

2 Literature Review

In this chapter, definitions of data and information quality as well as decision support systems and the decision-making process will be presented. In addition, there will be an overview of research related to each of these topics. In particular, there are various factors that can influence one's decision-making efficiency. One main assumption of this study is that accuracy of information or how data is presented has a major impact on the time it takes to make a decision as well as decision-making performance. At the end of the chapter, a summary of factors that can have an effect on decision-making efficiency will be listed. These factors were extracted from existing literature.

2.1 Data and Information Quality

Understanding a user's decision-making processes is imperative for the data analyst to understand, since data quality is dependent on the business need. For example, one data consumer might rate data quality as very low because there is no sufficient amount of data available to make a decision. Another data consumer rates data quality as high, even though no sufficient data is available. In this case, other data quality dimensions might be important for this data consumer. Wang and Strong's Quality Framework, which comprises 16 different dimensions of data quality, clustered into four categories, demonstrates this (Fisher et al., 2011: 41-43), as outlined in the next four sections.

2.1.1 Intrinsic Data Quality

Fisher et al. (2011: 42-45) found a strong correlation between accuracy, believability, objectivity, and reputation of data. *"The high correlation indicates that the data consumers consider these four dimensions to be intrinsic in nature"*. The quality of the data is intrinsic when the quality of the data is directly knowable of the data. Batini & Scannapieco (2006: 20-21) emphasize that there are two kinds of data accuracy. One, syntactic accuracy considers the closeness of a value to a definition domain. In other words, a value v will be compared to a set of values D . If D contains v , then v is syntactically correct. For example, one might compare $v = \text{Jack}$ with $v' = \text{John}$. Value v (Jack) would then be syntactically correct, even if $v' = \text{John}$, because Jack is a valid name in a list of persons' names. Two, there is semantic accuracy. This type of accuracy looks at how close a value v is to its true value v' . Semantic accuracy applies when there are relationships between sets of data. For example, one might consider a database with records about movies. For each movie title there is a director listed. If Peter Jackson was listed for The Lord of the Rings, then Peter Jackson would be considered semantically

correct. If Peter Jackson was replaced by Quentin Tarantino, then Quentin Tarantino would be semantically incorrect. In both cases, the name of the director would be syntactically correct, since both of them exist in the domain of valid directors.

Wang & Strong (1996) noted that companies are focusing too much on accuracy as the only data quality dimension. The authors suggest considering a much broader conceptualization of data quality. In regards to believability of data, Fisher et al. (2011: 44) talk about multiple factors determining this dimension of data quality. One's knowledge, experience, and the degree of uncertainty in related data are known to be the influencing elements on believability. Furthermore, the authors suggest that believability might be much more important than accuracy because people are driven by their beliefs. The degree of judgment used in the data building process negatively correlates with how people perceive data to be objective. Finally, reputation of data might prevent people from considering how accurate data is. Reputation of data is built over time, and as Wang & Strong (1996) noted, both data and data sources can build reputation.

2.1.2 Contextual Data Quality

This category includes relevancy, completeness, value-added, timeliness, and amount of data (Fisher et al., 2011: 45). Wang & Strong (1996) brought up that the value-added dimension of data quality can be understood as data that adds value to a company's operations and, thus, gives the organization a competitive edge. Timeliness refers to how old data is. This is a very important attribute of data in manufacturing environments, as Fisher et al. (2011: 45) point out. Furthermore, some data are affected by age, whereas other data are not. As an example, the authors refer to George Washington, who was the first president of the United States. This information is unaffected by age. Incorrect decisions are often the result of financial decisions that are based on old data.

The quantity of information is a serious issue in evaluating data quality. A study on the use of graphs to aid decisions and a phenomenon called information overload was once conducted by Chan (2001). The scholar assumed that processing too much information can lead to making poor decisions. An experiment was conducted to show whether business managers would perform differently when treated with different loads of data. One group of subjects was given information with high load, whereas the other group of subjects was given information with nominal load. The results demonstrated that business managers under nominal information load could make higher quality decisions than those under high information load. This demonstrates that having more information is not necessarily better, or, in other words, does not necessarily lead to higher decision-making performance. The phenomenon of information overload could be proven in this study.

