LUT
Lappeenranta
University of Technology

Data quality methodologies and improvement in a data warehousing environment with financial data

Niko Blomqvist

Examiner: Professor Mikael Collan

Second examiner: Maria Kozlova

ABSTRACT

Author:	Niko Blomqvist
Title:	Data quality methodologies and improvement in a data warehousing environment with financial data
Faculty:	School of Business and Management
Degree:	Master of Science in Economics and Business Administration
Master's Program:	Strategic Finance and Business Analytics
Year:	2019
Master's Thesis:	Lappeenranta University of Technology, 70 pages, 15 figures, 23 tables
Examiners:	Mikael Collan, Professor and Mariia Kozlova, Postdoctoral Researcher
Key words:	Data Quality, Data quality methodology, Data Quality Improvement

The goal of this thesis is to understand what is needed to successfully use a data quality methodology and give improvement suggestions with the given restrictions. The restrictions in this work are:

- it can be used with financial data and in a data warehousing environment
- it gives a quality score and it doesn't focus on a single issue or measurement
- it doesn't take a too big scope.

13 methodologies were found from the literature review from which one was chosen to be used in this thesis. Quality Assessment Using Financial data turned out to be the best methodology with the given restrictions. The methodology uses objective and subjective assessment methods and compares their results. Based on the results, the dataset under measurement gets a quality score.

Based on the empirical part we can say that there is a real world need for data quality evaluation and measurement. Unsupervised data quality can lead into massive losses in manual labor and money. We found several points you need to define and understand in order to use a data quality methodology successfully:

- what type is your data (numeric, string or binary) and what is its structure
- determine what you really want to measure and what results do you want to gain
- get acquainted with data quality literature

After using the context suitable methodology, you can improve your data based on the steps provided in the chosen methodology. If the chosen methodology doesn't provide improvement suggestions, you can use basic understanding and literature related to data quality issues in the environment your data is in.

TIIVISTELMÄ

Tekijä:	Niko Blomqvist
Otsikko:	Datan laadun menetelmiä ja parannuksia tietovarastoympäristössä rahoitukseen liittyvällä datalla
Tiedekunta:	School of Business and Management
Tutkinto:	Kauppatieteiden maisteri (KTM)
Maisteriohjelma:	Strategic Finance and Business Analytics
Vuosi:	2019
Pro Gradu -tutkielma:	Lappeenrannan teknillinen yliopisto, 70 sivua, 15 kuvaa, 23 taulukkoa
Tarkastajat:	Mikael Collan, Professori ja Mariia Kozlova, Tutkijatohtori
Hakusanat:	Datan Laatu, Datan laatu tarkistamismenetelmä, Datan laadun parantaminen

Tämän pro gradu -tutkielman tavoitteena on ymmärtää, mitä tietoja tarvitaan, jotta pystytään onnistuneesti käyttämään datan laadun tarkistamismenetelmää ja antamaan parannusehdotuksia laadun kehittämiseksi annetuilla rajauksilla. Tämän työn rajauksia ovat:

- sitä voidaan käyttää rahoitusta koskevalla datalla, joka on peräisin tietovarastosta
- tulokseksi saadaan laatumittari ja se ei keskity yhteen ongelmaan tai mittauskohteeseen
- soveltumisala ei ole liian laaja

Kirjallisuuskatsauksen perusteella löytyi 13 datan laadun tarkistamismenetelmää, joista yksi valittiin tässä työssä käytettäväksi. ”Quality Assessment Using Financial Data”-menetelmä osoittautui parhaimmaksi vaihtoehdoksi, sillä sitä pystyttiin käyttämään annetuilla rajauksilla. Valittu menetelmä käyttää objektiivista ja subjektiivista tarkistamismenetelmää ja vertailee niistä saatuja tuloksia keskenään. Tulosten perusteella tarkastelun kohteena oleva tietoaineisto saa laatupisteytyksen.

Empiirisen osuuden perusteella voidaan todeta, että datan laadun arvioimiselle ja mittaamiselle on todellinen tarve. Laaduttoman datan käyttö liiketoiminnassa voi johtaa liiketoiminnallisiin tappioihin ja tarpeettomiin työtunteihin. Työstä löytyi muutamia kohtia, jotka tulee määritellä ja ymmärtää, jotta datan laadun tarkistamismenetelmää voidaan onnistuneesti käyttää:

- datan tyyppi (numeerinen, merkkijono vai binaarinen) ja mikä on sen rakenne
- Määrittely siitä, mitä halutaan mitata ja saavuttaa datan laadun tarkistamismenetelmällä
- Tutustuminen datan laatua koskevaan kirjallisuuteen

Kun on käytetty datan laadun tarkistamismenetelmää sen antamissa rajauksissa, voidaan parantaa datan laatua tarkistamismenetelmän antamien ehdotusten perusteella. Mikäli valittu menetelmä ei tarjoa laadunparannusehdotuksia, voidaan käyttää yleistietoa ja kirjallisuutta liittyen datan laatuun ja eheyteen siinä ympäristössä missä data on.

ACKNOWLEDGEMENTS

This thesis has raised my understanding in data quality issues and helped with my own profession. The thesis process itself took longer than I personally expected but it was worth the trouble.

I want to thank my friends at LUT for good times. I want to thank my family for supporting me during the master's degree process. I also would like to thank the teachers at LUT for giving good quality teaching and sharing their knowledge on business analytics and finance.

Espoo, 23.4.2019

Niko Blomqvist

Table of contents

List of figures

List of tables

1 Introduction

In today's world companies have huge databases where data flows from various source systems. This data is then transformed and manipulated for further use and analyses for the end user. This is where the concerns with the data quality comes in. Data quality can mean different things to different users of the data. Based on literature there have been identified three main roles in information production: the ones who generate the raw data, the ones that maintain, store and secure the data and finally the ones that use the data. (Strong, et al., 1997) Data quality could mean for a system manager that the data that comes from the source system is unbroken and logically whole and the time stamps are correct. To an analyst the data quality might mean that historical data should be as accurate and correct as present data without the fear of having duplicate data items. The aspects stated above are all ways of viewing data quality. This means data quality isn't any more measuring how accurate or reliable the data from some survey is, but it can mean huge costs to companies if the data is flawed and it can't be used for analyses.

Based on a report by Laatikainen & Niemi (2012) in Finland it is estimated that costs due to data quality issues are around 10 billion euros per year. This has caused companies to not trust their data and causes difficulties to make business decisions based on poor data. It also influences smart models and machine learning techniques. Since data is the fuel for the machine and model, the results are as good as the data behind it. This has created a need for understanding and assessing the data quality to improve it and and making sure the used data is reliable. From these various assessment techniques there has formed data quality methodologies. These methodologies aim to measure and improve the data. (Batini, et al., 2009)

Some literature exists on the assessment of data quality. From these articles only, a handful give improvement suggestions based on the assessment. Also, many of these articles only list comparison of various methods and don't show how these methods are used or how the assessment results should be understood. This brings the need for my thesis. Successfully finding a correct methodology for a real-life data and suggesting ways of quality improvement based on the results.

The reason above gives reason for my work. In this thesis we concentrate on finding and using a data quality methodology in data warehousing environment that has financial data.

Based on the assessment results gained from the methodology we give improvement suggestions. This show how the assessment results can be utilized for the improvement.

1.1 Focus and research questions

My master's thesis concentrates on data quality. The focus is on methods how to asses/measure data quality and give improvement suggestions based on the assessment. The goal of my master thesis is to successfully find useful and informative measurement for the chosen dataset from a real-life company and using the measured data to find possible fixing points. These results would help in my day job and support the learning process. Also, this thesis helps understand what methodologies are available for data quality assessment and what steps to take to choose the correct methodology to use for your data.

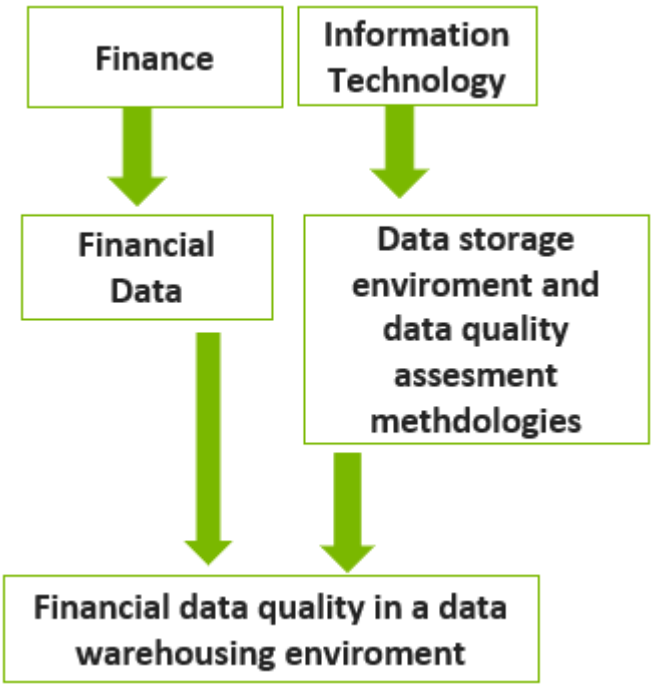


Figure 1 How data quality is linked to different fields of science

In figure 1 we show how data quality in this work is linked to different fields of science. In this thesis we focus on financial data which links to the world of finance. The methodologies and data storage come from information technology. This leads to the subject and focus of this thesis data quality for financial data in a data warehousing environment.

From the focus of my thesis and the desired results I formed one research question that is supported by three supporting questions:

1. How to successfully use a methodology to assess data quality and improve it based on assessment results?

sub question:

1.1 What methods can be used to measure data quality?

1.2 What do you need to know about your data to use data quality methodologies efficiently?

1.3 How can the findings from using the methodology help on data improvement?

The research question aims to understand what's a good method at assessing data quality relating to the data set chosen to be analyzed in the empirical part. The sub questions support the main question by answering what steps are needed to successfully use a data quality methodology.

1.2 Research methodologies and data set

This research is a qualitative research case study since it is related heavily on theory and testing if the methodology presented in previous research is still valid. At the same time, it is tested if the methodology is used into the correct purpose meaning it can be used to dataset chosen in this research. (Fick , 2009)

The data set used in the empirical section is a part of a real-world data set from a company, which is a financial institution. The data set is defined in more detail in chapter 4. Because the data set is taken from a financial institution, the data to be analyzed can't contain certain information, like customer specifying data. Therefore, the data must be anonymized so that it can be still used in the research without losing its reliability and integrity. The masking process on how the data is manipulated won't be discussed in this thesis. The details of this data set are introduced in the empirical part of my thesis starting with descriptive statistics.

After describing the dataset, the chosen methodology is used on the data set based on the literature review. Based on the results gained from the chosen methodology the data will get a score and there will be given suggestions on how the results could be improved. These improvement suggestions are given based on the suggestion found in broad theory of data quality improvement or they are given by the chosen methodology. The broad theory of data quality improvement is discussed in chapter 3.

1.3 Structure of the thesis

The thesis starts with an introduction to the topic. It aims to raise the interest towards the reader to read further on and understand what this thesis is about. It also aims to give an understanding what is the focus and what are we trying to find out. The research questions are introduced, and key concepts are explained. The summary of the structure of my thesis can be seen in figure 2.

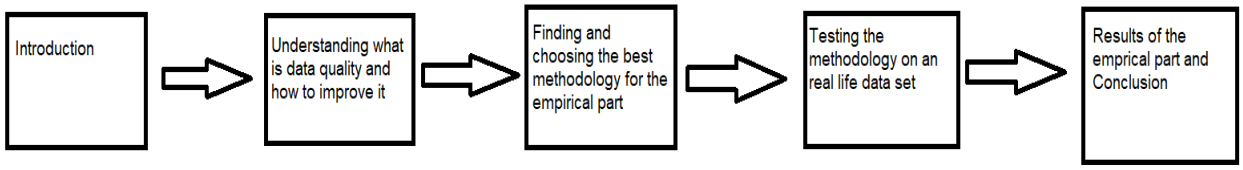


Figure 2 Structure of thesis

In the theoretical section consists of two chapters. In chapter two, we take a glimpse of what information is needed when data quality assessment is made in broad perspective. This chapter also discusses common data quality issues in a data warehousing environment and ways of data improvement. The third chapter begins with the literature review where we aim to explain how we found the articles and what literature have been written on data quality methodologies. Based on the review three methodologies are chosen for further analysis. These methodologies are first looked at in more detail and the chosen methodology is discussed in-depth at the end of chapter 3. This links the broad perspective to specific methods of data quality assessment.

Empirical part has a case example using a real-life data set that is gained from a real company using the chosen methodology. The goal is to successfully use the chosen methodology and suggest further actions on trying to improve the quality. The empirical part begins with basic statistics of the data set and their introduction. After this the chosen methodology is used on the data set and it follows steps that are described in it.

The discussion and conclusion are in the same chapter. discussion contains findings and results of the empirical part and summarizes the relevance of the measurement to real world need. Here we also look at the expert comments on the assessments they made. Conclusion sums up the thesis, discusses results and findings related to the research questions and suggests future research that might come up during the thesis process. This part also includes critical thinking towards the master's thesis process and gives suggestions what in my opinion could have been improved.

1.4 Key Concepts

Data Quality (DQ):

Data Quality is a way of examining your data from different points of view. These points of view can be referred to as data quality dimensions which will be discussed in chapter 2.1. Data quality has become an important factor in today's world since lots of data flows in companies and decisions and actions are taken based the data that is transformed into information. Your report is as good as the data that you have used to make it. (Technopedia, 2018) DAMA (2019) suggest data quality concept should be divided into two: process to improve data quality and characteristics related to high data quality. High quality is determined by the data consumer. This means that everyone has their own understanding on quality and the concept high quality is different depending on the context.

Information Quality (IQ):

Information quality is the way of examining the quality of the information from different viewpoints. Since information is created from your data, the quality of your information is as good as the quality of your data. In information quality the key is to understand the data to form information and ask what information is relevant. Like data quality, information quality can be viewed from different perspectives in other words dimensions. (Miller, 1996) Data Quality literature has shifted from talking about data quality issues to information quality issues.

Information System:

An Information system is a defined as a system that has many different parts. These parts together create the information system. Information system parts can be information itself, system connections, IT hardware and its network connections, software that is needed to store and handle the information and show it to the end users. There are different types of information systems for example: Management information system, decision support systems or operations support systems. The information system is usually built for a need or purpose. (Technopedia, 2018) In the literature review there were used concepts monolithic and distributed information systems. Monolithic refers to an information system that has structured data within one system. In methodologies that focus on monolithic information systems data quality issues happen when different systems exchange

information. A distributed information system means that the information comes from different sources or systems that are connected. (Batini, et al., 2009)

Total Data Quality Management (TDQM):

Total Data Quality Management is methodology for measuring data quality. It is a foundation methodology for many other data quality methodologies. This methodology can be implemented in any environment and it is not restricted or created for a specific purpose. TDQM has four main parts: define, measure, analyze and improve. TDQM is examined in more detail in chapter 3. (Francisco, et al., 2017)

Data Warehouse Quality (DWQ):

Data warehouse Quality focuses on the relationship between quality objects and design options in data warehousing. Data warehouse quality is Like TDQM, it contains the same four parts. The downside in this method is that the improvement phase is only mentioned and not given too much emphasis. DWQ is examined in more detail in chapter 3. (Batini, et al., 2009)

Quality Assessment of Financial Data (QAFD):

Quality Assessment of Financial Data is a methodology that focuses on defining standard quality measures for financial operational data. The goal of this method is to minimize the cost of measurement tools. Unlike the previous two methodologies QAFD contains five steps: Variable selection, Analysis, Object measurement, Qualitative subjective measurement and Comparison. This method is designed only for the use of financial data. Also, this methodology doesn't give improvement suggestions. QAFD is examined in more detail in chapter 3. (Batini, et al., 2009)

2 Understanding data quality

This chapter looks at data quality in the big picture, understanding what data quality is, where data moves and what data types are. After this we look at the different data quality dimensions which can be regarded as aspects looking at your data. The chosen book for the dimensions is written by Batini & Scannapieco (2016). The book looks at dimensions related to structured data and it reason why it was chosen. Even though the book is named information quality, it is related to data quality since it is just an updated version of a book the authors wrote in 2006, which was called Data Quality Concepts, Methodologies and Techniques.

