sample of customers to calculate their average revenue and profit. The lifetime of a customer depends on the organization performing the calculation. Retail customers are generally considered to have a five year lifetime, whereas banks often calculate a longer lifetime value, since the customer is more likely to stay with the same bank for a longer period of time. The costs of acquiring a customer, through marketing or otherwise, should be subtracted from the expected profits during his lifetime.

Once the customer lifetime value of a segment of customers has been calculated, this value can be used to calculate the costs of customer dissatisfaction due to data quality. After identifying sources of customer complaint data, like the organization's customer service department, help desk, or any other point of customer interaction, the customer complaints that have a data quality component should be identified. This entails the complaints regarding one of the aforementioned data quality errors, such as misspelled names, billing errors and incorect product data.

An estimate of customers who are dissatisfied due to data quality issues is now known, and the attrition rate of these customers should be determined. The attrition rate of dissatisfied customers is likely already known by the marketing department, and by multiplying the number of data quality related complaints by this attrition rate, the number of lost customers can be estimated, and should be multiplied by their customer lifetime value to calculate the cost of poor data quality to an organization.

Calculating the costs of customer attrition due to poor data quality is in fact a very powerful way to illustrate the impact of data quality on an organization, since it provides a solid value of lost revenue.

7.4 Lost revenue

In this impact overview, this impact consists of the costs made because poor data quality lead to errors in your invoices, or even an inability to properly bill your customers.

According to a survey of small and medium-sized enterprises by credit insurance company Atradius done in 2005, 67% of the companies regularly receive invoices which contain errors. 64% of the firms send the invoice back and demand a correct one, and more than half said incorrect invoices delay payment by at least a week; 15 % said the average delay was more than a month (Bray, 2005)). This delay in time and the time spent on handling the incorrect invoices can be translated into the costs caused by poor data quality. The costs of underbilling customers could be included in the calculation of the total cost of this impact, but this would be debatable since it is likely that overbilling occurs as well, canceling the underbiling issue.

However, some industries are affected more by this impact than others. Insurance companies deserve a special mention here, since they have a higher risk to encounter significant costs due to poor data quality causing billing errors.

In Katz-Haas and Lee (2005), the researchers looked at information quality in a managed care organization and found a very significant financial impact of poor data quality.

Their information-quality manager had found data in the data warehouse showing 40,00060,000 members per month as active when in fact their policies had been canceled. Further analysis showed that the company was paying approximately $4,000,000 annually in claims for members with canceled policies. Due to out-of-date information, this organization suffered a considerable loss, indicating the potential risk in this area, especially for organization such as this one.

7.5 Operational inefficiencies

Operational Inefficiencies cover costs that arise due to inefficient processes caused by poor data quality. This includes poor resource planning, inability to react in time to external developments, increased system workloads, and costs that arise due to incorrect or duplicate mailings to customers. In the case of duplicate mailings, the cost of this impact is simply the cost of sending obsolete mailings. But this impact also includes the costs of having to find the root cause of a data error and the time and money invested in remediating the problem caused by a data defect. When knowledge workers have to stop working to find missing information, they are spending time on a data quality related incident. This time can be used to calculate the costs attributed with this impact, by multiplying it by their wages. It should also be mentioned that when a a piece of faulty data is encountered, it can ‘break’ a larger process. The data might have to be corrected while the process is halted, wasting more time and money. By reducing the delays associated with detecting and correcting data errors, and the rework associated with that correction, more transactions can be processed, resulting in greater volume processing and lower cost per transaction.

Examples are the costs of having incorrect product price data listed on a website, which causes a significant loss. Having incorrect address data, which causes products to be sent to the wrong customer causes the cost of resending the product to the correct customer, as well as the cost in time wasted on finding the correct data for said customer. In manufacturing companies, a data error can be the cause of a defect product, which can not be sold or might have to be recalled, causing considerable costs. Costs caused by data quality related operational inefficiencies differ significantly per industry, and there is no standard way of calculating these costs. However, when a process fails due to a data quality related issue, the cost of this failure can be attributed to poor data quality and used to justify data quality improvement initiatives.