2.1.3 Representational Data Quality

This category reflects the importance of the presentation of data. It consists of the dimensions interpretability, ease of understanding, representational consistency, conciseness of representation, and manipulability. The category is “*based on the direct usability of data*” (Fisher et al., 2011: 47). Wang & Strong (1996) describe representational consistency as data that is continuously presented in the same format, consistently represented and formatted, as well as compatible with data that was presented previously. The scholars list clarity and readability as synonyms for the understandability of data. Attributes comprising the dimension of consistency are as follows: aesthetically pleasing, well-formatted, well-organized, and represented compactly. Fisher et al. (2011: 47) emphasize that there is a fine line between having troubles excerpting the essential point of an expression that is too long and having problems remembering what an acronym or short expression stands for when shortening long expressions. This could lead to errors in decision-making and, thus, it is suggested that data analysts work with users in determining the ideal version of data presentation. In addition, different users should be involved at different times.

2.1.4 Accessibility Data Quality

This category of data quality consists of the dimensions access and security. Questions to consider in this category are how and if data is available, and how well data is secured against unauthorized access. “*Accessibility and security are inversely related dimensions*”. As an example, time-consuming security features (e.g. login) that are added to restrict data access make it more difficult for users to get access to information they need for making decisions and, thus, lowers perceptions of data quality. Probably, increasing security decreases accessibility (Fisher et al., 2011: 47-48). Wang & Strong (1996) list the following attributes in connection to data access security: proprietary nature of data as well as inability of competitors to access data due to its restrictiveness.

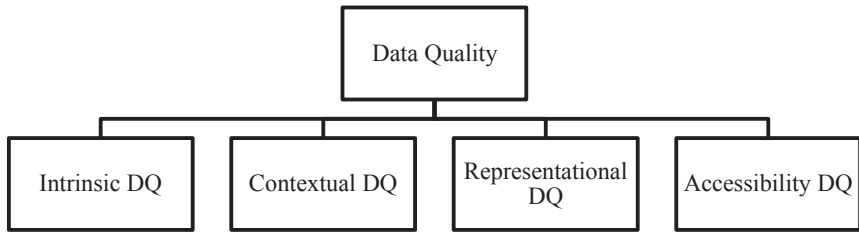


Figure 2: Data Quality Hierarchy – The four categories
Adapted from Fisher et al. (2011: 43)

Fisher et al. (2011: 44-48) illustrate some examples of different data quality dimensions. For instance, high data quality in terms of being accurate means that if there is an inventory database showing that 79 parts are in stock, then there should also be exactly the same amount of items in the stockroom. Another example is about the objectivity of data quality. The author states that *“users measure the quality of their data based on the degree of objectivity versus the degree of judgment used in creating it”*. Timeliness is referred to as how data is out-of-date. A strategic planner may perceive a data record as timely even if it is years old. The strategic planner might base their decisions on old information whereas a production manager might only value data that is within the hour. According to Sedera & Gable (2004), enterprise systems success is dependent upon attributes within the dimensions of system quality, information quality, individual impact, and organizational impact. In comparison to Wang & Strong’s Quality Framework, which was illustrated before, Sedera & Gable present the following attributes for information quality: Availability, usability, understandability, relevance, format, and conciseness. Moreover, system accuracy is mentioned to belong to the category system quality. Decision effectiveness, learning, awareness and recall, as well as individual productivity are classified into individual impact.

The Canadian Institution for Health Care Information (2009: online) follows a data quality framework which consists of five different dimensions:

- Accuracy: Does information from a data holding coincide with real information?
- Timeliness: Is data still current when it is released?
- Comparability: Are all data holdings collecting data in a similar manner?
- Usability: Can data be easily accessed and understood by its users?
- Relevance: How does data meet a user’s current potential future need?

The dimensions of the framework outlined before are part of an approach to “*systematically assess, document and improve data quality*” for all data holdings of the Canadian Institute of Health Care Information (2009: online). In this Master’s thesis, Wang & Strong’s (1996) data quality framework will be followed, since most research efforts have been undertaken into this direction.

Eppler & Muenzenmayer (2002) came up with a conceptual framework for information quality in the website context. They generally distinguish between content quality and media quality. For content quality, they further distinguish between relevant information and sound information. Attributes that can be associated with relevant information are as follows: comprehensive, accurate, clear, and applicable. Concise, consistent, correct, and current are attributes that make information sound. Media quality can be divided into the categories optimized process, with attributes like convenient, timely, traceable, and interactive, as well as reliable infrastructure, with attributes like accessible, secure, maintainable, and fast. Difficult navigation paths on a website are deemed an example of the convenience attribute.