In this chapter we also look at data quality issues and improvement suggestions. The aspect taken is issues that affect data quality in data warehouse environment. The reason why we are looking only at issues in data warehouse environment is due to the storage environment of the real-life data set used in the empirical part. This gives an understanding on what possible problems accrue in this environment and what possible fixes are suggested.

2.1 Data Quality, Information Systems and Data types

Data Quality can be referred to be part of information quality. The word information contains both data and information quality. Data is at the early stages of information and information is after the data has been assessed and understood as information on a later stage. To put it more clearly, information is data put into a context. Example being 323 which is a number or in this case data. But when it is given a context account balance, then it is information meaning the account balance is 323. (Strong, et al., 1997) If data quality is put into an even bigger perspective, it is one of the important parts of data governance. Data governance can be best understood by figure 3.

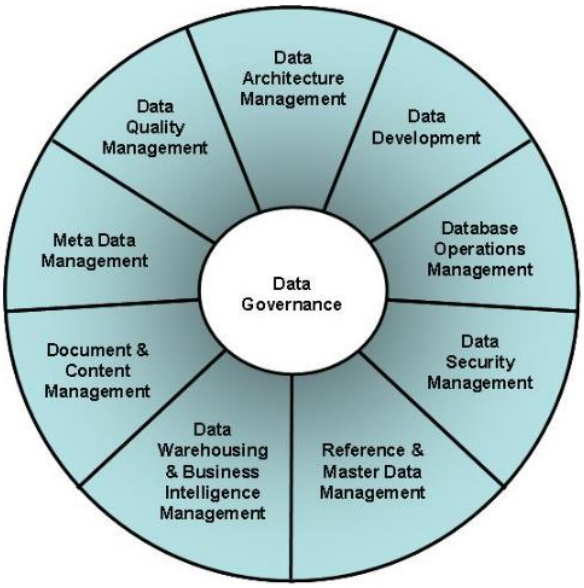


Figure 3 Data Governance and its aspects (Stiglich, 2012)

Data Governance can be understood that it is the thing that connects and contains all the aspects that are seen in figure 3. Basically, it is core component that is linked to the rest of the aspects or dimensions. Data governance sees data as an asset to your company and it makes sure that the asset is protected, and the quality level defined by your company is maintained at that level. (Stiglich, 2012) Related to data quality and data governance is master data and master data management. Master data is the core data of ones' business. It is the data that should be the same to all departments of the company and everybody should have access to it. Master data management refers to the managing of data quality and the quality should be ensured at the master data level. If master data quality is at an acceptable level all other data that is refined from it can be regarded as having significant quality. (Laatikainen & Niemi, 2012)

To understand how you want to evaluate your data quality, you must understand that there are various data types and information systems the data flows in. Based on the data type you can't use all methodologies in data quality assessment. Data types can be categorized into three main categories: structured data, semi structured data and unstructured data. (Batini, et al., 2009)

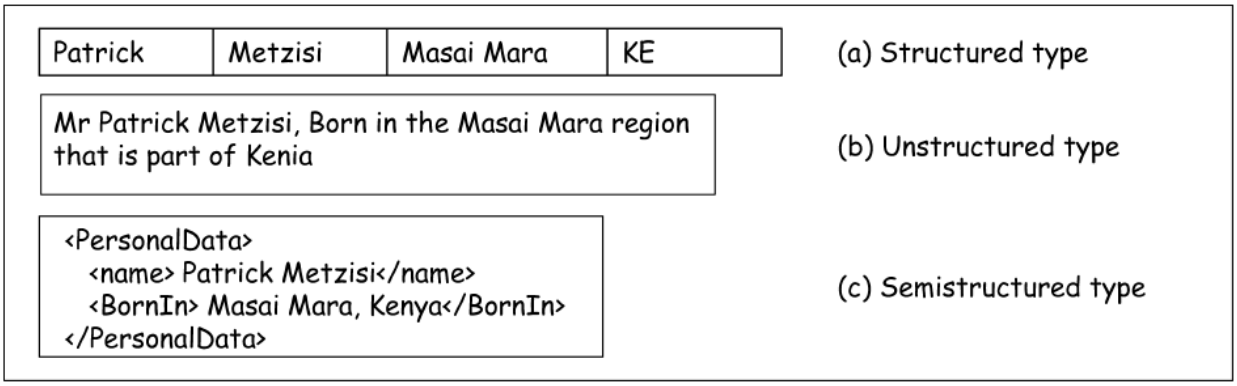


Figure 4 Different representations of same real-world object (Batini, et al., 2009)

From figure 4, we can see the basic structures of data and what they might look like. Structured data can be used easily for further analyses since it can be represented in separate fields and the columns the data is in have been formatted correctly. Also, it is typical that the length of the data is limited and known. Unstructured data usually is in form of free writing and as seen in figure 4 has no limitations or formatting. Semi structured data is a combination of structured and unstructured data. This structure type might have the same class the data belongs to but different attributes. (Robb, 2017) This thesis will concentrate on methodologies that can be used on structured data, because the dataset to be analyzed is considered a structured one.

Data flows in different kinds of information systems. These systems can be used to organize and manipulate the data from the source system to a desired structured form that is usable for the end user. Batini & Scannapieco (2016) introduce six main types of information systems from which two of them are defined further. This is due to the nature of the empirical part and the environment the data is stored in. The chosen systems are Data warehouse (DW) and Cloud Information systems.

Data warehouse is a centralized location to store data from different sources and it is designed to support some tasks. These tasks can include things like business analytics. The biggest problem in DW is the cleaning and integration of data from different sources. This means that once the data is stored there or it is no longer useful it must be cleaned from the warehouse for it to be removed. The other problem refers to a situation where all the different data from different source systems must be integrated so that no matter what is the source the data can be found and connected. Example being that one customer has only one

customer key. This key connects the customer to all the services he or she has no matter what the source is. (Batini & Scannapieco, 2016)

Cloud Information system is a system that enables centralized data storage and access to computer services and other resources via network. These systems have autonomy which refers to a hierarchical system in which there are different levels. Based on the level, the user is assigned rights to different tasks, locations, duties etc. These systems also have heterogeneity which means that it considers all types of semantic and technological diversities among different systems. (Batini & Scannapieco, 2016)

Since the data set to be used is from a financial institution, we should define what kind of data can financial data be. Luckily there is research on this subject and four main groups of data have been discovered:

1. Registry data
2. Daily data
3. Historical data
4. Theoretical data

Registry data is used for defining the chosen financial instrument or product say bond information. Daily data is things like price changes. Historical data is information across some time line and is meant for describing information on a specific date in time. Theoretical data refers to the result of some financial model that is used. (Batini & De Amicis, 2004) This thesis will concentrate on historical data, registry data.

2.2 Data quality dimensions

This section discusses different dimensions or aspects on viewing data quality. The idea here is to understand that looking at data quality in different dimensions means different things. Broadly Data quality dimensions can be imagined as an umbrella where each dimension captures a specific aspect of Data Quality. This is illustrated in figure 5. Data Quality dimensions can refer to extension of data or intensions of data. Extensions of data refer to the data values and intensions refer to a schema of a model. (Batini & Scannapieco, 2016) Batini & Scannapieco (2006) also state, that choosing the data quality dimension is the first step of any data quality related activity.

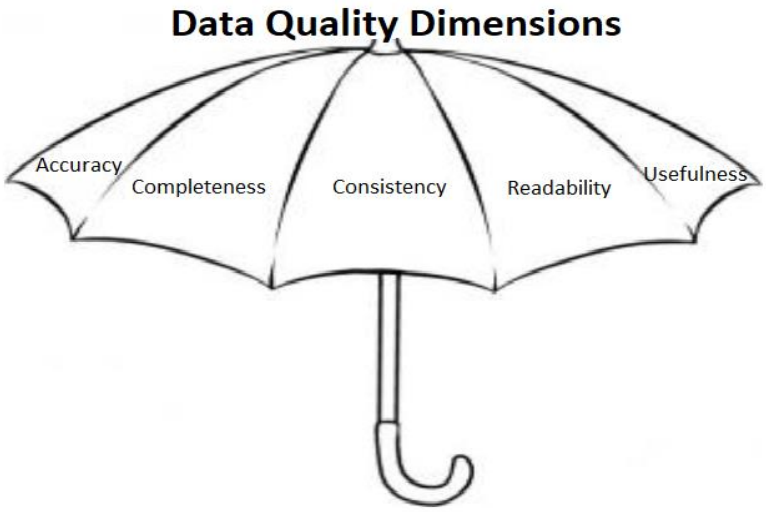


Figure 5 Data Quality Dimension umbrella

Batini & Scannpiece (2016) have generated eight clusters of data quality dimensions. The clusters are: Accuracy, Completeness, Redundancy, Readability, Accessibility, Consistency, Usefulness and Trust. Trust won't be discussed in this work since it is related to big data and web data which are not defined and related to this work. Also, accessibility won't be discussed further since it is related to access of data and the understanding of the language used in the data. The remaining six dimensions will be defined and given examples to better understand them. There will also be given additional information, from different sources, if it is seen relevant to this work and understanding the dimension. Each cluster is also given examples on how they could be measured.

Accuracy can be defined as how close the data is to what it is supposed to represent. Usually it is referred to be correct or incorrect. An example would be a customer's name is supposed to be "Niko", but data is shown as "Nko" which is incorrect. Accuracy can be further divided into structural accuracy and temporal accuracy. Structural accuracy refers to values that can be considered unchanged and not volatile like a person's social security number or a bonds name. The other form of accuracy, temporal accuracy, refers to volatile and time related data. Examples can be stock price which is relevant only in that specific time and the data changes all the time. This dimension can be measured by calculating the number of correctly populated columns divided by the total population in that column. Example you have a total population of 10 from which 5 of them are correctly populated this gives you an accuracy of 50 % or 0,5. (Batini & Scannapieco, 2016)

Structural accuracy can be defined even further into two sub sets: semantic and syntactic accuracy. The difference of the two can be best defined with an example seen in table 1.

Table 1 Semantic and syntactic accuracy

ID	Job Title	Name
1	Janitor	Mike
2	Cheuf	Bob

From table 1 we can see job titles and names in a data warehouse. Here we can see that job title for id 2 to is incorrectly spelled. There is no job title “Cheuf”, but Chef would be correct. This can be referred to as syntactic accuracy. It means that all attribute values relate to the attribute itself in this example all attribute values should contain job titles. Semantic accuracy is then related to the relations of two attributes. If Mike isn’t a janitor, then there would be an error in syntactic accuracy. If he is then there is no problem in this dimension. Basically, syntactic accuracy refers to the information value given by the relations of the attributes. (Batini & Scannapieco, 2016)

Completeness is defined as how complete the data is. The idea here is that you know what the data should contain and by that you can compare is the data complete or not. Completeness can be divided into three types: Schema, Column and population. Schema completeness refers to that the amount of non-missing values related to the schema. Column completeness refers to the amount of values missing related to that specific column. This means that the amount of missing values in that column divided by the total amount of values in that column times 100 gives the completeness of that column as a percentage. Population completeness refers to the amount of missing values related to the whole population of the chosen data set. Here you calculate the amount of missing values in the whole dataset divided by the total population of the dataset times 100. Example: you have a dataset that is supposed to have a total population of 100, but only 50 of them are populated so based on the calculation the completeness dimension of that dataset is 50%. (Batini & Scannapieco, 2016)

Consistency is related to semantic rules that are defined over data items. Integrity constraints are related to this dimension since they can be used to test the consistency of the data. An example of an integrity constraint can be a rule or constraint in which say Loan is a relational schema and there is an attribute called margin on that loan. The constraint can be that margin must be something between 0 and 20 percent. The constraints don’t

have to be just related to one attribute but can contain many attributes depending on the relational schema. These types of constraints can be referred to as intrarelation integrity constraints. Another form is called interrelation integrity constraints. An example of these constraints can be that the loan amount on the application must be equal to the loan amount drawn. In this example the loan application and drawn loan amount are their own relations. Consistency can be also viewed as internal or external consistency. Internal consistency means consistency within a data set or system and external means consistency between two system. External consistency is measured when viewing data that should be similar between two information systems. The measurement of consistency is quite simple since all you need to do is check if the data follows the given constraints. If you have a data from which 10 attribute values should follow the given constraint and only 5 of them do you have a consistency rating of 50% on that consistency constraint. If you have more than one consistency constraint you sum the results of each individual constraint together and divide by the total amount of constraints to get the consistency score on the whole data set. (Batini & Scannapieco, 2016)

Usefulness of dimensions is related how the user of the data sees the data. Is the data useful for that person in that situation? An executive at a contact center doesn't necessarily find data related to macroeconomics useful when measuring the performance of the unit. Another way to view usefulness is that a person gains advantage from the useful information for example a BI analyst gains a deeper insight of loans drawn from the bank when using data that is related to loans drawn. More often usefulness is used in picture quality. In this case it is simpler to say if the picture is useful or not. (Batini & Scannapieco, 2016)

Related to usefulness dimension can be also related timeliness relation, since the data might be useful only at a specific time. Timeliness can be simply defined as the age the data is appropriate or useful for that purpose. (Richard & Diane, 1996) In the timeliness dimension you can also include currency and volatility dimensions. Currency means in this perspective that how frequently the data is updated. Since currency is related to value of the data at that time it can be said that the data is either high currency or low currency. An update in the information of a customer's address where the person lives can be considered high currency since it is relevant and correct at that time. Low currency would be an address that hasn't been updated and the person doesn't live in the address known to us. Volatility in time dimension means the period the data remains valid. A person's address can be regarded as low volatility data and stock prices when the stock exchange is open can be regarded as

high volatility data. Also, things that are regarded as non-changing like birth date are regarded as stable. (Batini & Scannapieca, 2006)

Validity is a dimension which checks if the data attribute values are consistent with domain values. Domain values can be a table that defines what the data attributes should be, or a set of attribute values determined by data quality rules. If the attribute values don't meet with the domain values validity is lost. It is good to note that the attribute value can be valid, but it still might not be accurate. Recall that accuracy can be syntactic, which means that it is not accurate in that context. (Dama-DMBOK , 2019)

Readability dimension is related to the understating of the data. This means how easily person can read the data and understand it. Also, related to this dimension can be clarity of the data or simplicity. Readability dimension can be heavily related to data quality in inquiries since the person answering the query needs to understand the question i.e. the question been answered must be readable and understandable. Readability can also relate to schema quality dimensions. In this context it means the schema is understandable by any user, it is clear, and it isn't too complicated. For readability to be measured many users should be asked if they can understand say the schema or not. Based on their assessment an average can be built and that would be the readability measurement. (Batini & Scannapieco, 2016)

Redundancy is also related to schema quality dimension. Redundancy can be divided into minimality and normalization. when talking about schemas minimality means that every element in the schema is used/introduced only once and nothing can't be removed without taken the risk of removing some information aspect. Normalization is used in relation models. In a relational model elements or records are linked together with a specific key. This key allows the linkage between different data sets since the key is unique. For example, a customer has a unique key which can then be linked to an account, services, credit cards etc. Normalization is related to the functional dependencies in the model. (Batini & Scannapieco, 2016)

From the defined data quality dimensions, we can conclude that the most appropriate dimensions or aspects to look at in the empirical section are accuracy, completeness and consistency. Redundancy and readability dimensions are left out since they are dealing with schema qualities and this thesis is not concentrating on building or testing a schema. Usefulness isn't looked at since it depends on the user of the data. Timeliness is left out due

to the nature of the data set and we do not have access to the time factors related to the data set.

