Dung Hoang

THE EMPHASIS OF DATA QUALITY IN THE DATA HARMONIZATION PROCESS

Thesis
CENTRIA UNIVERSITY OF APPLIED SCIENCES
Business Management – Enterprise Resource Planning
May 2023

ABSTRACT

Centria University of Applied Sciences	Date May 2023	Author Dung Hoang
Degree programme Business Management – Enterprise Resource Planning		
Name of thesis THE EMPHASIS OF DATA QUALITY IN THE DATA HARMONIZATION PROCESS		
Centria supervisor Janne Peltoniemi		**Pages** 55 + 3
Instructor representing commissioning institution or company Centria University of Applied Sciences		

The thesis aims to analyze data quality and its impact on the data harmonization process. Its purpose is to discover problems, caused by the poor quality of the data, in the data harmonization process and the main goals are to find out methods used to guarantee the data immaculacy for the afterward process as well as to evaluate the room for developing methods that ameliorate the data harmonization process in the future.

In the theoretical framework, the definition of data and data harmonization are explained. Furthermore, the perception of data quality is discussed. From this proposition, the advantages of good data quality and the disadvantages of poor data quality, as well as the measuring methods for data quality, are investigated. Once the causes and the effects are determined, solutions for improving data quality are brought up, providing thought-provoking ideas for similar issues in the future.

The research adopted a qualitative method. An interview was held with a professional in the field of data management. Through the interview, a thorough and practical knowledge of the impact of data quality on the data harmonization process was acquired. Furthermore, the interview revealed criteria for a good data quality source and showed the interviewee's perspective on the Artificial Intelligence support in improving data quality in the future.

The findings of the research indicated that the impact of poor data quality on the data harmonization results in bad decision-making, which negatively affects the revenue of the company. For the purpose of providing a good quality data source, Completeness, Accuracy, and Consistency are obligatory criteria. However, the majority of firms are not aware of the importance of data quality. Moreover, AI can contribute a great support to humans on the data cleaning process in the future, depending on the human’s ability in innovation.

To conclude, firms currently do not take great consideration about data quality management, which is crucial to various business operations.

Key words
Data assessment, data cleaning, data harmonization, data quality

ABSTRACT
CONTENT

1 INTRODUCTION 1
1.1 Background 2
1.2 Objectives 2
1.3 Research problems 3
1.4 Research method and research process 3

2 DATA QUALITY 4
2.1 Definition of data 5
2.2 Definition of data quality 6
2.3 Quality dimension 7
2.3.1 Completeness 8
2.3.2 Timeliness 9
2.3.3 Accuracy 10
2.3.4 Consistency 11
2.4 Data quality assessment 11

3 DATA HARMONIZATION 14
3.1 Definition 15
3.2 Stages of the process 17
3.2.1 Capture 18
3.2.2 Define 18
3.2.3 Analyze 19
3.2.4 Reconcile 20
3.2.5 Review 21
3.2.6 Illustrate 22
3.3 Data quality problem and its impact on the harmonization process 22
3.3.1 Hierarchical analysis 23
3.3.2 Dimensional analysis 28
3.4 Data quality improvement 31

4 RESEARCH METHODS 34
4.1 Research design 34
4.2 Semi-structured interview 35
4.2.1 The interview subject 36
4.2.2 The interview content 36

5 RESULTS 39
5.1 The impact of poor data quality on the data harmonization process 39
5.2 Criteria for evaluating input data 43
5.3 The potential of AI in providing data immaculacy in the future 45
5.4 Findings 49

6 CONCLUSIONS 53

REFERENCES 56

FIGURES

FIGURE 1. The occurrences of particular dimensions in selected frameworks 7
FIGURE 2. Type of measurement used in each framework 12
FIGURE 3. Diverse types of date format 16
FIGURE 4. A step-by-step summary of the data harmonization process 17
FIGURE 5. The process of reconciliation 21
FIGURE 6. The data organization 23

TABLES

TABLE 1. Examples of false duplicate tuples.....26
TABLE 2. Comparison of improvement steps among several methodologies.....32
TABLE 3. List of details questions used in the interview.....37

1 INTRODUCTION

"Know the enemy and know yourself in a hundred battles you will never be in peril" is what Sun Tzu, an eminent Chinese military general and strategist, has said. This perspective has been acknowledged by other individuals, such as Francis Bacon, an English philosopher, and statesman, with the phrase "knowledge is power." All historical events and facts have proved that knowledge is the most powerful weapon that humanity can own, and the way they utilize it will lead to different outcomes. For businesses to maximize profit, the ultimate requirement is to efficiently utilize the information they have, convert it into useful knowledge, and thoroughly exploit its potential. Hence, information has a massive impact on every entity in general and on businesses in specific.