7.6 Delays in system/project deployment

This category consists of the costs made due to the delay or cancellation of new projects or systems, because of poor data quality. If many projects (such as business intelligence and data warehousing projects) fail as a result of poor data quality, then conversely, improving the quality of information should ensure that a project cannot be delayed or canceled as a result of bad data. This is not a predictable cost, a project either succeeds or fails; but having poor data quality hinders new projects and systems. Data quality improvement will help create a smoother environment for new projects, since it lessens the need to correct old data when converting to a new system.

7.7 Regulatory compliance

Regulatory Compliance costs are costs made due to being unable to comply to regulatory compliance. In case of banks, this can be a very significant impact. Since this adhering to regulatory compliance is essential in this industry, poor data quality can create a very large cost. In defining this impact, the ability to fulfill service level agreements is also included in order to make the impact more relevant to businesses which have less regulatory compliances to worry about. Failure to maintain proper environmental data for instance, can put an organization at risk for liability; inaccurate financial data or improper use of information can even put the entire organization at risk. Calculating the costs of this impact is not really viable. This impact represents a risk of creating situations in which an organization can not function at all. If poor data quality leads to this impact, the costs will be represented by penalties which depend on the magnitude of the failure to comply to regulations, and will be different for every organization.

7.8 Poor decision making

These are the costs made due to the inability to make correct long term decisions caused by poor data quality. Improper forecasts made due to incorrect or out-of-date data are a prime example of this impact. Having more compliant customer information allows the business intelligence process to provide more accurate customer profiling, which in turn can lead to increased sales, better customer service, and increased valued customer retention.

Most organizations understand the impact of data quality on analysis and decision support. The proliferation of business intelligence (BI), with data drawn from disparate systems and applications, can degrade data quality, lowering users confidence in BI reports. However, BI deployed with quality data can help an organization compete more effectively and decisively. Improving data quality creates the opportunity to make quicker and more correct decisions. Calculating costs related to this impact is not possible, but it is clear that better data quality will facilitate better decisions.

7.9 Lost business opportunities

Loosely related to lost customer opportunities, lost business opportunities cover the costs made due to missing business-to-business opportunities caused by the inability to correctly analyze internal and external data. Poor data quality can lead to a poor view of the businesses around you, leading to missed opportunities such as better procurement. If the data on the market around a company is incomplete, there is a risk that they are missing opportunities such as being able to buy resources cheaper. Having complete and up-to-date data enables companies to effectively analyse purchasing data, which in turn will provide a means of determining which suppliers offer the most value and will make it possible to identify and monitor saving opportunities. The value this creates consists of the savings created through better informed procurement, and will differ per organization.

7.10 Employee morale

These are the costs made due to dissatisfied personnel, dissatisfied because they have to correct data, and can not service customers properly. These employees might become less productive due to having to correct data while it is not their primary task, or become increasingly frustrated by being unable to conduct their daily activities properly due to poor data quality. If poor data quality causes lower employee morale, the cost of the loss of productivity of an employee can be contributed to poor data quality. If an employee is 10% less productive because of lowered morale, the cost will be 10% of that employee's wage. However, if the lowered employee morale leads to increased employee attrition, one can make a case that the costs associated with hiring and training new employees are partly caused by poor data quality. Interestingly, Mueller & Coppoolse (2013) found that incentive systems can be used to increase information quality in business intelligence systems as a way to enhance employee morale.

7.11 System credibility

This final impact covers the costs that are caused by low trust in data. An example of this is multiple departments within a company all keeping track of their own data, often in excel sheets which can not be accessed by others; this because they do not trust the data from another department or the general data of the company. Sometimes there are even multiple databases which contain the same data, but with different values. Besides the costs of running these separate systems, having multiple versions of the truth can lead to an array of other impacts, as described above. But to keep this impact categorization mutually exclusive, this impact will only cover the costs of running multiple systems and the costs of the inefficiency caused by having multiple data repositories. The costs of this impact can be calculated by estimating the time and money spent by employees on collecting data to include in a private database and the cost of running this database next to the now obsolete older databases.

8. Discovering the interdependencies between data quality and business impacts

The goal of this research is to find the relation between the identified data quality characteristics and the business impacts of poor data quality. This chapter will show the method used to identify these relations, and present the results of the study.

8.1 Method

Using the theoretic basis established in the previous chapter, expert interviews were conducted. For this study, ten experts in the field of data quality were interviewed using a semi structured interview, combined with a task. Table 5 shows that the ten experts have worked on data quality projects in several different industries, including telecom companies, banks, insurers, and education.