2.3 Common data quality issues in data warehousing and improvement suggestions

This section introduces suggestions on data quality issues and improvement. Some of the suggested methods seem to be very self-explanatory, but improvement doesn't always have to be difficult. The main idea is to understand what the correct way is to improve the quality without risking losing quality while trying to improve it. Basically, the best way to fix a data quality problem is to find the main source of the problem to make sure the problem doesn't happen again.

Data quality issues can generate almost at any stage of the data's life cycle. Singh & Singh (2010) have classified four main classes of data quality issues regarding data warehousing: Issues at data sources, Issues at data profiling stage, issues at data staging or ETL and Issues at data modeling (database schema design) stage. Since schemas aren't discussed we won't go into further detail with issues regarding them. The focus is to understand the first three classes. Figure 5 illustrates the four main classes and how the classes are linked. Like a data quality problem that is found already at the data source flows all the way to the schema. This means to fix the problem it must be done at the source.

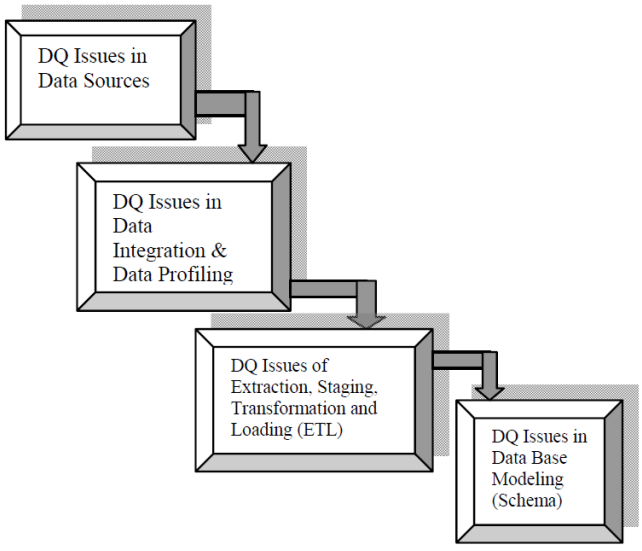


Figure 6 Stages of data warehouse problems (Singh & Singh, 2010)

The most common problem related to data quality come from issues at the data's source. The sources where data comes from can best seen from figure 7.

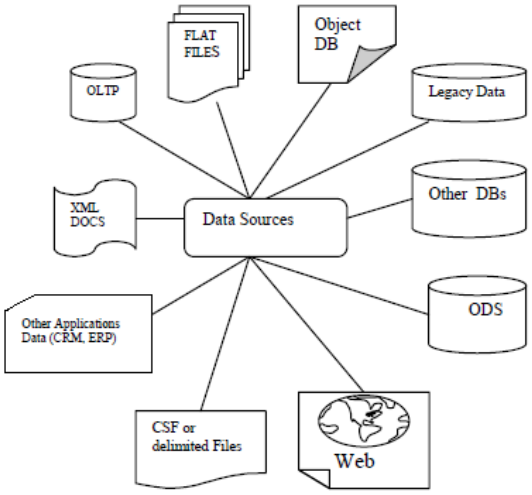


Figure 7 Possible sources of data (Singh & Singh, 2010)

As we can see from figure 7, there are a lot of different sources where the data might come from and this is one of the reasons that might cause the problems. Since the location, supplier and format of the data is different these can cause problems. The issues might relate to timing of data when suppliers of the data deliver the data at different times. Incorrect formatting of the data and misspelled or missing data from the source are also common issues that happen. (Singh & Singh, 2010)

Usually it requires the company or process/system owner to be in touch with the data supplier to fix the issue. The problem can also, be in the source system itself so the information can be in some field that needs to be corrected to get the data problem fixed. An example could be gender information ticked in a wrong box in the source system and this triggers a data quality problem regarding this one suspect. Another would be the supplier sending yesterday's data again the next day. This would mean the same data would be warehoused twice and it would influence data as information. Here we assume we have not quality controls to check if it is yesterday's data or not.

The data profiling stage means the analysis of your source system. This stage is usually ignored since it is more important when new source systems are brought. An example could be a new source system brought as a new part of the warehousing environment or new data sets are implemented into the ETL process. Issues at this stage can be lack of analysis and numerical data for inbound files like min, max, file size, standard deviation etc. In general, data should be documented, meaning we can see from the document what data should be loaded and what not. This document can be referred to as the technical document, which is

usually supplied by the process owner or data provider. Also, problems like user written data profiling or SQL queries cause data quality issues at the profiling stage. (Singh & Singh, 2010)

Possible ways to improve the quality at this stage could be making sure every project that brings new systems or data to the warehouse environment understands data quality and potential issues regarding it. It should be implemented to the companies' culture, so it is understood by people who work with data. Another crucial improvement is making sure you have a technical document of the dataset and it is regularly updated during the dataset's lifecycle. This means the data should be regularly tested against the technical document to make sure the data is what it is supposed to be stated in the document.

The ETL (Extract, Transform, Load) process is the most crucial stage where data quality problems can happen. It is also the best place to perform checks on the quality of your data. This is because the data is transformed and formatted during the ETL process. Here is also executed data cleansing which refers to cleaning data to improve the accuracy or completeness dimensions. For example, not storing columns that only contain Null or missing values. (Singh & Singh, 2010) A common issue during ETL process can be key or attribute level problem which can be seen in figure 8.

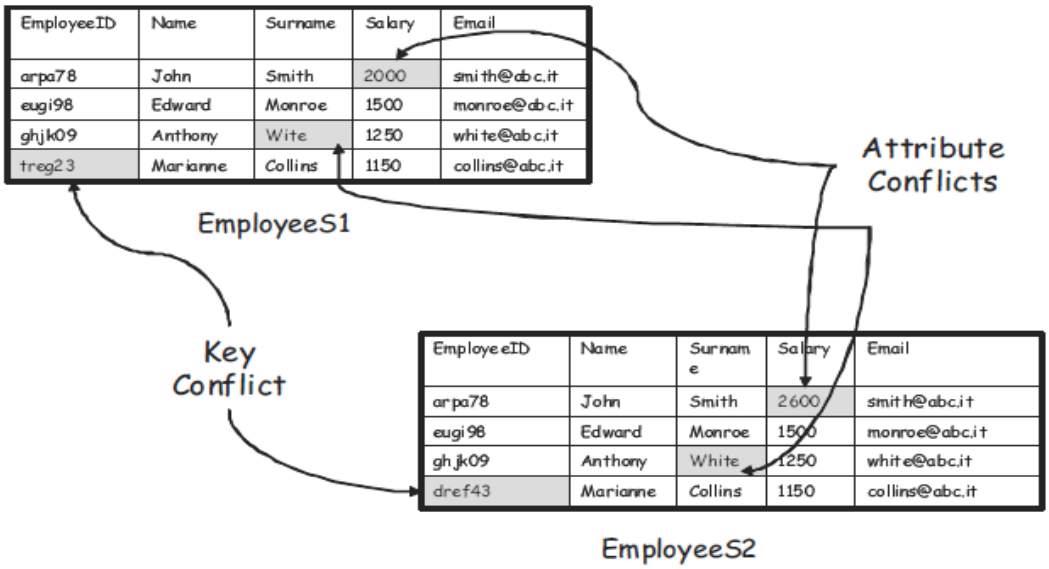


Figure 8 An example of key and attribute conflicts (Batini & Scannapieca, 2006)

In figure 8 the problem is usually generated during the ETL-process where for the same id is generated two unique keys. This means that in the example case the same employee has

two unique ID's. This can happen due to failures in the ETL-process for example logic failures. The attribute issues are related to information that is not matching between different data sets. From figure 8 we can see that for EmployeID arpa78 the salary amount the person gets is not the same between the two date sets EmployeeS1 and EmployeeS2. As we learned from previous section, these issues affect dimensions like accuracy of the data set. (Batini & Scannapieca, 2006) Singh & Singh (2010) also suggests data quality issues like lack of proper extraction logic and loss of data during the process. This loss refers to data that is disregarded. Example could be the data not meeting some rules in the ETL process or it being rejected due to quality problems that come from an earlier stage.

The problems in the ETL process can be fixed for example by generating business rules and checks at different parts of the ETL process. The checks are for us to make sure the data is transformed according to the required rules. The rules help supporting IT functions to understand the whole ETL process. If the support understands the process and the rules, they can fix problems regarding the process logics if they are broken or need to be adjusted. Also, Singh and Singh (2010) suggest that no user written code should be used with the ETL process because it usually causes quality issues. Instead of user written codes tools that are designed to be used for the ETL process should be used.

Usually the improvement suggestions only fix the found issue. To fix the issues in big picture a whole company wide data governance program should be initiated, so that everybody understands their role with data and the source problems can be fixed. It is noted that the human element in data quality improvement is as important as the technical part. This means that you can't fully eliminate the human, since in the end the data owner is responsible for the data. (Dejaeger, et al., 2010)

3 Methodology assessment

In this chapter we choose three data quality methodologies from all possible data quality methodologies that are found by the literature review. Then we look in deeper detail at the chosen methodologies. In the end of this chapter one methodology is chosen for the empirical case-study on the real-life data set to evaluate the quality of it and suggest improvements. The chosen methodology is defined in full detail.

There are common phases that can be found in each data quality methodologies. These phases can be divided into Assessment and Improvement phase. Each phase contains steps. The steps the methodologies take can be different, but the structure in each methodology is similar.

Assessment phase starts with the analysis of the data by looking at the data as data and then asking from specialist and forming the data to information. This is analysis is done to completely understand the data and IT related to the data forming. A summary of assessment phase steps can be seen in figure 9.

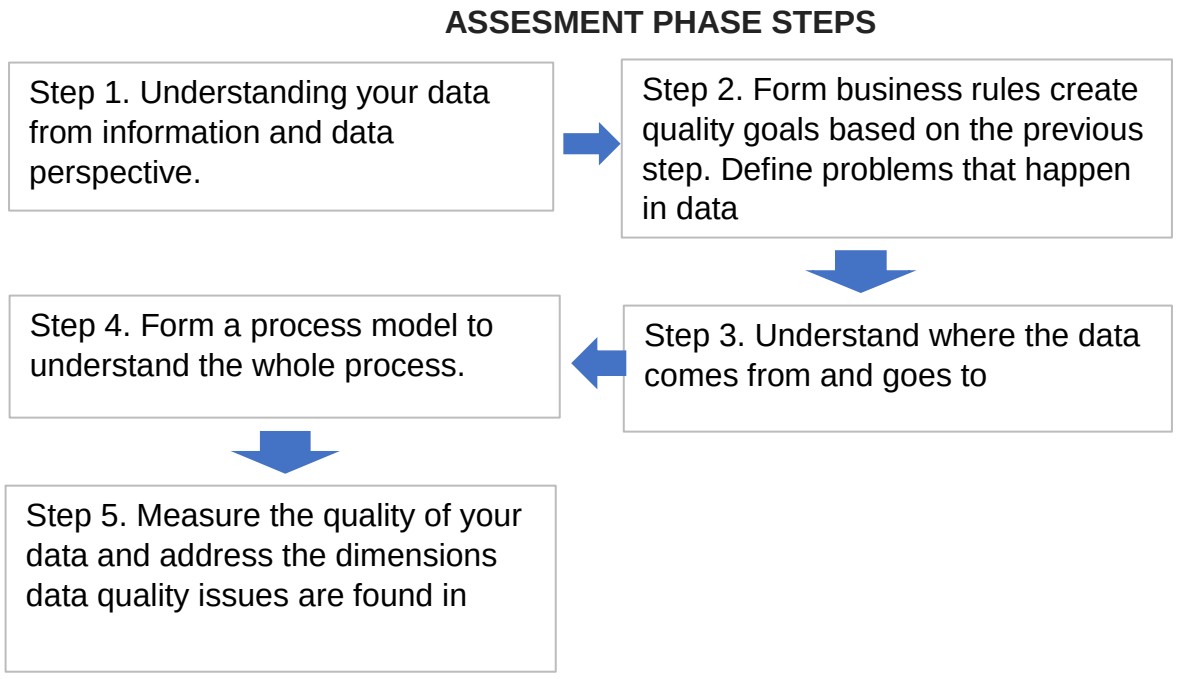


Figure 9 Assessment phase steps

Based on the first step we move on to step 2 by creating rules and understating issues related to the data. In step 2 you can also set targets to reach regarding the quality of the data. After knowing the key issues within the data, you identify the key areas where the data comes from and goes to. After this, in step 4, you form a process model to illustrate what is

really happening. The model should contain everything related to the data and its movements. Then you measure the quality and define which data quality dimensions influence the issue in phase 5. The measurement in the step 5 can be objective meaning it is done based on quantitative metrics and doesn't need deep understanding of data or it can be done subjectively where you measure the information quality of the data, basically a qualitative approach. (Batini & Scannapieco, 2016)

The improvement phase contains steps on the improving the data based on the results of the assessment phase. The improvement phase can be divided into different steps which are summarized in figure 10.

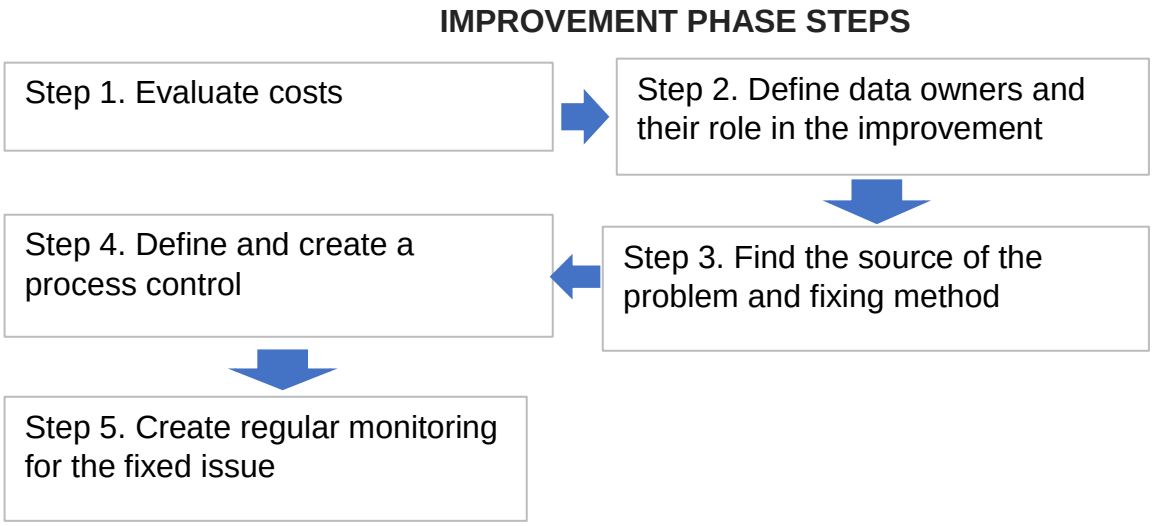


Figure 10 Improvement phase steps

Step 1 begins with evaluating the costs. These costs can be direct costs meaning the issue has a direct impact on the information quality or indirect meaning the costs are not directly related to the issues but affect other things as well. Example might be a customer's address is written wrong. The direct cost is the issue that the address is wrong and indirect cost would be on an analysis of where customers live. The issues effects directly to the customers information and indirectly to the analysis. After evaluating the costs, you assign and define the information/data owners, the people responsible for the process and define what is there role on the improvement process and make sure they are aware of the issue at hand. Then you find the source of the problem and find the right strategy or technique to fix it. Then you create process control to check that it is fixed, and it won't happen again, or you must redesign or adjust the process to make sure the issue doesn't repeat. After this the

improvement end with monitoring so that it will not happen again. (Batini & Scannapieco, 2016)

3.1 Data Quality Assessment methods from the literature

Based on the research question the literature review focuses on finding literature on data quality assessment methodologies. We look at what literature have been written on data quality methodologies and aim to find a list of them for further analysis in the following sections. We look at what the methodology is meant for and can it be used in data warehousing environment and with financial data. Financial data is regarded to be structured data and it can be numeric or character format. Based on the literature review the decisions on what methodologies will be looked at in further detail and what are left out of the scope. We also look at literature that discusses results of data quality methodologies and how these methodologies can be used.