2.2 Research Areas of Data and Information Quality

Batini & Scannapieco (2006: 16-17) talk about research areas that are being discussed in relation to data quality:

- Statistics: Making predictions and formulating decisions in different sets of contexts even if there is inaccurate data available is possible due to the development of a wide variety of methods and models in this field. Statistical methods help to measure and improve data quality.
- Knowledge representation: Rules and logical formulas are needed as the basis of a language that helps to represent knowledge. For improving data quality, reasoning about knowledge and the provision of a “*rich representation of the application domain*” are becoming more important.
- Data mining: This is the analytic process to find relationships among large sets of data. Exploratory data mining, which is defined “*as the preliminary process of discovering structure in a set of data using statistical summaries, visualization, and other means*”, can be used to improve data quality as well (Dasu & Johnson, 2003: 23).
- Management information systems: This research area is probably the most relevant for this Master’s thesis. Data and knowledge in operational and decision business processes are resources that are gaining in value and importance.

- Data integration: Distributed, cooperative, and peer-to-peer information systems own heterogeneous data sources that need to be integrated so that a unified view of data can be provisioned.

Research studies that have been done in the fields mentioned above will be introduced in the following text.

2.2.1 Impact of Data Quality on Organizational Performance

Madnick et al. (2009) note that there are technical and nontechnical issues that may cause data and information quality problems:

“Organizations have increasingly invested in technology to collect, store, and process vast quantities of data. Even so, they often find themselves stymied in their efforts to translate this data into meaningful insights that they can use to improve business processes, make smart decisions, and create strategic advantages. Issues surrounding the quality of data and information that cause these difficulties range in nature from the technical (e.g., integration of data from disparate sources) to the nontechnical (e.g., lack of a cohesive strategy across an organization ensuring the right stakeholders have the right information in the right format at the right place and time).”

A literature review about previous research in data quality reveals that these technical and nontechnical issues have been frequently focused on by various scholars. Research of data and information quality is wide-reaching and affects many areas in the industry, as Tee et al. (2007) show in their article that can be found in the Accounting and Finance Journal. The scholars examined factors that influence the level of data quality in an organization. Senior managers as well as general users were sampled through interviews and surveys in a target organization. One key insight that is very relevant to this Master’s thesis is that the perceptions of the relative importance of data quality dimensions were measured among and between senior managers and general users in this company. It turned out that there were no differences between the two groups across the three dimensions tested – accuracy, relevance, and timeliness. Accuracy was rated almost twice as important as the other two dimensions. In addition, management commitment and the presence of a champion for data quality both had a positive influence on the levels of data quality achieved in the target organization.

IBM’s white paper talks about solving data quality issues through improved data quality management. Retail and manufacturing businesses constantly expand their channels for reaching customers. With increasing global economic complexity, maintaining high levels of data quality becomes a problem. In the production industry, it is important to maintain high levels of data quality as a means of reducing waste in the supply chain (IBM, 2010).

An intensive literature review about impacts of factors on the success of information systems was done by Petter, DeLone & McLean (2008). The scholars point out that there are six major dimensions that are known to have an influence on the successful usage of information systems: system quality, information quality, service quality, use, user satisfaction, and net benefits. In comparison, many of the attributes within these six dimensions are very similar to the attributes that can be found in Wang & Strong's (1996) framework of data quality dimensions. For example, understandability and user friendliness of a system are two attributes of system quality. These might be closely related to ease of understanding as well as interpretability of data in Wang & Strong's framework.

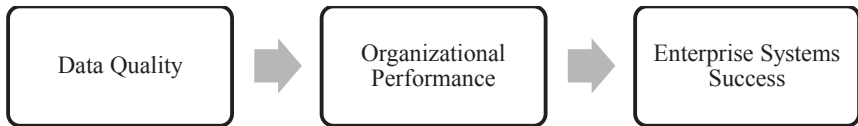


Figure 3: Dependencies between data quality, organizational performance, and enterprise systems success

The figure above demonstrates the importance of data quality. Considering previous research on dependencies between data quality dimensions, information systems success, and organizational performance, a big picture can be drawn. Sedera & Gable (2004) argued that overall productivity of an organization has an impact on the success of enterprise systems, whereas Fisher et al. (2011: 4) summarize that data quality in organizations has an influence on productivity.