Table 5. Interviewed expert profiles.

#	Expert profile	Industry
01	2 years software development data matching software. 5 years of experience in reporting, KPIs, data warehousing, data modelling, cleansing and proling.	Consultancy - multiple
02	5+ years experience in data management and architecture	Consultancy – multiple
03	Data migration tooling, 20+ years of experience in the field	IT – multiple
04	Business intelligence, 15+ years of experience in data quality	Consultancy – multiple
05	Data warehousing and business intelligence	Telecom
06	Leading Dutch expert on Customer data management	Education, research, marketing
07	10+ years of experience with data analysis	Insurance
08	8 years of experience with ETL and data architecture	Financial
09	Data quality software development, 10+ years	IT – multiple
10	Over 5 years of experience with business intelligence, master data management	Consultancy - multiple

During the interview, the experts were asked about their experience with data quality projects and they were inquired about their knowledge of the theories behind data quality. This part usually lasted about fifteen minutes and established an insight about the experts' knowledge and experience on the subject of data quality. This unstructured start of the interview also gives the interviewer and participant the chance to establish a good relationship, and it gives the expert the opportunity to describe the domain in ways he is familiar with (Schreiber et al., 2000).

The second part of the interview consisted of a task, based on a technique called 'card sorting'. Card sorting is a user-centered design method used in web-design for increasing a systems findability. The process involves sorting a series of cards, each labeled with a piece of content or functionality, into groups that make sense to users or participants.

According to Rosenfeld and Morville (1998), card sorting can provide insight into users' mental models, illuminating the way that they often tacitly group, sort and label tasks and content within their own heads. Card sorting is described as a reliable, inexpensive method for finding patterns in how users would expect to find content or functionality. However, this technique is mostly used to discover what the best architecture is of a web site: where should menu items be and how should they relate.

This technique was used as a foundation for a task, which aimed to reveal three things:

1. Which impacts are the most costly to an organization?
2. What are the links between the data quality characteristics and the business impacts?
3. Are the chosen characteristics, impacts and the business impact model valid?

The materials to accomplish this task consisted of ten cards with data quality characteristics, eleven cards with business impacts, and an answer sheet (see appendices). The terms on the cards were explained to the expert, so there would be no misinterpretation of the concepts on the cards.

To discover the first item, which impacts are the most costly to an organization, the experts were asked to sort the cards with the impacts on them from one to eleven. Placing all the cards on the table, the participant had a good overview of all the impacts at all time. The result of this part of the task was a ranked list of all the impacts.

After this, they were asked to group the impacts into several categories, without telling the participant what kind of categories. The goal of this question was to indirectly investigate the usefulness and validity of the grouping that has been made according to the three strategic vectors in the previous chapter.

During the last part of the task, the experts were handed a random business impact, and asked to pick three data quality characteristics which they thought were the most relevant to that impact. This was done for all the impacts to discover the links between the characteristics and the business impacts.

By using this relatively simple variation of the card sorting technique, combined with the unstructured part of the interview, it was possible to investigate the three subjects of interest in this study in a relatively short time. The interviews lasted one hour on average, but provided a lot of information and data suitable for this research.

8.2 Results of ranking the business impacts

The following sections will present the results of this task. First, the experts were asked to rank the business impacts in terms of costs. The goal was find out which impacts were viewed as being the most costly to an organization. The final results of this ranking task are shown in Table 2.

Table 6. The ranked business impacts in terms of costs, and their weighted impact score.

#	Business impact	Rank:	1	2	3	4	5	6	7	8	9	10	11	Score
01	Lost sales opportunities		3	3		1			2	1				7.62
02	Operational inefficiencies		2			3	2	3						7.15
03	Customer dissatisfaction		2	2			1	1	1	1	1	1		5.79
04	Lost business opportunities				3	2		2		2			1	5.69
05	Increased customer service costs		1				6		1	2				5.67
06	Lost revenue		1	2				2	2	2			1	5.08
07	Employee morale		2	1	1				2				4	4.57
08	System credibility				2	2		2				3	1	4.44
09	Poor decision making			1		2			3	1		3		3.90
10	Regulatory compliance				2					3	4		1	2.86
11	Delays in project development				1					3	5	1		2.09
		Weight multiplier:	1	1	0.9	0.8	0.7	0.5	0.4	0.2	0.1	0	0	

Table 6 shows the number of times an impact was ranked on a specific rank. In the case of the impact lost sales opportunities, three experts ranked it as having the most impact on an organization in terms of cost (hence the value 3 in the top left cell), three as the second most important impact, one experts ranked it as being the forth most important impact, et cetera.