The literature was found using databases found in LUT Finna: SpringerJournals, ACM – Association for computing machinery, Science Direct and Google scholar. The search results were narrowed to research / articles that could be accessed. Literature was also found by looking at references in other articles related to data quality. The key words used to find literature from the databases were: Data Quality, Data Quality Improvement and Data Quality Methodologies.

Data quality methodologies can be used in every environment where is data. Only restriction being the ones methodology itself sets. The environment doesn’t have to have real or production data it can even be tested on test data. J. Held (2012) made a research on how data quality methodologies could be used on test data. Here he wanted to make sure the data was useful for development purposes. This means new things could be developed using test data before moving it to use production data. Cases like GDPR might have an effect where development should be made. The research also stated that methodologies are good at finding the dimensions data has problems in, but the expert opinion is needed as well.

Christoph Samtisch (2014) discussed data quality assessment in general without the use of a methodology. In his book he suggested prior research done on data quality assessment. An example would be implementing data quality checks on a query. The quality constraints could be embedded to database queries. By these queries the quality improves. The problem is that the people who create these queries are not the data users. Another

assessment technique suggested was comparing the stored value against its real world counterpart. This is because the stored values might not be up to date and therefore lack quality. This method is especially useful when comparing data warehouse information to their real world counterpart say loan details in warehouse versus current real world situation of the loan. The idea of data quality methodologies is usually to give a framework for analyzing the quality. The framework usually contains different data quality dimensions that are suited for that situation. (Samtisch, 2014)

Wang (1998) Found a methodology named Total Data Quality Management. This methodology has been used as the foundation methodology for many other data quality methodologies. The methodology, as the name suggests, can be used in any information system environment. The goal is to create, improve and understand the information product (IP) cycle. The cycle is fully covered why the word total is used. The methodology doesn't set any restrictions on the data (Wang, 1998)

Jeusfeld, et el. (1998) created a methodology called Datawarehouse quality methodology (DWQ) which focus entirely on quality improvement in data warehousing environment. In this methodology the people using the data define the quality and try to achieve the set goals. The methodology can be used with structured data and it is specifically designed to be used in a data warehouse environment. The downside here is that it doesn't suggest improvement suggestions and to use the methodology the person needs access to the data warehouse and the ETL-process that transfers the data there. (Jeusfeld, et al., 1998)

English (1999) Total Information Quality Assessment (TIQM). TIQM can be used in monolithic and distributed information systems and it focus on the architecture perspective. It was also, the first methodology to evaluate costs in the assessment. The methodology is especially useful in data warehouse projects. (English, 1999) A similar methodology that evaluates the costs is Cost-Effect of Low Data Quality or COLDQ that was developed by Loshin in 2004. In this methodology the goal was to create a quality scorecard that supported the evaluation of costs. This methodology is also for monolithic information systems. The methodology provided the first detailed list of costs and benefits that can happen or can be gained by good or bad data quality. (Batini, et al., 2009)

TIQM and TDQM have been used in testing the quality of data in companies that use customer relationship management (CRM) as customer retention and profit increase. These methodologies were chosen because they give most detail on the whole data quality

process. The aim here was to assess the pros and cons of the two methodologies in CRM environment. The results were that TDQM succeeded at improving the data quality in the long run but failed give explicit metadata, treatment to data quality metrics and didn't discuss costs related to poor data. TIQM succeeded in giving attention to metadata and its weakness was the need of at least one expert. The research concluded that the best way would be to use a combination of the two methodologies. (Francisco, et al., 2017)

A methodology for distributed information systems and structured data is Data Quality in Cooperative Information Systems of DaQuinCIS developed by Scannapieco et al. in 2004. This methodology assesses quality issues between two information systems that work cooperatively. This methodology suggested two modules that would help on assessing and monitoring cooperative information's systems: the data quality broker and quality notification. (Scannapieco, et al., 2004)

Lee, et al. (2002) developed a methodology named A methodology for information quality assessment (AIMQ). The methodology focuses on benchmarking and is especially useful when evaluating the quality questionnaires. The methodology uses PSP/IQ model which is a 2x2 matrix that focuses on quality based on the users and managers perspectives. The downside of this methodology that it doesn't suggest any improvement tools or perspectives based on the result. The methodology can be used with structured data. (Lee, et al., 2002)

Long and Seko developed a methodology called Canadian Institute for Health and Information methodology or CIHI in the year 2005. The methodology was developed to improve data regarding health information. It tries to find and eliminate heterogeneity in a large database. The methodology supports structured data, but the data in this methodology is regarded to be related to healthcare. Another methodology that was designed for specific data is ISTAT or Italian National Bureau of Census Methodology found by Falorsi et al in 2004. This methodology was made to measure data quality in multiple databases. The goal was to maintain high quality statistical data on Italian citizens and businesses. Both methodologies can be used with structured data, but the environment and data are restricted for the specific need. (Batini, et al., 2009)

Data quality assessment methods in public health information systems have been researched and tested. Based review done by Chen, et al. (2014) the methods that were used are quantitative and qualitative methods. Quantitative methods included descriptive surveys and data audits. Qualitative methods included documentation reviews, interviews

and field observations. Their review found out that data quality in public health information systems has not been given enough attention. Other problems found were that there were no common data quality attributes/dimensions found, data users' issues were not addressed and there was near to none reporting over data quality. This review proved that data quality methodologies need to be further enhanced and companies need to really use them to improve data quality.

Pipino et al. (2002) Created a methodology called Data Quality Assessment (DQA). This was the first methodology to guide with and define data quality metrics. These metrics could be used in multipurpose situations instead of single-issue fixes. This methodology can also be used with structured data. This methodology first suggested the subjective and objective approach in doing data quality assessment. (Pipino, et al., 2002) The subjective and objective approaches in data quality methodologies were later used in Quality Assessment methodology developed by Batini and De Amicis in 2004. This methodology focused on measuring data quality of financial data. Here was defined what is financial data and how the subjective and objective measurement could be done. The down side of this methodology was that it doesn't suggest any improvement suggestions. Both methodologies could be used with structured data and there weren't any environment restrictions. Unfortunately, the original research was not found in the databases or internet. but Batini, et al (2009) and Battini & Scannpieco (2016) have defined how the process works.

Data quality effect on firm performance has been tested in the financial sector in Korea. The research didn't use any methodology to test the effect. It was conducted using regression model. The research showed that Korean commercial banks had high data quality and credit unions low. Also, the results showed that having good data quality improves the revenue of sale and adds operating profit. (Xiang, et al., 2013)

IQM or Information Quality Measurement methodology was created by Eppler and Munzenmaier (2002). This methodology was defined to asses quality issues in web context. The methodology can be used with structured data. Here they developed the methodology based on five different measurement tools used in web context. Here they comment that only continuous information or data quality-measurement can tell if the chosen methods of improvement have worked or not. Out of the five four of the tools are technical tools like web page traffic analyzer or data mining tool. The fifth one that is equally important is user feedback. An example how feedback can be collected is user polls. (Eppler & Muenzenmayer, 2002)

AMEQ or Activity-based-Measuring and Evaluating of Product information Quality was developed by Su and Jin in 2004. This methodology was designed to be used in product information quality. It is especially useful for manufacturing companies. The methodology gave the first data quality tool to use when product information quality is considered. The methodology can be used with structured data. (Su & Jin, 2004)

CDQ or Complete data quality created by Scannapieco and Batini in 2008. This methodology uses existing techniques on data quality assessment and improvement, so it can be used with any type of data. This methodology doesn't assume that contextual information is needed for it to be used. Since the methodology aims for complete data quality assessment and improvement it may be hard to model and evaluate a whole business. The methodology aims to be complete, flexible and simple to use. Completeness is argued to be achieved by using existing technology and knowledge that is used to a framework that can be used in and out of the organization context. The methodology can be used with structured data. Simplicity is achieved since it is explained step by step. Flexibility is because the methodology supports the user in choosing the right tool in each of the steps. (Batini & Scannapieco, 2016)

Batini, et al. (2009) collected and compared all available data quality methodologies that were available at that time. This research was found to be the best source for finding data quality methodologies and understanding what they were meant for. In this research they evaluated their pros and cons based on two aspects: assessment and improvement phases. In total there were 13 methodologies used in their comparison, meaning all the methodologies that were already introduced.

Based on the literature on data quality methodologies I created table 1 which shows how each of the 13 methodology fits the criteria. The criteria are, the methodology has can be used with structured data, the methodology isn't restricted in a specific environment and the methodology isn't made for a specific purpose.

Table 2 Data Quality methodologies with criteria

	Can the methodology be used with structured data?	Is the methodology restricted in a specific information system or system type?	Is the methodology created for a specific purpose?	Can the Methodology be used with the real-life data set in this thesis?
TDQM	YES	NO	NO	YES
DWQ	YES	YES	NO	YES
TIQM	YES	YES	YES	NO
COLDQ	YES	YES	YES	NO
DaQuinCIS	YES	YES	NO	NO
AIMQ	YES	NO	YES	NO
CIHI	YES	NO	YES	NO
ISTAT	YES	NO	YES	NO
DQA	YES	NO	YES	NO
QAFD	YES	NO	YES	YES
IQM	YES	YES	NO	NO
AMEQ	YES	YES	YES	NO
CDQ	YES	NO	YES	NO

Based on table 2 we can say each methodology can be used with structured data. The methodologies that should be used regarded for further analysis have to be found looking at the usage of the methodology and the environment the information system can be used in. Since CIHI was created for medical data, ISTAT for Italian government statistical data and AMEQ for manufacturing industry they can be dropped out immediately. AIMQ was created for the usage in questionnaires it doesn't fit the scope of this work. DaQuinCIS and IQM are meant for different information systems than data warehouse they are dropped out. TIQM and COLDQ are more of measuring costs related to data quality so they are dropped out. Even though CDQ sounds good on paper the actual usage of the methodology to completely analyze a company's data quality doesn't fit into this scope. We are left with four methodologies from which DQA is dropped out since it is more of creating metrics than measuring the quality and improving it. The methodologies that will be discussed further are: TDQM, DWQ and QAFD. The process of elimination can be seen in figure 11.

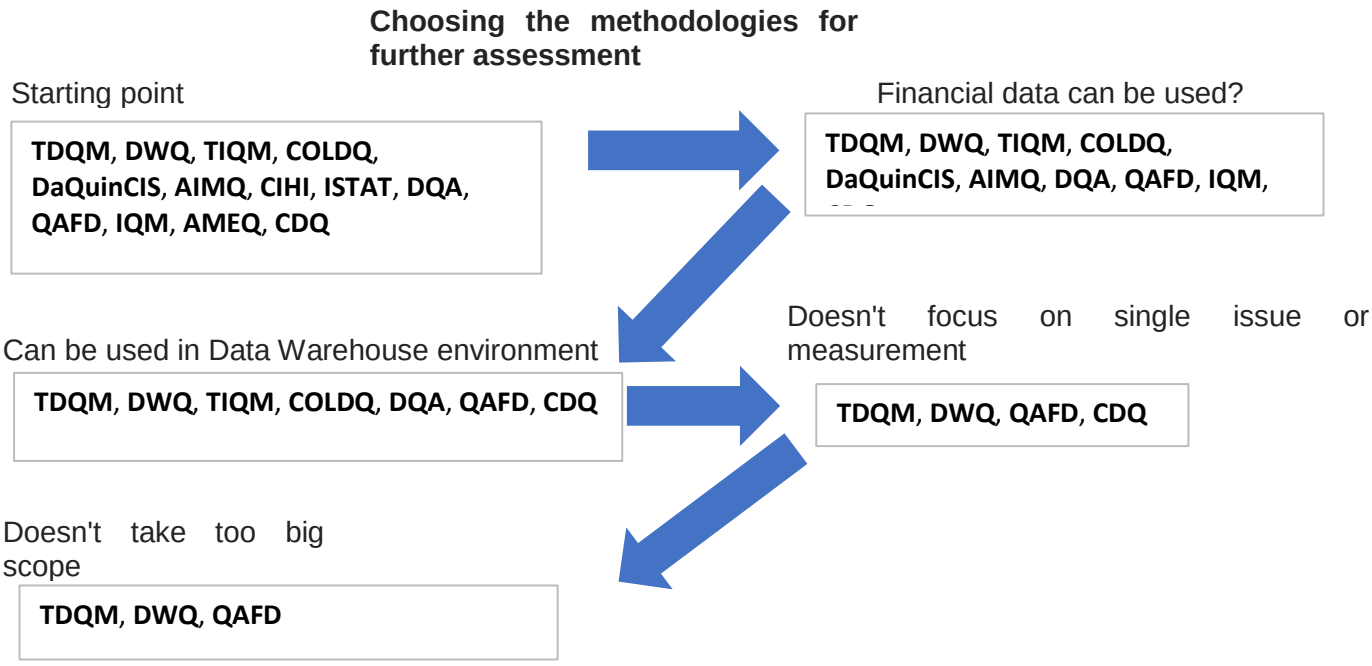


Figure 11 Process of choosing the best methodologies

Figure 11 shows how we end with the chosen three methodologies. If the name of the methodology can't be seen under the assumption it means it has been eliminated. The chosen three methodologies fit the scope of this thesis since they can be used with structured data, they can be used with financial data and they can be used in a data warehousing environment. These three also methodologies also don't focus on single issues and don't make the scope too big.

3.2 Data Quality Methodologies

Batini, et al (2009) have listed various data quality methodologies and created a graph of them. This list can be seen in table 3. Here are listed all the methodologies that were looked at in the literature review regarding data quality.

Table 3 List of different methodologies assessing data quality (Batini, et al., 2009)

Methodology Acronym	Extended Name	Main Reference
TDQM	Total Data Quality Management	Wang 1998
DWQ	The Datawarehouse Quality Methodology	Jeusfeld et al. 1998
TIQM	Total Information Quality Management	English 1999
AIMQ	A methodology for information quality assessment	Lee et al. 2002
CIHI	Canadian Institute for Health Information methodology	Long and Seko 2005
DQA	Data Quality Assessment	Pipino et al. 2002
IQM	Information Quality Measurement	Eppler and Münzenmaier 2002
ISTAT	ISTAT methodology	Falorsi et al 2003
AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology	Su and Jin 2004
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality	Loshin 2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco et al. 2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis and Batini 2004
CDQ	Comprehensive methodology for Data Quality management	Batini and Scannapieco 2006

In table 3 the chosen three methodologies for further analysis are highlighted. Next, we will look at the three methodologies deeper to find the best of them for usage for the real-life data set. Since all the methodologies can be used with the real-life data set, the focus is on getting a “quality score” on data quality for the financial data. Also, it is important that the methodology looks the problem from different aspects.