According to Cuesta (2016, 9-10), knowledge comprises meaningful information. The organization and integration of information, coupled with insightful analysis, results in the creation of comprehensive and valuable knowledge. Consequently, the caliber of knowledge is contingent upon the caliber of information. Assuming that the data provided is both comprehensive and transparently aggregated, it is also being scarce in nature. In this scenario, possessing knowledge is undoubtedly valuable; conversely, inadequate information will result in knowledge that lacks worth.

Data serves as the foundation of information. Data is ubiquitously available across the globe. Perceptions that can be visually observed, audibly detected, tactually sensed, olfactorily perceived, gustatorily experienced, and emotionally felt by an individual are classified as data. Data refers to the information that an individual possesses in their cognition. Individuals have the ability to assimilate information from multiple sources and subsequently analyze and synthesize it to form meaningful insights, which can then be utilized to facilitate decision-making processes. Furthermore, the process of writing a dissertation can be perceived as a systematic approach to collecting data, managing information, and exchanging knowledge. Hence, data is attached solidly to the existence of humanity.

As stated in the second paragraph, information quality influences knowledge quality, and the same goes for the conversion from data to information; data quality influences information quality. It is inferred that data quality has a major influence on the data processing process, with the result reflected in information quality. This thesis focuses on a single stage of the process, specifically the work of data harmonization.

1.1 Background

Data is unquestionably playing an essential role in every business activity. The acquisition and examination of data are fundamental procedures for organizations to ascertain their objectives, formulate tactics, and acquire comprehension of the intended markets and their competencies. In one phrase, data is vital for firms to constantly improve their performance. Hence, it is imperative to be cognizant of the significance of data quality.

In the current era of digitalization, an increasing number of technological advancements are being created to aid individuals in the field of management. This is due to the exceptional mathematical proficiency, extensive memory capacity, and capacity to generate and exhibit real-time data possessed by machines. The presence of numerous multinational corporations with dispersed subsidiaries across the globe results in voluminous and asynchronous data that poses challenges for firms during the collection and analysis stages.

While pursuing the academic program at Centria University of Applied Sciences, the author was presented with an opportunity to engage in a project that involved the collection, analysis, and visualization of data. She encountered a significant obstacle during the Extraction-Transformation-Loading procedure within the SAP BW/4 HANA application. Ultimately, the author successfully resolved the matter. The aforementioned incident piqued the author's curiosity regarding the data cleansing procedures employed by corporations in practical settings, primarily due to the fact that the information which her team handled is no longer merely 'data,' but rather, 'big data.' What is the impact of data quality on the process and stages of data harmonization in situations where organizations gather external data from diverse sources and internal data from various subsidiaries? The primary objective of this dissertation is to resolve this enigma.

1.2 Objectives

The primary topic of the thesis pertains to the investigation of data quality and its impact on data harmonization. The thesis focuses on various aspects related to the evaluation of the impact of insufficient data quality on data harmonization. These include the assessment of input data, identification of issues,

and implementation of appropriate solutions. Furthermore, the thesis endeavors to examine the potential of Artificial Intelligence (AI) in facilitating the data cleaning process for seamless data harmonization, as well as serving as a knowledge repository for upcoming scholars interested in this domain.

1.3 Research problems

The goal of the thesis is to answer these following questions:

Question 1: What are the detriments in the data harmonization process caused by poor data quality?

Question 2: What are the criteria for assessing input data?

Question 3: How far can AI support organizations in providing data immaculacy in the future?

1.4 Research method and research process

The research process includes three phases. Firstly, scientific articles, books, and other literature sources were studied to gain the necessary information. Subsequently, an interview was conducted and its findings were analyzed in conjunction with previously acquired knowledge. Finally, the outcome was recorded, and a deduction was drawn.

Given the primary focus of the study on a singular data analysis process, it was inevitable that the scope of the study would be constrained and preclude the exploration of alternative processes. Furthermore, this subject matter entailed information that surpassed the extent of the writer's understanding. Therefore, there might be material for which the author is unable to transmit the original author's message.