In a distributed project setting, the quality of aggregate project-status data that needs to be sent between organizations can be a major problem in a lot of companies. Managers must make sourcing decisions in distributed software projects based on data that are flawed, even though the data are inaccurate. Most of the erroneous data is mainly caused by *“exchanging data across multiple systems, combining data from multiple sources, and from legacy applications”*. Furthermore, collecting project-status data in a timely manner is often an issue. The importance of timeliness as a data quality dimension which is correlating with accuracy suggests that project-status data should be updated in a real-time manner. Performance is also negatively influenced by inaccurate data. One approach for improving data quality in distributed project settings is to apply a Kalman-Bucy filter to present more accurate data to managers who need to make sourcing decisions (Joglekar, Anderson & Shankaranarayanan, 2013).

Xu et al. (2002) report improved information quality as one of the benefits of implementing ERP systems, whereas Cao & Zhu (2013) view it from a different perspective and talk about data quality problems in ERP-enabled manufacturing.

Changes in the Bill of Materials (BOM) require adjustments in calculating materials required, and in generating product, purchase, as well as work orders. The scholars found that adjustments of these data were especially difficult if the Bill of Materials had to be changed frequently. Inaccurate data in processes such as production planning, logistics, or manufacturing would be the result of these frequent changes. It was also found that are two characteristics of ERP systems that are very hard to extinguish, but they can cause data quality problems: 1) Complex interactions of components, and 2) tight coupling due to the implantation of an ERP system.

Furthermore, inconsistencies and inaccuracies in data sets can pollute a data source. This might cause difficulties in performing data analysis. For transactional systems, it means that orders taken incorrectly, or errors occurring in packaging, documentation, or billing, can cause dissatisfied customers, or can result in additional material and labor costs. In a case study that involved the implementation of an ERP system, it was found that a cross-departmental increase in ERP system usage had increased overall data accuracy in the company (Vosburg & Kumar, 2001).

In general, data inaccuracies in sets of data seem to be a crucial issue in companies. Moreover, it seems as if manufacturing firms and organizations that are utilizing ERP systems need to be very aware of data quality issues. In the next subsections, light will be shed on data quality issues in the health care industry, as well as options for assessing data quality. Knowledge about how to measure data quality will be needed in this study so that possible impacts on decision-making efficiency can be determined.

2.2.2 Data Quality Issues in Health Care

According to McNaul et al. (2012), assisted living technologies use artificial intelligence and automated reasoning to understand the behavior of people who need care due to chronic diseases, and people who need health and social care provision due to their age. Inherently, ambient intelligence-based systems, or Ambient Assisted Living (AAL) technologies make it possible for people to extend the time they live at home by providing feedback to users and carrying out particular actions based on patterns that these systems are able to observe. There are certain data quality issues that may cause these systems to provide assistance based on inaccurate data and, thus, the person using such a system may be detrimentally affected. It is essential that information in these systems is sent and received in a timely manner as events are happening. Moreover, poor data quality can lead to poor information quality, which furthermore is closely linked to poor-quality contextual knowledge. The authors of this paper suggest a model to implement quality-control measures into Ambient Assisted Living systems. The way it works is to feedback knowledge gained during the system's reasoning cycle and using it for conducting further data quality checks.

Curé (2012) emphasizes the importance of high data quality in drug databases which are often exploited by health care systems and services. Poor data quality, e.g. the inaccuracy of drug contraindications, can have a severe negative impact on a patient's health condition. The author notes that data quality should be ensured in terms of data completeness and soundness. In his study, Olivier Curé presents special technologies to represent hierarchical structures of pharmacology information (e.g. the technology of the Semantic Web). Moreover, SPARQL is presented in the article as a query language for resolving issues of conditional dependencies (CINDs – conditional inclusion dependencies) for these graph-oriented structures. In Curé's study, an experiment was conducted in which CINDs in a drug database with both real and synthetic datasets were investigated. The author describes attempts to improve data quality in this drug database.