3.3 Total Data Quality Management (TDQM)

The TDQM process was originally designed for data warehouse projects on ensuring that quality is kept at a certain or desired level. It was first introduced by Richard Wang in 1998 in his work called A product perspective on total data quality management. (Batini & Scannapieca, 2006). The goal is to create, improve and understand the information product (IP) cycle. IP stands for in this context the output of an information system. If possible TDQM suggests forming an IP-map of the whole process. (Wang, 1998)

In the TDQM methodology information that is formed from data is considered to have a life cycle. The cycle defines also different roles within the cycle: Collectors, Custodians, Information Quality managers and Customers. Collectors are the people in the beginning of the cycle who collect data. Custodians are the ones who maintain and develop the information systems the collected data flows in. Information quality managers are the ones who verify the data and are responsible of the quality management through the whole cycle. Customers are then ones who use the data in their work. (Francisco, et al., 2017) An example of the different roles can be as follows: A regulatory reporter must collect data for

a report he has to send to Bank of Finland. The regulatory reporter is in this case the collector of the data. The Custodian in this process is the person responsible for maintaining, developing and producing the information system for the report to be sent in. The Customer is Bank of Finland since it is the one who reads the report. The Quality manager is the one responsible for the creation, collection and delivery of the report so in this case the reporter's manager.

The methodology consists of four main processes: Define, Measure, Analyze and Improve. In the defining phase you define the information product characteristics and information quality requirements. The IP characteristics mean that you know what information product you are focusing on and making sure the quality of it is good. The IQ requirements are according to Wang found through data quality dimensions and their analysis. Wang suggests using an IT-tool that is meant for that purpose to visually see what dimensions have possible issues. It is important to point out you can't define the IQ requirements unless you have defined the IP characteristics and collected the data. In the defining phase it is also recommended to define the information system where the data comes from to understand the different phases how the data moves. (Wang, 1998)

In the measurement phase, IQ metrics are developed to track the quality. These metrics can be designed based on the data quality dimensions for example accuracy. An example of a metric could be the amount of incorrect city names in a database related to customer information. Another example could be the amount of missing information concerning customer addresses in the same customer information database. The former relates to the completeness dimension. These metrics are so called "basic" metrics. There are also "advanced" metrics which are related to business rules that measure things for example across time. An example would be following the amount of missing data related to customer gender. If there is model that predicts customer actions and one attribute in the model is gender, then if the amount of missing values related to gender changes drastically it will affect the model also in the long run. These metrics can be applied to an existing information system as an add-in or if there is purchased a new system it can be implanted into the system itself. (Wang, 1998)

The analyzation phase is done based on the results of the measurement. In this phase we try to find the original source of the problem. If looking at the previous examples, we could test that if we create a dummy customer and see if the data flows correctly. Another way is to check if the information system works, but the problem is with workers who don't follow

guidelines, or the guide is faulted. The key is to find the source of the problem and understand why it happens. This phase begins the improvement of the data by defining the source. The previous two parts were regarded as assessment phases. (Wang, 1998)

The final phase is improvement. Here the key is to fix the problem to improve the quality. The improvement should be done so that it doesn't happen again. Therefore, the analysis must be done carefully. The improvement must be done to meet the business needs. Suggested methods are forming the Information Manufacturing Analysis Matrix or using a methodology for allocating resources for IQ improvement. (Wang, 1998) The later methodology was created to so because companies have usually limited resources or funds on quality improvement, so the methodology helps on allocated the resources to problems that have a bigger impact on quality. The methodology is based on a mathematical formula which tries to minimize the penalty costs and accuracy costs. (Ballou & Tayi, 1989)

From this process, especially from the defining phase, Wang (1998) suggests that a company has two options how to improve the data quality or information quality. Since the company has defined the information product it can either purchase or produce a whole new information system or use the results to improve and refine the existing information system. It is also noted, that purchasing a new system is better in the long run, since the requirements can be implemented to the system itself, so that it checks for certain IQ before sending the actual data. Whichever path the company chooses it is good to point out that if the requirements of the IQ customer change it might mean that the information system needs to be changed.

The information product map or schema of the whole process is an important part of this methodology since it is important to understand the different points at which could cause the quality problems. Especially if using a dummy customer to test the process you must understand how the data flows and what effects what. The summary of the methodology can be found in figure 12.

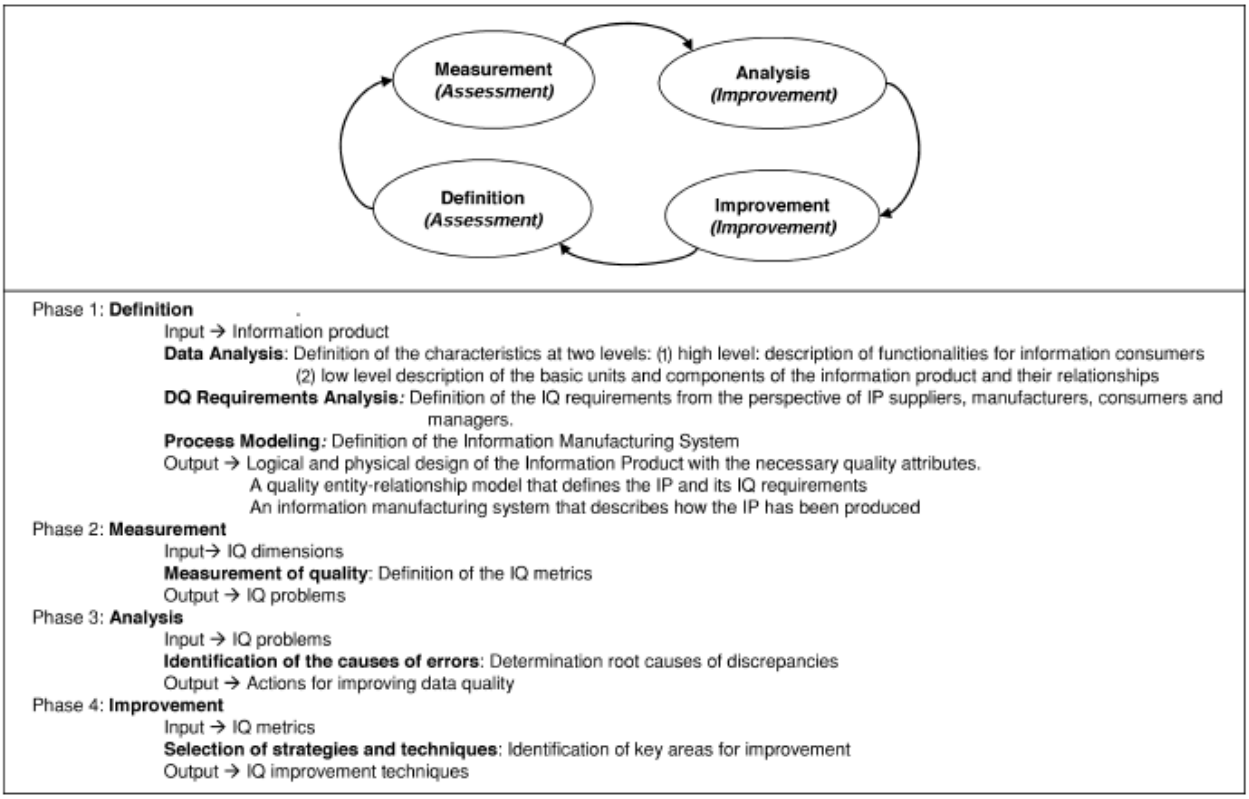
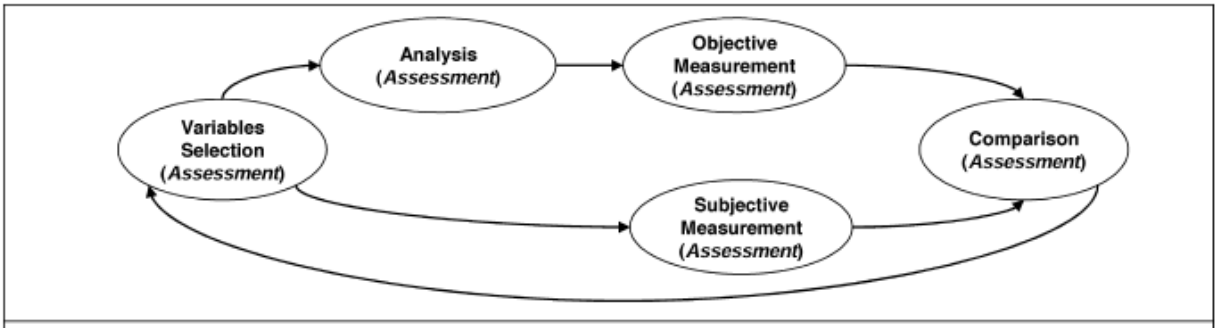


Figure 12 A Summary of the methodology created by Battini et al. (2009)

3.4 Quality Assessment of Financial Data (QAFD)

The quality assessment of financial data is designed to be used with data related to the world of finance. Even though it is a restricted methodology in that sense it is appropriate one to further discuss since the empirical part is data related to the world of finance. The methodology was found by Batini and De Amicis in 2004. It is regarded as an assessment methodology since it tries to provide the current state of the data in the information system. When QAFD methodology is discussed it is more related to information quality than pure data quality since it relies heavily on expert opinions what the information should be like. (Batini & Scannapieco, 2016) Figure 13 shows a summary of QAFD and its steps.



Phase 1: **Variables selection**
- Input → financial variables and the financial context
- **Data analysis**: identification and description of the most relevant financial registry variables, classification of the variables along three classes: qualitative/categorical, quantitative/numerical, and dates
- Output → classification and groups of primary variables

Phase 2: **Analysis**
- Input→ classification and groups of primary variables, data quality analysis techniques, business rules, process analysis, observation
- **DQ Requirements Analysis**: Analysis of data quality dimensions: syntactic and semantic accuracy, internal and external consistency, completeness, currency, timeliness, and uniqueness
- **Identification of critical areas:** Application of data quality rules
- Output → Identification of errors

Phase 3: **Objective measurement**
- Input → Identification of errors
- **Measurement of quality**: Definition of indices for the evaluation and quantification of the global data quality level
- Output → Quantitative objective assessment

Phase 4: **Qualitative subjective measurement**
- Input → Business and data quality expertise, measurement, and classification and groups of primary variables
- **Measurement of quality**: Business expert assessment, Financial operator assessment, Data quality expert assessment, Comparison of subjective assessments
- Output→ Qualitative subjective assessment

Phase 5: **Comparison**
- Input→ Quantitative objective dimensions, qualitative subjective dimensions
- **Measurement of quality**: Calculation of discrepancies
- Output → Assessment and improvement

Figure 13 Summary of QAFD methodology created by Battini et al. (2009)

Unlike TDQM, quality assessment of financial data contains five phases in the methodology: variable selection, analysis, objective measurement, qualitative subjective measurement and comparison. The variable selection phase where financial variables related to same or similar role. The categorizes can be for example time, numerical or character. The second phase contains analysis the financial data is analyzed. This can be done using statistical methods for example mean, min, max standard deviation etc. type analysis. The goal is to understand what might cause the possible errors. The desired result of this phase is a report which contains the identified errors based on used statistical techniques. The report also contains possible dimensions which the errors might be related to. (Batini & Scannapieco, 2016)

Phase three contains objective assessment on the chosen data. Batini and De Amicis (2004) suggest a mathematical formula to calculate quality rankings between the dimensions identified in phase two. The mathematical formula contains the identification of the amount of erroneous data and the information attributes are evaluated. Example could be giving a

score from 1 to 10 in the accuracy dimension for three different attributes that are linked together. After all attributes are rated you move on to the next dimension, say timeliness, and do the same. In the end you sum all the attributes together and divide them by the number of attributes to get a total score of that attribute or variable. The higher the final score the worse the data since the amount of erroneous of the data is being evaluated. (Batini, et al., 2009)

Phase four, subjective assessment, is done by interviewing three different groups to gain different perspectives. The groups are: business experts, customers and data quality experts. Customer refers to in this case the personnel who use the data in their work. Each group is asked to rank the quality of the data based on each dimension found in phase 2. This can be stated: “what do you think is the completeness of this data: poor, mediocre or excellent”. The final verdict is the mean of the answers of the three groups. Even though we are talking about groups here it is enough to get the opinion of one expert, customer and data quality expert, since it might be the company has only one person working in that field. (Batini, et al., 2009)

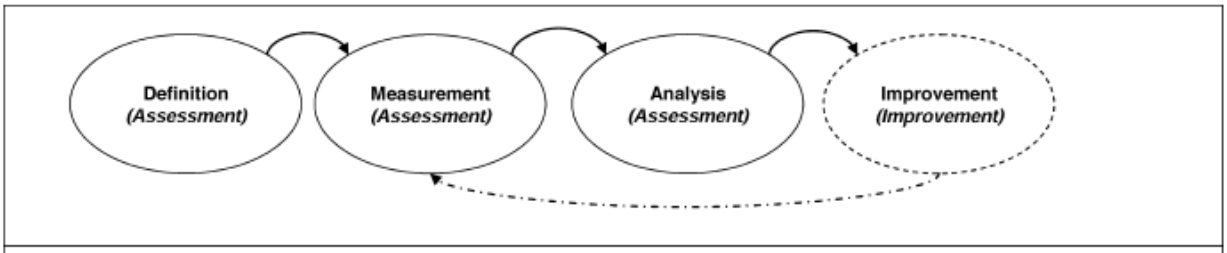
The final phase, comparison, is done between the subjective and objective measurement. This is done by calculating the percentage of erroneous observations in the objective measurement and converting the results to match with the subjective measurement. Then the results are compared and based on them the data quality expert gives possible suggestions or ways on how the data could be improved. The subjective measurement has a bigger weight meaning it is more important than the objective measurement. unfortunately, the methodology doesn’t suggest any concreate suggestions on data improvement. the methodology seems to assume that the data quality expert has the needed tools for the improvement. (Batini & Scannapieco, 2016)

Even though the methodology doesn’t state how the data could be improved a common understanding on what causes data quality issues is usually enough to find the source of the problem at hand. Also, this methodology doesn’t suggest process maps, so it is in a sense lighter process and it can be done by anyone who has understanding on data.

3.5 Data Warehouse Quality (DWQ)

Data warehouse quality methodology was created due to the users not understanding the quality of the data warehouse. The methodology suggests the best place to represent quality

goals is the metadatabase of the data warehouse. The stakeholder set the quality goals and transform them into usable queries to give quality measurement. The query refers to checking if the quality goals are met or if the quality change after improvement. The results are then stored into the metadatabase. Metadatabase refers to information that defines the data in the data warehouse. Metadata can be data information like formats, names etc. The metadatabase contains information on the data warehouse for example links in the data warehouse. Stakeholders in this methodology are the end users of the data that have a need for it. (Jeusfeld, et al., 1998) Figure 14 shows a summary of DWQ.



Phase 1: **Definition**
- Input → Data Warehouse project and context information, operational databases, stakeholders' perspective
- **Data Analysis**: Identification of relevant data quality dimensions in the data warehouse context and identification of the relations between data quality dimensions and data warehouse objects
- **DQ Requirements Analysis**: Definition of the House of quality in which stakeholders assign weights to quality dimensions in order to express their relevance
- Output → List of the quality dimensions and related assessment methods, quality requirements

Phase 2: **Measurement**
- Input→ List of the quality dimensions and related assessment methods
- **Measurement of quality**: Hierarchical Quality Assessment and identification of dependencies among data quality dimensions
- Output → DQ values, dependencies among data quality dimensions

Phase 3: **Analysis**
- Input → DQ values, dependencies among data quality dimensions, DQ Requirements
- **Identification of critical areas**: Comparison between data quality values and data quality requirements
- Output → List of dimensions that need improvement, effects of dependencies

Phase 4: **Improvement (only mentioned)**
- Input → List of dimensions that need improvement, effects of dependencies
- Output → Improved data

Figure 14 A Summary of DWQ methodology created by Battini et al. (2009)

The methodology gives data and software quality dimensions in the data warehouse context. These dimensions are put into three categories: Design and administration -, software implementation - and data usage quality. Design and administration quality refer to two separate things. Design refers to a model and its ability to show information accurately and efficiently. In financial context an example could be a model that predicts customer behavior concerning credit usage. Administration refers to way the model changes during data warehouse operations. Software implementation quality refers to ISO 9126 standard which contain quality dimensions. The ISO 9126 standard contains dimensions like functionality and usability. The data usage quality dimension refers to the queries made to the data in the data warehouse. (Batini, et al., 2009) Jeusfeld, et al. (1998) state that each dimension hierarchy looks different since it is depended on the stakeholder and his preferences. This

means that even when look at the same data quality issues in the data warehouse environment the model might look different due to the preferences of the stakeholder.