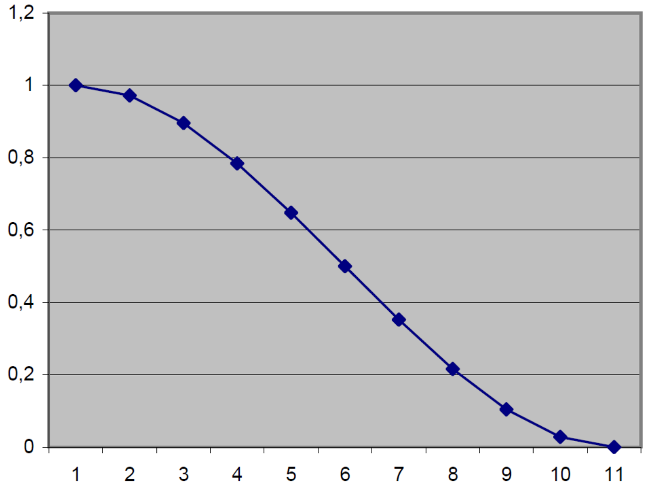

Fig. 4. The blending curve of our ranking mechanism.

Sorting the impacts based on most and least costly to an organization was perceived as a diffcult task, mostly because some of the impacts were seen as being equally important. This occurred mostly while ranking the top three and bottom three impacts. This lead to the decision to use a ranking curve, as shown in Figure 4, to analyze the results of this task. By using a blending curve, the differences between ranks 1-3 and

9-11 were slightly adjusted in a way to reduce their difference, to reflect the comments of the experts. This curve yields multipliers for each rank, as shown in Figure 4, which were used to calculate the final ranking of the business impacts. These combined weights led to a final ranking of business impacts of poor data quality, as shown in the bottom row of Table 6 labeled *Weight multiplier*.

Lost sales opportunities and operational ineffciencies were perceived as being the most costly impacts of having poor data within one's organization, while impacts as regulatory compliance and delays in system and project development were seen as having little impact in terms of costs.

9. BIDQI: The Business Impacts of Data Quality Interdependencies model

This section presents the Business Impacts of Data Quality Interdependencies (BIDQI) model which helps identify the most relevant characteristics per business impact, and shows the interdependencies between data quality characteristics and business impacts. It shows the important characteristics per impact, such as accuracy, timeliness and relevancy in the case of lost sales opportunities. More importantly, it shows which areas of a business are affected by having problems with a certain data quality characteristic.

Table 7. The Business Impacts of Data Quality Interdependencies (BIDQI) Model relates data quality characteristics with business impacts.

Business impact / *Data quality characteristic:*	Accuracy	Accessibility	Objectivity	Security	Conciseness	Consistency	Understandability	Timeliness	Relevancy	Completeness	Accuracy
Lost sales opportunities	6	3	1		2		2	7	9		6
Increased customer service costs	6	2	2		1	4	3	6	2	4	6
Customer dissatisfaction	8	3	3			1	4	8	1	2	8
Lost revenue	6	3			2	1	2	7	3	6	6
Operational inefficiencies	8	5	1			3		4	2	7	8
Delays in project development	1	3		2	7	6	5	1	2	3	1
Regulatory compliance	6	1	1	5	2	4	1	6	1	3	6
Poor decision making	5	1	3	1		1	5	3	8	3	5
Lost business opportunities	4	5	2				4	6	7	2	4
Employee morale	7	2			1	2	6	8	1	3	7
System credibility	3	4		3	2	5	4	4	4	1	3

Table 7 displays the results of the task in which the experts were asked to identify the three most important data quality characteristics per business impact. The amount of shading of the cells indicate to what extent a link was found between a certain characteristic and a business impact. Therefore, white cells annotate that none of the experts linked that characteristic to the business impact. The results of a validation of the BIDQI model are presented in the next section.

10. Evaluation of the BIDQI model

We evaluated the BIDQI model in practice through a single case study at a Dutch industrial company. The market in which the company operates is currently getting increasingly dynamic, raising the complexity and importance for portfolio and risk management. However, the organization's IT landscape could not fully support these developments. They were addressing this issue by creating a new IT landscape that would support the company in successfully operating in a market that demands an ever increasing flexibility.