The research method used in the thesis was the qualitative method so as to acquire a deep and meaningful perception, which was crucial for the thesis that needed professional's experience for the delivery of current and practical information (Chapman 2022)

2 DATA QUALITY

The quality of data holds significant importance in both personal and professional domains. Instances exist where individuals receive erroneous messages or peculiar correspondence delivered to their residences, bearing the names or addresses of disparate recipients (Scannapieco, Missier & Batini 2005.) The instances mentioned above exhibit substandard data quality and they show that disseminating inaccurate information has a restricted impact on both the originators and recipients of communication in typical discourse and everyday circumstances. Nevertheless, in the corporate sphere context, the deficiency in data quality will result in significant repercussions. The dissemination of even a small amount of inaccurate information has the potential to precipitate the downfall of a company. The phenomenon perplexes individuals in positions of authority, leading to erroneous interpretations of events and misguided choices.

Firms need data to review and evaluate performances, forecast the future, and make decisions. Poor data quality can lead to imprecise circumstances, which in turn can result in missed business opportunities or incorrect decisions. Furthermore, it is noteworthy that companies are compelled to allocate additional resources, including time, labor, and expenses, to rectify deficient and/or inaccurate data (Cichy & Rass 2019.) Additionally, IBM's estimation in 2016 revealed that inadequate data quality resulted in an annual loss of $3.1 trillion for companies in the United States (Redman 2016.) In 2018, Gartner research stated that firms believe around $15 million per year is lost for the same reason (Moore 2018.) Organizations hold the belief that inadequate data quality results in significant costs and reduced operational effectiveness. Consequently, there has been an increase in recognition of the importance of data quality.

With the aim of mitigating the adverse monetary impact resulting from suboptimal data quality, companies implement a benchmark for evaluating the worth of data. The standards' indicators are intended to satisfy users' requirements, and therefore, they are subject to variation and modification as long as they continue to meet those needs. However, specific quality dimensions have been established to enhance the transparency and comprehensibility of this concept. This enables enterprises to efficiently create their data quality assessment framework. Regarding the quality dimension, it is an element that reflects a distinguishing feature of the data (Scannapieco, Missier & Batini 2005.) In this chapter, the definition of data quality, quality dimensions, and data evaluation methodologies are investigated.

2.1 Definition of data

Data might have several multiple meanings or only one at a time. Data objectively speaking, is the core of what people perceive. The essence of an object or a phenomenon is data. That is set in stone and will not alter. However, different individuals have different perspectives, so subjectively, each person has their explanation about the definition of data. Human beings cannot alter what has happened or is happening, but they can change their viewpoint, which has since changed the facts gathered by the individual. Hence, collecting insufficient or imprecise data will result in a variety of outcomes.

To convey this concept more understandably and transparently, the author used a Vietnamese fable about five blind fortune tellers and an elephant. The elephant's appearance piqued the interest of five blind fortune tellers. They paid the elephant handler to touch the elephant when someone walking an elephant came across them. The first person touched the elephant's trunk and compared it to a leech. The second person who touched the ivory described the elephant as a yoke, the third person who touched the elephant's ear characterized it as a fan, the fourth person who touched the leg outlined it as a pillar, and the fifth person who touched the elephant's tail referred it as a dull broom (Truyện cổ tích.) The elephant's trunk, ivory, ear, leg, and tail are its features, which can be viewed as a set of data describing the elephant. Yet these blind people in the fable failed to gather entire data, and the result was controversy in identifying who was right and who was wrong. In conclusion, it is the bad reception of data that leads to disparities in information and knowledge.

In addition, there are other definitions of data. According to Cambridge Dictionary, data is "information, especially facts or numbers, collected to be examined and considered and used to help decision-making or information in an electronic form that can be stored and used by a computer." In the book Practical Data Analysis (2016), Cuesta stated that data represented the facts of the world or a phenomenon. It appeared in several forms, such as alphanumeric, images, and sounds.

For Fox, Levitin and Redman (1994), their aim is to seek a definition that can support them in enhancing data quality, which means the definition should be able to meet two sets of criteria: Linguistic criteria and Usefulness criteria. The Linguistic criteria consist of clear and straightforward definitions, not involving information to avoid circular definitions, and familiar usage tightness to avoid getting lost in another definition. The Usefulness criteria consist of these components: conceptual and representational data features, comprehensive applicable function, and the ability to suggest quality dimensions. The

explanation of data, which defines it as a collection of data items, is adopted from Tsichritzis and Lochovsky [1982.] This assemblage comprises a triple <e, a, v>, where the value v is chosen from the attribute a's domain to denote the value of the attribute for the entity e (Fox, Levitin & Redman 1994.) Each party explained their understanding about the definition of data but in the end, it can be summarized that data reflects what exists in the world under several forms.