Like total data quality management data warehouse quality can be divided into four parts: Definition, measurement, analysis and improvement. The difference here is that unlike in TDQM the analysis part here is still assessing the quality and not starting to improve it. The definition phase contains the finding and defining the relevant dimensions in the data warehouse environment. In this phase the stakeholder's preferences are also considered and based on them weights to the dimensions are assigned. The goal of the definition phase is to gain a set of data quality dimensions each having its own weight based the preferences. (Batini, et al., 2009)

The measurement phase contains the measurement of the data based on the quality dimensions assigned with a hierarchy. The goal here is to find dependencies between the dimensions to find the ones that have the biggest effect on quality. Here are also gained data quality values. These data quality values refer to a score say completeness dimension how many fields are empty of the total. The analysis phase contains the actual comparison between the data quality values regarding the required/desired results. The output here is the list of values or dimensions that need to be improved and an understanding of the dependencies between the dimensions. The final improvement phase has the data improved based on the results in the analysis phase. The improvement phase is only mentioned as further research and not any constructive and concreate suggests on how the results of the analysis phase could be improved. (Batini, et al., 2009)

Since this methodology relies heavily on data warehousing, it requires a deeper understanding of data warehousing process. This is especially crucial on where the quality queries are implemented. This makes the process more technical in that sense it easier for a person working in IT to understand the problems than a person working with the data or owning the data. Also, to use this methodology, you would need the access to the ETL-process.

3.6 Summary of three chosen Methodologies

All the presented methodologies have positive and negative sides. Below is a table that summarizes the three methodologies based on the introduction to this chapter in other words assessment and improvement phase. If the methodology suggests something to that step

or supports, the step it gets an ‘X’ and if not ‘- ‘. Tables 4 and 5 were created based on Batini et al. (2009) research on different data quality methodologies.

Table 4 Assessment phase of all chosen methodologies

	Data Analysis	**IQ/DQ requirement analysis**	**Identification of Critical Areas**	**Process modeling**	**Measurement of Quality**
TDQM	X	-	X	X	X
QAFD	X	X	X	-	X
DWQ	X	X	X	-	X

As we can see from Table 4 there is no method that is better than the other, since TDQM doesn’t suggest anything for IQ/DQ phase and QAFD and DWQ don’t suggest any process modeling. This said based on the assessment phase, QAFD and DWQ would be most suitable for usage in the empirical part, since the focus was not on process modeling. Next is table 5, which contains the difference between the methodologies in the improvement phase.

Table 5 Improvement Phase of all chosen methodologies

	TDQM	**QAFD**	**DWQ**
Evaluation of Costs	X	-	X
Assignment of Process Responsibilities	X	-	-
Assignment of Data Responsibilities	X	-	X
Selection Strategies and Techniques	X	-	X
Identification the Causes of Errors	X	-	X
Process Control	-	-	-
Design of data Improvement solutions	-	-	X

Process Redesign	X	-	-
Improvement management	X	-	X
Improvement Monitoring	X	-	-

As we can see from Table 5 QAFD has all values not suggested. This is since QAFD method doesn't suggest any improvements and is only concentrating on the assessment phase. When comparing TDQM and DWQ the results are that TDQM suggests more improvements based on the improvement phase. The differences come from assignment of process responsibilities, Process redesign, improvement of monitoring and design of data improvement solutions. It is good to point out that none of the chosen methodologies give any suggestions to process control.

Based on the findings and the pros and cons of each methodology I decided to choose QAFD as the methodology to be tested in the empirical part. This is because QAFD gives the most informative measurement of data quality regarding financial data. If QAFD proves to be effective based on the empirical results it clearly shows in what dimensions issues are found and it uses the expert's opinions as well. Also, QAFD is the best methodology at giving your data a score value how good or bad it is. The criteria for the selection are seen in table 6.

Table 6 Empirical part methodology selection

	Does the methodology give the data a "quality score"	**Does the methodology look at different aspects of the data**	**Can the methodology be used with resources given in this thesis?**
TDQM	NO	YES	YES
DWQ	NO	YES	NO
QAFD	YES	YES	YES

Even though TDQM has a better take on improvement compared to QAFD, the methodology itself focuses on schema building to understand quality issues. The schema related

dimensions were not in the focus of this work hence it is not appropriate to use the following methodology. Also, TDQM focuses on the big picture rather than just digging deeper into the issue at hand. DWQ was not chosen since it is more technical, and the testing of the methodology would require creating the queries in warehousing process which is not possible in this thesis. This would have meant that the empirical part would have been more of a written suggestion how it should be done rather than using it and finding the results.

3.7 How the QAFD methodology process works

Since QAFD was chosen as the best alternative for the analysis we will now look in-depth how it works. Based on this in-depth look of the methodology we can then use it in the empirical part. From figure 13 we can see the summary of the methodology and we will go phase by phase through the methodology.

Phase one of the whole methodology process is choosing the variables. The variables must be related to each other in financial context meaning that the variables can't be taken one attribute is taken Corporate rating and the other attribute is Private customer loan margin. A correct context is defined to be having the same risk, business and descriptive factors. The methodology calls the different attributes that are objectively and subjectively measured as variables. (Batini & Scannapieco, 2016)

In phase one it is also highlighted, that the variables should be the most relevant financial variables. Unfortunately, it is not clearly defined how the most relevant variables are selected. Batini, et al (2009) suggest that the variables should be based on previous assessment. This means that previous knowledge of the chosen variables and their quality can be used as a basis of the decision. After identifying the data, the data should be categorized into data types. The methodology states three data types:

- Qualitative / Categorical
- Quantitative / Numeric
- Date / Time

Based on this the data should be now identified what data is looked at, what data types do the variables represent and the context of the financial data itself.

Phase two of the process contains the analysis of the variables. Here is also chosen the dimensions to be used for the objective and subjective measurement. Based on the analysis there is also created business rules or consistency constraints. The aim here is to find

possible causes of the error. Simple statistical inspection is suggested to look at the data, to understand what it is and what would be good dimensions to be looked at. The chosen dimensions are highly related to your data and the data types. This means that it is not logical to choose timeliness dimension if you don't have data related to timeliness. Another example could be choosing currency dimensions, which relates to value of the data if the data doesn't have so called "money value" or "time value". (Batini & Scannapieco, 2016) The dimensions seen in figure 13 were explained in chapter 2.2.

After looking analyzing the variables you should find out the dimensions that could be useful for further analysis. You should have created a set of dimensions to be looked at and set business rules to measure the consistency of your data. The business rules can be designed based on previous knowledge or business logic. Another way is to test the logic if the statistical analysis finds irregularities in the data. Having the dimensions selected you should be now able to guess where the possible errors have come from. This means that you can say that the errors might become from bad information from the source system or there might be a problem in the ETL-process. The results should be collected in a so-called data quality "report" that defines the possible error locations and dimensions to be analyzed. (Batini & Scannapieco, 2016)

In phase 3 is the objective assessment. It means analyzing the data based on the chosen dimensions and finding how much of the data is erroneous. The result is summed together and the data gains objective measurement score for each variable based on each quality dimension. The mathematical formula itself that was mentioned in the short introduction of this methodology isn't shown in any research paper, however Batini and Scannapeico (2016) have shown with examples how the objective measurement can be done. The example result can be seen in table 7.

Table 7 Example of objective measurement (Batini & Scannapieco, 2016)

	Variables		
Quality dimensions	Moody's Rating	Standard's & Poor Rating	Market Currency Code
Syntactic Accuracy	1.7	1.5	2.1
Semantic Accuracy	0	0.1	1.4
Internal Consistency	2.7	3.2	1.3
External Consistency	1.6	1.1	0.1
Incompleteness	3.5	5.5	8.1
Currency	0	0	0
Timeliness	8.6	9.2	2
Uniqueness	4.9	4.9	9.3
Total (average)	3.6	3.2	3.0

In table 7 the chosen attributes are rating codes for companies. Each attribute is analyzed based on each dimension and a score is given. The figure represents the amount of erroneous data based on each dimension i.e. the higher the score the worse the data. Each dimension is calculated based on the dimension i.e. each dimension can be calculated in different ways. The ways how to calculate the dimensions, if given, are described in this thesis in chapter 2.2. After getting a score for each dimension the dimension scores are normalized to the same scale. The normalization can be done immediately after calculating the score. Normalization means for example that results are on the scale of 0-10. After this the results are summed together and divided by the number of dimensions to get the final score. (Batini & Scannapieco, 2016)

To summarize phase 3, you first must calculate a score for each variable in each dimension. The calculation formula for each dimension is given in data quality literature. You also define a scale for the normalization phase so that each result is on the same scale. Finally, the results are summed together and divided by the number of dimensions used. This gives the total score for each individual attribute. The higher the final score the worse the quality of your data.

Phase 4 is related to the subjective assessment. Here the assessment for each dimension is done by experts. The expert needed for the assessment to be made are the following: Business expert, financial operator and IQ / DQ expert. The business expert is defined as to a person who uses and analyzes the information from business process point of view. The financial operator is regarded to be person who uses daily financial information and who

works hands on with the data. The Information or data quality expert is regarded to be a person who has access to the data and analyzes the quality of the data. (Batini & Scannapieco, 2016) Batini, et al. (2009) listed that the experts were: business expert, customer and data quality expert. Even though the expert names have changed the meaning or description of the role is the same.

Now the experts individually answer to each attribute based on the dimension and how they see the data quality. The metrics how they give their opinion are given. It is suggested by the methodology that they give a written assessment based on a scale given to them. The scale can be good, mediocre and poor or similar scaling. An example of what the results look like after the assessment can be seen in table 8.

Table 8 Example of subjective assessment

	Rating Moody's	Rating S&P	Market Currency Code
Syntactic Accuracy	H	H	H
Semantic Accuracy	H	H	M
Internal Consistency	H	H	H
External Consistency	H	H	M
Incompleteness	L	L	L
Currency	H	H	H
Timeliness	M	M	H
Uniqueness	H	H	H
Total	H	H	H

From table 8 we can see how the final domain values are formed. The total result is based on the average of the answers. Already in this phase the domain values can be converted to same scale numeric scale as the objective assessment.

The final phase of the QAFD process contains a comparison between the subjective and objective assessment. Here is calculated the difference between the subjective and objective assessment. Meaning how much do the results differ between the two assessment. The calculation is done by subtracting the objective assessment score from

the subjective assessment score. If the difference is positive the objective assessment is overruled by the subjective measurement. In general, this means that the objective assessment shows issues that the experts do not regard as having big effect on their work and the quality of the data. The expert opinion for quality question in each dimension has a bigger value. If the difference is negative the assessment results agree with each other. (Batini, et al., 2009)

As already mentioned, the methodology doesn’t give improvement suggestions, but these can be made based on the environment the data is in and the issue at hand. An example could be that there is an issue in the consistency dimension. The environment is the key in this issue. If it is a data warehousing environment, then it might happen due to bad ETL or source itself is invalid.

4 CASE: Data quality assessment using QAFD

This chapter discuss how the chosen method is used with the real-life data set. This chapter begins with introducing the chosen data set and continuing with the step by step process of using the QAFD methodology on the data set. The chosen dimensions to be used are: accuracy, consistency and completeness as already chosen in chapter 2.

4.1 Data set

The data set contains three columns which are named A1, B1 and C1. The data is financial data related to financial accounts. Each column consists of 10000 attribute values so the whole data set contains 30000 attribute values. Attribute A1 is character format and has attribute values A and S which describe the account. Attributes B1 and C1 are numeric and include values based on attribute values in A1. Both B1 and C1 can contain positive and negative values. The assumption is that there are no missing values and constraints are set to test the data. The data is masked so that the real context where the data is taken from isn't affected. The labels used to describe the data were removed during data masking process. In table 9 we can see the first 9 rows of data in the data set.

Table 9 First 9 rows of the data set

A1	B1	C1
A	-854284,58	0,00
A	0,00	0,00
S	0,00	0,00
A	0,00	0,00
A	-569,30	-0,03
A	-877,94	0,00
A	-877,94	0,00
A	-847,46	0,00
A	-847,46	0,00

4.2 QAFD illustration using the data set

As stated in phase one of the QAFD methodology the data must have a connection for the methodology to be used. The chosen data for data quality testing has similar role and is related to each other. The data was chosen based on previous knowledge that there might be some issues regarding the data. No previous assessment was made on the data prior to this assessment. The variables are categorized in the following way: A1 is categorical and B1 and C2 are numerical. There are no variables related to date/time. Since the data is

related to each other there can be set consistency constraints on it. A1 should only contain values A or S and if A1 is S then attributes in B2 and C3 on that same row should both be zero i.e. 0.

Phase two of the methodology stated at giving basic statistics on the data you are analyzing to understand what it contains. Also, here we aim to understand what dimensions might be related with the erroneous data. Since we already defined restrictions and believe the data set is complete, we must test if these dimensions hold. Tables 10 and 11 shows basic statistics that have been gained using SAS Enterprise Guide (EG) characterize data. All values are shown with 1 decimal accuracy.

Table 10 Basic statistics of Variable A1

Variable	Label	Value	Frequency Count	Percent of Total Frequency
A1		A	7309	73,1
		S	2691	26,9

Table 11 Basic statistics of Variables B1 and C1

Variable	N	NMiss	Total Sum	Min	Mean	Median	Max	Std Dev
B1	10000	0	6529680278	-18724944954	-652968	-32,4	22324033554	305626771,5
C1	10000	0	4270,9	-43,5	0,43	0	3114,1	32.6

Table 10 is separate from table 11 since the values in it are character and not numeric as it was defined in phase one. From table 4 we can see that roughly 73 % of the data has a value of A and 27% has the value of S. None of the values have labels attached to them. Here would read a description of the if it would have been added. In table 11. we can see that there are 10000 attribute values in variable B1 and C1 as seen in column N. there are no missing values which is indicated by NMiss. Total sum adds up all the attribute values. Based on the total sum, min, max and mean we can see that variable B1 has much bigger

attributes than C1. B1 has also a bigger standard deviation compared to C1 as seen in standard deviation.

Since A1 has two unique values A and S we can drill the statistics further by using SAS EG summary statistics. Here we have separated the basic statistics based on if the attribute value in A1 is A or S. This also helps on checking if the consistency restriction holds when A1 has the value S. This can be seen in table 12.

Table 12 Basic statistics when looking at variables B1 and C1 with variable A1 attribute equal to A or S

A1	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	N
A	7309	B1 C1	-889873,80,6	35751289538,2	-18724944954 -43,5	22324033554 3114,1	7309 7309
S	2691	B1 C1	-9510,4 -0.009	306875,5 0.08	- 14715636,6 -2.2	500019,4 0	2691 2691

In table 12 we can see how the population is divided based on the attribute value of A1. We can see that A has a population that is over two times as big as the population of S. We can also see that values in C1 are way smaller than B1 as seen from minimum and maximum and the mean. Also, the mean is negative which indicates there are more values that are <0 than values that are > 0.