10.1 Evaluation method

Although the project under investigation was not identified as a data quality project by the organization itself, data quality was being addressed thoroughly during this project, and recognized as being important to the succes of the project. Therefore, their IT landscape renewal project was selected as being a proper candidate to test the interdependency model.

Yin (1994) states that a single case study can be an appropriate research approach if the organization under investigation can be considered a critical case (that is, one which meets all the conditions needed to test the theory). The company was actively addressing data quality issues while renewing their IT landscape, indicating a relationship between the drivers of the renewal project and data quality. By comparing both the status of the business impacts from the interdependency model and the status of data quality in the organization, we set out to evaluate the applicability of the model and its predictive qualities.

Testing the interdependency model consisted of three steps. The separate parts of this test analyze the current state of the organization in terms of business impacts and the current state of data quality within the organization, serving as a canvas on which the model can be projected, and subsequently validated.

First, the applicability and state of the business impacts described in the model were related to the organization. By combining document analysis with a structured interview with a project consultant regarded as having extensive knowledge of the project drivers, the business impacts directly leading to the IT renewal project were discovered.

Second, two experts, an information analyst and the technical architecture manager of the company were asked to indicate the data quality characteristics they identified as being of poorest quality at that time. They were presented with the same cards representing the data quality characteristics used in the interviews leading to the interdependency model, randomly spread out on a table. The goal of this step was to identify the data quality characteristics that were perceived as less than optimal in this organization. Additionally, they would validate the findings of step 1.

Third, using these data quality characteristics, the business impacts that had a high inclination of occurring according to the interdependency model were identified, and the experts were inquired about these impacts. They were asked to indicate whether the business impact(s) predicted as likely to occur were applicable to their organization. The input gathered from this exercise was used to interpret the results of the evaluation described below.

10.2 Identification of business impacts

From the document analysis and interview with the business impact expert, the following business impacts were important within the company and had a role in the business case of renewing the IT landscape: Operational Inefficiencies, Poor Decision Making, Lost Business Oportunities, and Lost Revenue.

Regarding *operational inefficiencies*, due to its highly competitive market the organization needs the ability to react to external influences as fast as possible. This includes external influences such as market fluctuations, competitor actions, and resource availability. Furthermore, resource tracking and allocation are essential due to the large number of customer transactions. *Poor decision making* is likely to have a big impact due to the number and complexity of external influences, but forecasting and (long term) resource planning is considered challenging because of fragmented data collection, data integration and data mining endeavors. *Lost business opportunities* loom as market competition increases and procurement contracts increasingly have shorter durations. This leads to a increased need of high quality market data, supplier data and contract data. Finally,the high number of transactions implicate a high importance of accurate and timely billing of customers. A relatively small discrepency in price information, volume information, time information or interest rates, will lead to a high accumulated *loss in revenue*. The next step in the evaluation was to discover the state of data quality in the organization.

10.3 Identification of data quality characteristics

The experts were asked to pick four data quality characteristics out of the ten shown to them and to comment on why they chose these data quality characteristics. This number of data quality characteristics does not completely reflect the methods used in building the model (choose three characteristics per business impacts); an extra characteristic was added to improve the completeness of the validation. The four data quality characteristics identified by the experts as being of less than optimal quality were: Conciseness, Accessibility, Timeliness, and Completeness.

Conciseness: conciseness was identified as having room for improvement. The challenge this characteristic describes is the large volume of data being processed and presented in a manageable manner. As the data within this company was highly fragmented, it posed a challenge integrating this data into a manageable and concise report. *Accessibility:* not surprisingly, as this seems to be an omnipresent practice within many organizations, many data were stored in Microsoft Excel files, locally stored on the desktop computers of individuals. This has several consequences (multiple versions of the same data or 'truth', a data analysis tool used as data source), but the one focused on by the experts was limited accesibility of data. *Timeliness:* as explained by the experts, the large volumes on data and the processing time to create reports lead to the data not always being delivered in a timely manner. *Completeness:* finally, completeness was added, explaining that the distribution (or better: fragmentation) of the data lead to incomplete data. They also confirmed and validated the results of step one of the validation: discovering the business impacts most relevant to this organization.