2.2.3 Assessing Data Quality

Pipino, Lee & Wang (2002) tried to answer the question of how good a company's data quality is. The authors describe principles to develop data quality metrics that are useful for measuring data quality. The core of their study was the presentation of three functional forms for developing objective data quality metrics. As an example, the Simple Ratio *"measures the ratio of desired outcomes to total outcomes"*, subtracted from 1.

Embury et al. (2009) talk about variability of data quality in query-able data repositories. Data with low quality can be useful, but only if data consumers are aware of the data quality problems. Quality measures computed by the information provided have been used to incorporate quality constraints into database queries. The authors describe the possibility of embedding data quality constraints into a query. These constraints should describe the consumer's data quality requirements. The problem that the research team attempted to address was that poor data quality is a consequence of information providers who define quality constraints. Their idea was to increase the level of data quality by incorporating quality constraints into database queries whereas users define quality such that domain-specific notions of quality can be embedded.

Heinrich & Klier (2011) propose a novel method for assessing data currency (one dimension of data quality, as mentioned earlier). Data currency is an important aspect of data quality management. In terms of quality in information systems, the authors distinguish between quality of design and quality of conformance. The latter is essential for this study, since it refers to *"the degree of correspondence between the data values stored in a database and the corresponding real world counterparts"*. As an example, data values stored in the database might not be up-to-date and, thus, lack quality of conformance. In other words, these data sets do not correspond with their real world counterparts.

Berti-Équille et al. (2011) propose a novel approach for measuring and investigating information quality. The scholars developed a model which can be transversally applied by users, designers, and developers. In their study, the quality of customer information at a French electricity company and patient records at a French medial institute were analyzed to create a multidimensional notion of multidimensional information exploration. Measures of data quality were stored in a star-like database. Quality dimensions (e.g. accuracy, response time, readability) are complimentary to analysis dimensions, which are analysis criteria such as date of data quality assessment, quality goals, and actors involved in the assessment. The model uses a GQM paradigm, which stands for “Goal-Question-Measure”. Applied to the multidimensional model of data quality, goals are set at a conceptual level (e.g. “*reduce the number of returns in customer mails*”), questions are asked at an operational level (e.g. “*which is the amount of syntactic errors in customer addresses?*”), and measures are defined for the quantitative level to quantify answers to questions (e.g. “*the percentage of data satisfying a syntax rule*”). A method, or a set of measurement methods to compute the measures are included in the model as well. Indicators as well as analysis criteria are included in the multidimensional data model for analysis of quality measures. A brief description of the indicators and analysis criteria is presented after the graph.

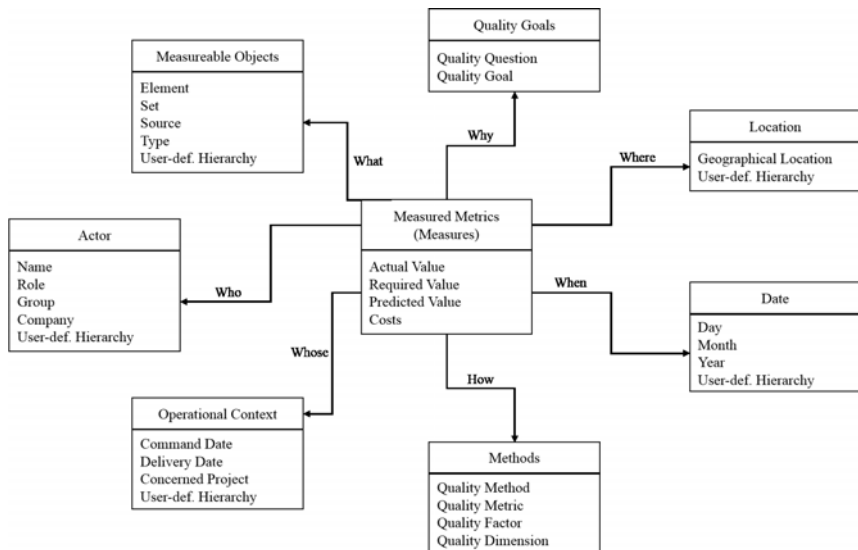


Figure 4: Multidimensional Data Model for Analysis of Quality Measures
Adapted from Berti-Équille (2011)

Data Quality and its Impacts on Decision-Making

How Managers can benefit from Good Data

Samitsch, C.

2015, XIII, 59 p. 13 illus., Softcover

ISBN: 978-3-658-08199-7