Based on table 12 we can also see that the consistency dimension on the restriction does not hold and we are interested of the total 2691 S attributes how many of them does meet the restriction criteria. Table 13 shows this where N means it doesn't meet the criteria and Y means that it meets the criteria. The table was created first by filtering only the values that have value S and then using a SAS EG advanced computed column builder with the following code:

```
CASE
When t1.A1 = 'S' and t1.B1 = 0 and t1.C1 = 0
then 'Y'
else 'N'
END
```

Table 13 Statistics on the consistency dimension

Restriction	Frequency	Percent
N	446	16,6
Y	2245	83,4

Since the desired results of phase to was to understand possible dimensions related to the data and create a report on the data error, we can conclude here that consistency dimension is heavily affected by this error. We choose dimensions Completeness, Accuracy and Consistency. Here Accuracy is regarded as semantic accuracy since syntactic accuracy errors can't be found. This means that A1 all attribute values are character and A or S and B1 and C1 all values are numeric and positive or negative. Consistency is related to how consistent the data is regarding the business rules set on the data. A1 should have only values A or s and B1 and C1 should be 0 if A1 is S. The other dimensions seen in figure 13 are left out. The data doesn't have time related data, so timeliness is out of the scope. We don't have unique keys or attribute values that or assumed to be unique, so it is left out. Currency is left out because we don't have time related data in the data set so there can't be "time value". The data hasn't been defined with a money value.

The possible errors in this case can happen from bad ETL-process or the erroneous data can come directly from the source. With these findings we can go forward to phase three which gives the data an objective score based on statistical measurement.

4.3 Objective measurement on the data set using QAFD

The objective measurement ranking is done based on the basic statistics and the fulfillment of the chosen three dimensions. The final score on each variable is given based on the average score gained from each dimension. The scores on how quality points are gained in the objective and subjective measurement are seen in table 14.

Table 14 Scales on how the scores are given in the objective and subjective measurement

Scale	Points
0-10%	10
10-20%	9
20-30%	8
30-40%	7
40-50%	6
50-60%	5
60-70%	4
70-80%	3
80-90%	2
90-100%	1
100%	0

Expert opinion	Points
Poor	10
Mediocre	5
Excellent	0

In table 14 the left-side scale is used for the objective measurement and the left-side scale is used in the subjective measurement when converting the verbal opinion into the same scale as the objective measurement. Remember that we are measuring the amount of erroneous data which means the higher the error the higher the score. If the data has no error, it is good, and it gets less points. As stated in the methodology the higher the final score the worse the data is.

We first start with completeness and it is divided into column and population completeness. As already explained in chapter 2, column accuracy is calculated based on the number of rows populated in that column. Population completeness looks at the whole population of the data set, how many of the individual attributes are populated in the data set. Here we know the maximum number of rows for each column is 10000, so if all the individual attributes are populated the completeness would be 100% which is zero points in the erroneous measurement. We can use the basic statistics here and see easily from table 11 that Nmiss is zero in both B1 and C1. This means that they have 100% completeness. A1 can be calculated by summing the frequency count of both Values A and S i.e. 7309 + 2691 which can be found in table 10. This gives a total of 10000 which means also A1 has completeness 100%. Based on this and if summed together we can say that the whole population which is 30000 is also 100% complete. The results can be seen in table 17.

Next up we have accuracy dimension. Here we want to know how many of the total 30000 individual values are correctly populated. Since we know what is incorrect based on the

consistency dimensions, we can create table 15 which shows how all the values follow the accuracy dimension. We use once again advanced computed columns in SAS EG on our data set to get the desired results.

Table 15 Accuracy Dimension on each variable in the data set

A1 Accuracy	Frequency	Percent
Y	10000	100

B1 Accuracy	Frequency	Percent
N	442	4,4
Y	9558	95,6

C1 Accuracy	Frequency	Percent
N	148	1,5
Y	9852	98,5

From table 15 we can see that A1 is the only variable with 100 % accuracy. It can be proved already in table 10 since it only contains the values A or S. B1 has a total accuracy of 95,6 % with 4,4 % of the values being inaccurate stated by N. C1 has a better accuracy than B1 with a total accuracy of 98,5% with 1,5% being inaccurate. This information is added to table 17.

The final dimension chosen for the analysis was consistency and we defined two rules that should apply in this data. The first one was that A1 should have only values A or S. This one holds as we can see in table 10. The other rule was that if A1 had the value S then B1 and C1 should have value 0 on that row. Based on table 13 we can see that the number of incorrect rows where the consistency restriction doesn't apply is 446. The total amount of where it should apply is 2691 which can be seen in table 12. We can calculate 100 – (446 / 2691 * 100) = 83,43%. This means 83,43% is consistent with the restriction. Now that we know the total population we want to dig deeper and find out is there a difference in B1 and C1 meaning does the other column only contain the erroneous values. Based on Table 12 it seems that both columns do not follow the consistency dimension. We build table 16 to see how C1 and B1 follow the dimension.

Table 16 Variable B1 and C1 Consistency dimension

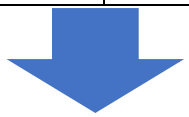
B1 Restriction	Frequency	Percent
N	442	16,4
Y	2249	83,6

C1 Restriction	Frequency	Percent
N	148	5,5
Y	2543	94,5

From table 16 we can see that B1 doesn't follow the restriction in 442 values and C1 doesn't follow it in 148 values. This means that variable C1 gets a consistency percentage of 83,6% and C1 gets a 94,5% consistency percent. The values are added to table 17 which is seen below.

Table 17 Results of objective measurement

	A1	B1	C1
Completeness	100 %	100 %	100 %
Accuracy	100 %	95,60 %	98,50 %
Consistency	100 %	83,60 %	94,50 %



	A1	B1	C1
Completeness	0	0	0
Accuracy	0	1	1
Consistency	0	2	1
Average score	0	1	0,7

As already stated previously table 17 uses the scale introduced in table 14. In the end the average score for each variable was created. The average score is the one to be used in the final phase of QAFD when comparing with the subjective measurement. We can see from table 17 that B1 has the highest average score. A1 doesn't have any problems in the chosen dimension and it can be regarded as having total quality in the data set. We shall find out next if the subjective measurement is like the findings of the objective measurement.

4.4 Subjective measurement on the data set using QAFD

Phase four of the methodology is done by using expert statements on the data. Here the expert is asked to evaluate the data based on each dimension and give it a rank of Poor, Mediocre or Excellent. After this the verbal score is converted into numbers based on the scale introduced in 4.3. The experts used are a data analyst (data quality expert), a business intelligence analyst (customer) and representative of banking business (business expert). To make sure the experts understand the data they are shown the unmasked data. This makes sure that the expert opinion is reliable. Also, to not make the experts bias they aren't shown the results of the objective analysis.

The first step of phase four is making sure the experts understand what is meant by each dimension. This is to make sure they know what perspective they are measuring when estimating the quality of the data. Here we also asked for comments for their answers if they wanted to comment on their assessment. The comments on the business intelligence analyst and data analyst are seen in table 18. The comments have been translated from Finnish to English. These comments were not written as part of the QAFD process, but it gives extra value for the results.

Table 18 Experts comments on their subjective measurement

	Completeness dimension	
	Finnish	English
BI Analyst	Datassa ei ole puuttuvia arvoja, joten siksi paras arvosana	Data doesn't have missing values, so that's why it gets the best grade
Data Analyst	A1 muuttujan osalta data on ollut melko laadukasta ja vastannut odotettua	Concerning A1 variable, the data has pretty good quality and has represented what it is supposed to represent
Banking expert	-	-

Accuracy dimension

	Finnish	English
BI Analyst	A1 pitää sisällään vain sallittuja arvoja. B1 Osalta kentän arvot ovat järkeviä, mutta C1 osalta datassa on muutamia outliereitä, joten siksi sen laatu sai arvosanan mediocre.	A1 contains only values that are allowed. Concerning B1, the values in them seem reasonable, but concerning C1 in the data it has a couple of outliers so that is why it only gets grade mediocre
Data Analyst	Sama vastaus kuin johdonmukaisuus dimensiossa, koska ehtolausekkeet vaikuttavat myös datan tarkkuuteen.	Same answer as in Consistency dimension, because the restrictions effect on the accuracy of the data
Banking expert	-	-

Consistency dimension

	Finnish	English
BI Analyst	A1-kentän osalta datassa on vain kahta arvoa, eli se on ok. B1 ja C1 kentissä on kuitenkin arvoja silloinkin, kun ei saisi olla. B1 pitää sisällään enemmän virheitä, joten siksi se saa arvosanan poor.	A1 field has only two values, so it is ok. B1 and C1 fields have values in them when they should not have. B1 has more errors so that it why it gets the grade poor.
Data Analyst	B1 ja C1 osalta datassa on tiedossa haasteita, jotka ovat tunnistettu ja niihin ollaan työstämässä korjaus toimenpiteitä, jotta haluttu eheys varmistetaan.	Concerning B1 and C1, the data has challenges that have been located. There is an ongoing working progress on fixing the quality issues so that the desired consistency is achieved.
Banking expert	-	-

The business expert gave no comments on their assessment. Table 19 shows the subjective measurement results from all three experts.

Table 19 Subjective results of three business experts

Business intelligence Analyst			
	A1	B1	C1
Completeness	Excellent	Excellent	Excellent
Accuracy	Excellent	Excellent	Mediocre
Consistency	Excellent	Poor	Mediocre

Grade conversion to score		
A1	B1	C1
0	0	0
0	0	5
0	10	5

Data Analyst			
	A1	B1	C1
Completeness	Excellent	Excellent	Excellent
Accuracy	Excellent	Mediocre	Mediocre
Consistency	Excellent	Mediocre	Mediocre

Grade conversion to score		
A1	B1	C1
0	0	0
0	5	5
0	5	5

Business Expert			
	A1	B1	C1
Completeness	Excellent	Excellent	Excellent
Accuracy	Excellent	Mediocre	Mediocre
Consistency	Excellent	Mediocre	Mediocre

Grade conversion to score		
A1	B1	C1
0	0	0
0	5	5
0	5	5

As the methodology suggested, now we must convert the expert analysis to a single value. This is done by summing the values from each cell and dividing by three sense there are three experts i.e. A1 completeness (0+0+0) /3 = 0. The final grades are shown in table 20.

Table 20 The final score on the subjective measurement

	Final Score		
	A1	B1	C1
Completeness	0	0	0
Accuracy	0	3,3	5
Consistency	0	6,7	5
Average Score	0	3,3	3,3

In the bottom of table 20 we have the average score for each value from the subjective assessment phase. A1 having an average of 0, B1 having 3,3 and C1 3,3 average score.

4.5 Comparison and improvement

Since QAFD didn't give any concreate suggestion on how the data could be improved it is done by the knowledge gained in chapter 2. Like QAFD suggests the expert opinion given in the subjective measurement is more important than the objective measurement. This is due to the knowledge and information quality factors. In table 21 we can see the results of both objective and subjective measurements.

Table 21 Results of subjective and objective measurement

	Subjective assessment			Objective Assessment		
	A1	B1	C1	A1	B1	C1
Completeness	0	0	0	0	0	0
Accuracy	0	3,3	5	0	1	1
Consistency	0	6,7	5	0	2	1
Average Score	0	3,3	3,3	0	1	0,7

From table 21 we can see that the objective measurement shows that the biggest issue is in variable B1. Subjective measurement supports the findings of the objective assessment where consistency dimension of value B1 is the worst of the dataset. We can say based on

the results that the experts now their data. This might be one of the reasons why QAFD supports the expert opinion and its usage. From table 22 we see the differences between the objective and subjective measurement calculated.

Table 22 Differences between the objective and subjective measurement

	Measurement differences		
	A1	B1	C1
Completeness	0	0	0
Accuracy	0	-2,3	-4
Consistency	0	-4,7	-4
Average Score	0	-2,3	-2,6

Table 22 shows that all the differences on the results are negative meaning that the subjective analysis doesn't have to overrule the objective measurement. The results on both measurements agree with each other. There is a problem in the accuracy and consistency dimension for attributes B1 and C1.

The following improvement suggestions can be made based on the results. First the same analysis must be done on the full data set to see if there is an even bigger issue. Then the location of the issue must be located. Recall that the most common places of the data quality issues are in the core system or the ETL process in a data warehousing environment. The problem could rise do to the lack of proper extraction logic or missing data quality checks. Since we have the consistency rule it could be implemented into the ETL process to check if the extracted data follows the restriction or not. Once this is done, the source of the problem can be easily crossed to either ETL or source system. If the problem is in the source system, then the process owner should be in contact with the data supplier and discuss the possibilities of fixing the issue. The other option here is to add into the ETL process a data manipulation step which follows the consistency rule.

The other problem source of the data quality issue would be the ETL. If the location was there it should be located at which point of the process the problem happens i.e. extractions phase, transformation phase, or load phase. Most likely place the issues happen is in the

transformation or load phase. Since it is not a key issue it can be narrowed down to transformation. This can be easily checked does at any point of the process the fields B1 or C1 go through any transformations. If it was located there, then the transformation part of B1 and C1 must be fixed.

4.6 Research Discussion

Based on the findings in chapter 4 we can say that there is a problem in column B1 and C1. To find the core reason and understand why gets attribute values that are not correct we would have to create a schema to understand the whole process. By building the schema we could determine precisely at which point of the data lifecycle it fails. As suggested in section 4.5 the most likely place for this error is straight at the source.

It is also worthy to note that the dataset seems to have more negative values than positive values in columns C1 and B1. This wasn't in the scope of the analysis but could use some further investigation to understand if the data is correct also in that perspective. That could be done by adding a new consistency rule, but usually you need to know the data to make logical rule base for it. The business intelligence analyst commented also, on the outliers on that field: "C1 in the data it has a couple of outliers so that is why it only gets grade mediocre". (table 18) These outliers should be further investigated so that we can understand why they are there or is there an issue.

Based on the comments and subjective measurement on each dimension we can say that the experts have a good understanding on the data. This supported by the objective measurement results. Also, the experts gave similar assessments. There was a consensus regarding column A1 that the data represented what it is supposed to represent, and it was complete. Both analysts commented on the issues regarding columns B1 and C1. These issues were also located with the objective measurement. As the data analyst commented: "Concerning B1 and C1, the data has challenges that have been located and there is working progress on fixing the quality issues so that the desired consistency is achieved." (table 18) In table 23 we can see the summary of the objective, subjective assessments and the distance between the results.

Table 23 Summary of the results of the empirical section

	Subjective assessment			Objective Assessment			Measurement differences		
	A1	B1	C1	A1	B1	C1	A1	B1	C1
Completeness	0	0	0	0	0	0	0	0	0
Accuracy	0	3,3	5	0	1	1	0	-2,3	-4
Consistency	0	6,7	5	0	2	1	0	-4,7	-4
Average Score	0	3,3	3,3	0	1	0,7	0	-2,3	-2,6

From table 23 we can summarize that the subjective measurement agrees with the objective assessment. The differences are negative, so the assessments agree with each other. If there would have been points in the A1 column in the subjective measurement, then subjective measurement would have overruled the objective measurement. In this case, subjective measurement does not have to be used to overrule the objective measurement.

The findings show that there is real world need for data quality assessment. If nobody investigates data quality issues it means that the quality is most likely bad or not near desired level. This adds additional costs to the company due to fixing costs and working hours when employees located the issue and its source. Worst case scenario would be that the company makes business decision based on poor data. The best and most efficient way in my opinion is to implement quality checks into the ETL-process by using some tool meant for it or user written code. The user written code isn't encouraged in the ETL environment because it usually means that when the person leaves nobody understands what he or she has really created. It also makes upgrades and further development work much more difficult.

Data quality is a real thing and it needs to be understood and assessed. These quality checks can be so called business rules, which are created by data quality experts. A tool sold on the market for managing data quality is for example SAS Data Management Studio. This software allows the user to create business rules, profile data and set the quality checks part of the ETL-process. (SAS, 2019) Even though it costs more to use a ready built tool for quality purposes it becomes cheaper in the long run since user written code is usually user dependent.

5 Summary and conclusions

Finally, we look at the results of the illustration using the QAFD methodology. We discuss what the findings tell us about the data and mirror the results to the real world need of data quality. After the discussion we answer to the research questions and reflect on this work what steps should be followed when choosing a data quality methodology for usage. We also, suggest further research on data quality and reflect on the master thesis writing process and this work in general.