10.4 Interpreting the findings

We are now ready to interpret the results. Table 8 shows the identified business impacts and the identified data quality characteristics of our case study.

The first thing that comes to attention is the low association of conciseness with the identified business impacts. The experts named this data quality characteristic as a very important one in their organization, yet the interdependency model does not reveal a linkage between it and the business impacts derived during the evaluation research. An explanation to this can be found in the process of building the model. Conciseness, at least to Dutch speakers, is not a widely known term.

Table 8. The BIDQI subset of identified business impacts ranked by the identified data quality interdependencies in the case study.

Business impact / *Data quality characteristic:*	Accessibility	Conciseness	Timeliness	Completeness	*Ranking*
Lost revenue	3	2	7	6	4.50
Operational inefficiencies	5		4	7	4.00
Lost business opportunities	5		6	2	3.25
Poor decision making	1		3	3	1.75
…					

During the interviews that lead to the model, it was often needed to point out what conciseness meant more than once (even though the card used had the explanation on it). This could have let to the experts overlooking or not considering the card often. It ranks low among almost all the business impacts, with the exception of 'delays in system/project development', which in turn was an impact that was ranked low in importance. Furthermore, the business impact poor decision making seems to have a low correlation to the mentioned data quality characteristics. This will be discussed in the next and final section.

If we calculate the average scores of the combination of data quality characteristics identified by the experts as sub optimal, it results in the ranked model shown in the rightmost column of Table 8. As can be seen, this model shows three of the four identified business impacts in the top four most likely business impacts to occur when there is lack of quality in the four identified data quality characteristics. This confirms the descriptive qualities of the interdependency model, at least for this organization.

Regarding the predicting capabilities of the BIDQI model in Table 7, we can also investigate, for example, customer dissatisfaction with its ranking score of 3.25 on this subset of characteristics as a likely business impact when these four data quality characteristics are of poor quality.

In conclusion, it was shown that the BIDQI model has a high descriptive value and reflected a real world case to a satisfying degree. The relationships between data quality characteristics and business impacts as described in the BIDQI model appear valid for the case study organization. However, due to time constraints we could not yet perform a more elaborate evaluation of the BIDQI model as presented in this paper.

11. Conclusions

A thourough literature study has refined the definition of data quality using a combination of established data quality models. This definition was validated during the research by twelve experts in total, who deemed it as complete and logical. Case analysis has produced eleven business impacts which were validated by the same experts as a complete set of impacts leading from poor data quality. This alone has produced an added value to the field of data quality.

This research has also developed the Business Impacts of Data Quality Interdependency (BIDQI) model in which the interdependencies between data characteristics and their business impacts were succesfully reflected in the case of a single case study at a large organization. The BIDQI model adds value by examining the impacts of data quality in a structured and novel way, by deconstructing data quality and its impacts into smaller components and uncovering the relations among the two concepts. As an analytical tool, these interdependencies can be used to identify likely business impacts related to specific data quality characteristics. These likely business impacts, in turn, have costs related to them, as illustrated through this paper.

References

Ackoff, R. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16:3-9.

Ackoff, R. (1996). On learning and the systems that facilitate it. Center for Quality of Management Journal, 5(2):27-35.

Alexander, J. and Tate, M. (1999). Web Wisdom: How to evaluate and create information quality on the web. Mahwah, NJ.

Alves de Freitas, P., Andrade dos Reis, E., Senra Michel, W., Gronovicz, M., De Macedo Rodrigues, M. (2013). *Information Governance, Big Data and Data Quality.* Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference, pp.1142-1143.

Ballou, D. and Pazer, H. (1985). Modeling data and process quality in multi-input, multi-output information systems. Management Science 31, 2:150 - 162.

Ballou, D.,Wang, R., Pazer, H., and Tayi, G. (1998). Modeling information manufacturing systems to determine information product quality. Management Science 44, 4:462-484.

Bray, P. (2005). Retrieved February 12 2009: http://www.telegraph.co.uk/nance/2955224/Survey-reveals-the-high-cost-of-bad-invoicing.html.

Dedeke, A. (2000). A conceptual framework for developing quality measures for information systems. In Proceedings if 5th International Conference on Information Quality, pages 126-128.

Eckerson, W. W. (2002). Data warehousing special report: Data quality and the bottom line. Achieving business success through a commitment to high quality data.