5.1 Research results

The goal of my thesis was to find successful methodology for the data set and give improvement suggestions based on the results of the measurement. The key idea was to understand what you need to know so that you can choose a methodology for usage. The chosen methodology was quality assessment on financial data. This was due to the fact it was more of a micro analysis which drilled into the data set and looked at it from different dimensions. This was not an analysis that tried to focus on the big picture. This helped understand what possible data quality problems could be in the data set. The methodology also had the subjective measurement which asked the expert opinions. This was in my opinion the most important part of the assessment since the experts should now best if there are data problems. The objective measurement is only there to prove their point or possible show new problems if bias happens. Also, the fact we had financial data and the methodology was meant to be used on it gave the final decision for it to be used. The only downside was the fact that it doesn't give any suggestions on improvement, but that it didn't matter since this thesis gave the basic idea of data quality issues and their improvement.

This thesis had one main research question which was supported by three sub questions. Next will answer them one by one starting with the sub questions:

What methods can be used to measure data quality?

There have been developed many different methodologies for data quality measurement all of them can be seen in table 3. The important thing to understand about the method what is it meant for. This means you should ask yourself what your data is and what do you want to measure. To answer this question, you need to get acquainted with the methodologies and based on that you find the correct one for you. This is since some are meant for schema building, survey quality analysis, special purposes etc. If the purpose of your measurement is to measure survey quality the methodology that suits you might be AIMQ. If you want to

understand the total process where the data comes from and goes to and want to create the schema in the process of it then TDQM might suit you best. The method itself is suggested by the methodology and it can be from statistical analysis, expert opinion analysis to more complex formulas or building schemas.

What do you need to know about your data to use data quality methodologies efficiently?

This question is already slightly answered in the previous sub question. Before you can choose the methodology, you need to know your data. This means you must define what is the source of your data. Here is important to understand what information system environment you are working with. What type is your data, numeric, character, date/time. The structure of your data is important as well since some methodologies aren't usable with some structures. Also, you need to understand the purposes of different data quality dimensions. This helps you at the analysis when you know what perspectives are especially important in the chosen methodology.

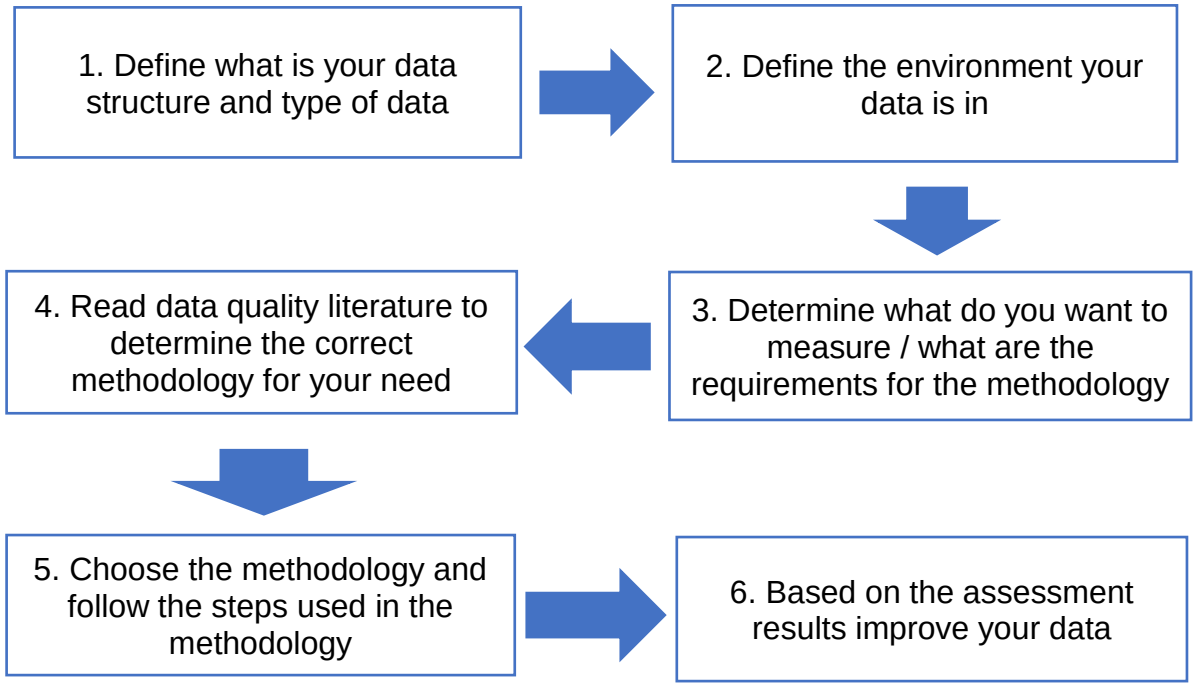
How can the findings from using the methodology help on data improvement?

The findings and results of your analysis can help you directly or indirectly on data quality improvement. Direct means in this case that the methodology itself suggest improvement strategies or methods. Indirect means that the methodology doesn't give any suggestions, but it gives results that can be used to understand where the issues is. Here the expert knowledge is put to test can he or she narrow the possible sources of the problem to the core issue. Understanding what data quality issues are the most common and where can they happen gives a person enough understanding to trace the problem based on the issue at hand. There are basically 4 places where data quality issues can happen in a data warehousing environment: The source, data integration and profiling, the ETL process or database model issues. If the issues are related in the schema the issue most likely is in database model. If the issue is already in the data warehouse it is either in the ETL or the source which is causing the issue. The data profiling and integration is the source most likely when brining new data to the data warehouse environment. In the case of the data set used in this thesis the problem was narrowed down to ETL or source system. From which the most likely reason was the source due to the nature of the problem.

How to successfully use a methodology to assess data quality and improve it based on assessment results?

My main research question is answered by following my sub question. But here to summarize, you need to understand and define your data, define the environment you want to use a methodology in and have a common sense of problems that might be possible in that environment. Once you know these things you can look at data quality methodologies and choose the one that you see fit for use. The methodology itself can give suggestions or further methods how to analyze and improve your data based on the results, but if not, then the common understanding is usually enough.

Reflecting on the empirical part we first defined that we are working looking at data in a data warehouse environment and the data is financial data. We further defined that the data is historical and registry data. After this we could look at the different methodologies and choose few possible options for further research. Next, we looked at the dimensions and chose the once that seemed to be logical for usage. Here the key was to understand that some dimensions were related to schemas and time, which weren't related to this thesis or the data. Then by understanding what the methodology is really meant for and how it is used we could make the final decision to test the quality with QAFD. Based on QAFD results I made suggestions on where the issue might be and how it could be improved or fixed. Since QAFD doesn't give any suggestions in the methodology itself we went through common issues and improvement suggestions in data warehouse environment. Figure 15 shows the steps found in choosing and using a data quality methodology.



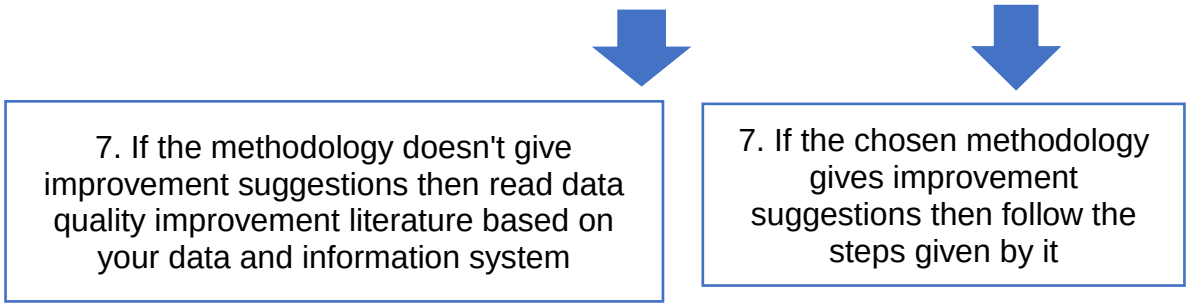


Figure 15 Steps for choosing and using a data quality methodology

If you follow the steps shown in figure 15 you can successfully use a data quality methodology and improve it. Steps 1-4 can be done in any order, but it is important to collect your needs so that you can clearly understand what you want from the data quality methodology. Steps 7 are divided into two since some methodologies, like the one used in this research, don't give improvement suggestions. That is why you need to understand the information system your data is stored in.

5.2 Further research and critical thinking

Based on the results of my thesis there could be done some further research. One possibility could be to create a combination of TDQM and QAFD methodologies so that you will do the objective and subjective measurement using QAFD and based on those result you will create the information product map which is the outcome of TDQM. This would give you an even better understanding of your data and you have a higher-level understanding of the whole process in the map and a deeper understanding of the data based on QAFD analysis. The important thing in the combination would be that there wouldn't be any overlapping when using the combination of the methodologies.

Another interesting thing would be extending the QAFD by giving concreate suggestions on how the quality could be improved based on the analysis. This could be done by adding possible phase 6 in the QAFD process. Right know it is left to the quality specialist and his or her skills to know how to improve the data. I believe that in the future there will be extensions in the current methodologies especially in the improvement phases. The problem is that data quality and in the bigger picture data governance is still at merge and companies and researchers need to understand the importance of it for it to be understood better.

This brings me to another point which is literature related to data quality in general. The same authors have created the methodologies and extensions of them which can lead to a bias world. Many papers bring nothing new to the table and just refer to authors like Batini who has made dozens of data quality related research. This supports my point that data quality is still at its merge. This is especially true when looking at the world of finance. If QAFD methodology is left out there is almost no scientific papers related to data quality in the financial sector. In my opinion, through data governance, the financial sector will start to understand their data as an asset. By understanding data as an asset, it means that it must have a quality and protection and that is the point where data quality comes in. Data quality is stating what is the data asset worth. No quality means the asset is worthless.

Regarding the empirical section of the thesis, one could argue that consistency restrictions are more related to information quality since the data must be put into a context to input logical business rules. But as stated in the beginning of the thesis you can't talk about data quality without talking about information quality. For the analysis to be really related to real life, I decided to take the dimension in to the analysis. Another point is that the dataset was only 10000, which is quite little when the real data set can be in the hundreds of millions. This is since it is from a financial institution and they don't want their data to be shared publicly. This is due to general data protection regulation and the fact that you don't give to your competitor's free information from your database.

The goal of my thesis, which was to find a successful methodology for the data set and suggest improvements based on the results was successful in my opinion. This thesis has helped me link data quality in to the big picture and understand that the quality of your data can be different things. That is why we have data quality dimensions. I have successfully filled my personal goal and am glad for that. I hope that financial world will also start understanding data and its importance even better in the future. Only time will tell.

6 References

Ballou, P. D. & Tayi, K. G., 1989. Methodology for Allocating Resources for Data Quality. *Management of Computing,* 32(3), pp. 320-329.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A., 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys,* 41(3), pp. Article 16 pages 1-52.

Batini, C. & De Amicis, F., 2004. A Methodology for Data Quality Assesment on Financial Data. *Studies in Communication Sciences,* 4(2), pp. 115-136.

Batini, C. & Scannapieca, M., 2006. *Data Quality Concepts, Methodologies and Techniques.* 1 ed. Berlin: Springer-Verlag.

Batini, C. & Scannapieco, M., 2016. *Data and Information Quality.* 1 ed. Switzerland: Springer International Publishing.

Chen, H., Hailey, D., Wang, N. & Yu, P., 2014. A Review of Data Quality Assesment Methdos for Public Health Information Systems. *International Journal of Environmental Reaserch and Public Health,* 11(1), pp. 5171-5027.

Dama-DMBOK , 2019. *Data Management Body of Knowledge.* 2 ed. Delaware: Technics Publications.

Dejaeger, K., Hamersb, B. & Baesens, B., 2010. *A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector.* Singapore, DAMD, pp. 1-9.

English, L. P., 1999. *Improving Data Warehouse and Business Information Quality.* 1 ed. New York: Wiley & Sons.

Eppler, M. J. & Muenzenmayer, P., 2002. *Measuring information Quality in the Web Context: A Survay of State-Of-The-Art Instruments and an Application Methdology.* s.l., Proceedings of the Seventh International Conference of Information Quality.

Fick , U., 2009. *An Introduction to Qualitative Research.* 4 ed. London: Sage.

Francisco, M., Alves-Souza, S., Campos, E. & De Souza, L., 2017. *Total Data Quality Management and Total Information Quality Management Applied to Costumer Relationship Management.* Barcelona, ACM, pp. 40-45.

Haug, A., Arlbjørn, J. S., Zachariassen, F. & Schlichter, J., 2013. Master data quality barriers: an empirical investigation. *Industrial Management & Data Systems,* 113(2), pp. 234-249.

Held, J., 2012. *Towards Measuring Test Data Quality.* Berlin, ACM, pp. 233-238.

Jeusfeld, A. M., Quix, C. & Jarke, M., 1998. *Design and Analysis of Quality Information for Data Warehouses.* s.l., 17th International Conference on Conceptual Modeling.

Laatikainen, T. & Niemi, E., 2012. *Data Quality Johtavissa Suomalaisissayrityksissä: Nykytila ja tarvittavat rakenteet,* Helsinki: DAMA Finland Ry.

Lee, Y. W., Strong, D. M., Wang, R. Y. & Kahn, B. K., 2002. AIMQ: A methodology for information quality. *Inform Mange,* 40(2), pp. 133-460.

Miller, H. E., 1996. The Multiple Dimensions of Information Quality. *Information Systrems Management,* 13(2), pp. 79-82.

Pipino, L. L., Lee, Y. W. & Yang, R. Y., 2002. Data Quality Assesment. *Communications of the ACM,* 45(4), pp. 211-218.

Richard, Y. W. & Diane, M. S., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* Vol. 12(4), pp. 5-33.

Robb, D., 2017. *Semi-Structured Data.* [Online]
Available at: https://www.datamation.com/big-data/semi-structured-data.html
[Accessed 3 October 2018].

Samtisch, C., 2014. *Data Quality and its impacts on Decision-making.* 1 ed. Innsbruck, Austria: Springer Gabler.

SAS, 2019. *Data Management Software.* [Online]
Available at: https://www.sas.com/en_us/solutions/data-management.html#data-quality
[Accessed 4 3 2019].

Scannapieco, M. et al., 2004. The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems,* 29(7), pp. 551-582.

Singh, R. & Singh, K., 2010. A Descriptive Classification of Cause of Data Quality Problems in Data Warehousing. *International Journal Of Computer Science,* 7(3), pp. 41-50.

Stiglich, P., 2012. *Data Governance vs. Data Management.* [Online]
Available at: https://blogs.perficient.com/2012/06/12/data-governance-vs-data-management/
[Accessed 11 1 2019].

Strong, D. M., Yang, L. W. & Wang, R. Y., 1997. 10 Potholes in the Road to Information Quality. *Computer,* 30(8), pp. 38-46.

Su, Y. & Jin, Z., 2004. A Methdology for Information Quality Assement in the Designing and Manufacturing processes of mechanical Products. *Proceedings of the Ninth International Conference on Information Quality,* 1(1), pp. 447-465.

Technopedia, 2018. *Data Quality.* [Online]
Available at: https://www.techopedia.com/definition/14653/data-quality
[Accessed 9 11 2018].

Technopedia, 2018. *Information System (IS).* [Online]
Available at: https://www.techopedia.com/definition/24142/information-system-is
[Accessed 9 11 2018].

Wang, R. Y., 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM,* 41(2), pp. 58-65.

Xiang, J. Y., Sangho, L. & Kim, J. K., 2013. Data quality and firm performance: empirical evidence from the Korean financial industry. *Information Technology Management,* 14(1), pp. 59-65.