English, L. P. (1999). Improving Data Warehousing and Business Information Quality. John Wiley & Sons, Inc.

Eppler, M. and Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In Proceedings of 7th International Conference on Information Quality, pages 187-196.

Huang, K., Lee, Y., and Wang, R. (1999). Quality Information and Knowledge. Upper saddle river, NJ: Prentice hall.

HumanInference (2006). Press release on 'de kosten van gebrekkige kwaliteit van relatiegegevens'. Retrieved January 21, 2009 from the World Wide Web http://www.humaninference.com/company/press/262.

Jain, D. and Singh, S. (2002). Customer lifetime value research in marketing: A review and future directions. Journal of Interactive Marketing, 16(2):34-46.

Jashapara, A. (2004). Knowledge Management. An Integrated Approach. Prentice Hall.

Juran, J., Gryna, F., and Bingham, R. (1974). Quality Control Handbook. McGraw-Hill Book Co, New York, NY.

Kahn, K., Strong, D., and Wang, R. (2002). Information quality benchmarks: Product and service performance. Commun. ACM, 45:184-193.

Katerattanakul, P. and Siau, K. (1999). Measuring tinformation quality of web sites: Development of an instrument. In Proceedings of the 20th international conference on Information Systems, pages 279-285, Charlotte, North Carolina, United States.

Katz-Haas, R. and Lee, Y. W. (2005). Understanding interdependencies between information and organizational processes. In Wang, R., Pierce, E., Madnick, S., and Fisher, C., editors, Information quality, pages 167-178. M.E.Sharpe, New York.

Klein, B. (2002). When do users detect information quality probelems on the world wide web? In American Conference on Information Systems, page 1101.

Leung, H. (2001). Quality metrics for intranet applications. Information and Management, 38(3):137-152.

Levitin, A. and Redman, T. (1995). Quality dimension of a conceptual view. Information Processing & Management, 31:81-88.

Mueller, R., & Coppoolse, D. (2013). *Using Incentive Systems to Increase Information Quality in Business Intelligence: A Quasi-Experiment in the Financial Services Industry*. System Sciences (HICSS), 2013 46th Hawaii International Conference, pp.3827-3836.

Naumann, F. and Rolker, C. (2000). Assessment methods for information quality criteria. In Proceedings of 5th Internation Conference on Information Quality, pages 148-162.

Poeppelmann, D., & Schultewolter, C. (2012). Towards a Data Quality Framework for Decision Support in a Multidimensional Context. *International Journal of Business Intelligence Research 3*(1), 17-29.

Polanyi, M. (1967). The Tacit Dimension. Doubleday, New York.

Redman, T. (1996). Data Quality for the Information Age. Artech House: Boston, MA.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Commun. ACM, 41(2):79-82.

Rosenfeld, L. and Morville, P. (1998). Information Architecture for the World Wide Web. O'Reilly, Sebastopol, CA.

Ryle, G. (1949). The concept of Mind. Hutcheson, London.

Sadiq, S. (2013). Handbook of Data Quality – Research and Practice. Springer.

Schreiber, G., Akkermans, H., Anjewierden, A., De Hoog, R., Shadbolt, N., Van de Velde, W., and Wielinga, B. (2000). Knowledge Engineering and Management, The CommonKADS Methodology. MIT Press, Cambridge, Massachusetts.

Shanks, G. and Corbitt, B. (1999). Understanding data quality: Social and cultural aspects. In Proceedings of the 10th Australasian Conference on Information Systems, page 785.

Spruit,M., & Pietzka,K. (In press). MD3M: The Master Data Management Maturity Model. *Computers in Human Behavior*.

Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997). Data quality in context. Communications of the ACM, 40(5):103-110.

Turban, E., McLean, E., and Wetherbe, J. (1997). Information Technology for Management. John Wiley & Sons, New York.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Commun. ACM, 39(11):86-95.

Wang, R. Y. (1998). A product perspective on total data quality management. Commun. ACM, 41(2):58-65.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst., 12(4):5-33.

Zeist, R. and Hendriks, P. (1996). Specifying software quality with the extended iso model. Software Quality Management IV - Improving Quality, BSC, pages 145-160.

Zhu, X. and Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In Proceedings of the 23rd annual international ACM SIGIT conference on Research and development in information retrieval, pages 288-295, Athens, Greece.