2.2 Definition of data quality

Data is the basis for information. High-quality data brings valuable information and generates profits for businesses. However, the value of data is variable based on context and utilization. To eliminate the amorphous meaning of the data quality, a number of researchers determined it, and they could be divided into two different definitions, which are appropriate to meet user needs or fitness for use (Sidi et al. 2012 [Alizamini, Pedram, Alishahi & Badie 2010; Wang 1998].) As a result, requirements for data quality differ among sectors and features. For example, the data displayed to sales managers should focus on periodic sales income, regional sales revenue, and cost of sales. In contrast, logistics managers want data such as regional sales volume, average delivery time between facilities, and logistical charges.

In addition, poor data quality leads to severe repercussions. It not only reduces productivity but also harms the brand's image (Anodot 2019.) Customers and business partners may lose faith in the company because they question its efficacy and genuineness. Hence, it will take the business a huge amount of time and effort to regain the trust of its customers and business partners. This is an illustration of the consequences of poor data quality: A seafood company needs to deliver ten tons of fish to a restaurant. However, because of a mistake by the person in charge, one ton of fish is purchased instead of ten tons. When the restaurant receives the goods, they are outraged and feel defrauded. As a result, the seafood company must take time to apologize and compensate its customers for the error caused by inaccurate data. As it is shown, poor data quality has a tremendous impact on enterprises and can lead to an organization's demise. According to research, corporations cost $3.1 trillion per year owing to poor data quality, and this figure was documented only in the United States (Redman 2016.)

2.3 Quality dimension

The definition of data quality is elusive and lacks established frameworks; however, scholars have identified the essential dimensions for evaluating data quality. The concept of data dimension refers to a set of various metrics that can be employed to assess the quality of data, thereby enabling the determination of the data's overall quality level. There exists more than one methodology for evaluating data quality. Numerous frameworks consist of diverse criteria that have been developed within a particular methodology to ascertain crucial outcomes for a specific organization or industry. It is comprehensible that the framework's consistency may vary depending on the context and characteristics of the data (Cichy & Rass 2019; Batini, Cappiello, Francalanci & Maurino 2009.)

In the majority of literature sources, the quality dimensions which were used the most are completeness, timeliness, accuracy, and consistency. Additionally, they are components of the basic set of data quality dimensions (Batini, Cappiello, Francalanci & Maurino 2009 [Scannapieco & Catarci 2002].)

In general, accuracy is the first factor that comes to mind when discussing data quality. As a result, accuracy is considered a dimension that cannot be overlooked in any framework. However, FIGURE 1 illustrates the contrary while not accuracy but completeness is the most commonly utilized dimension in many frameworks. Following them are timeliness, accuracy, accessibility, and consistency (Cichy & Rass 2019.) It can be observed that accessibility has been alluded to more in numerous frameworks subsequently, as this literature source was written after the source used to establish the essential data quality dimension set.

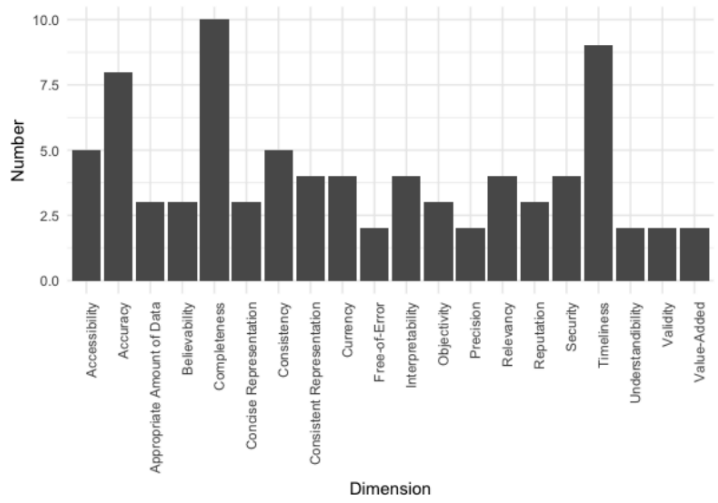

FIGURE 1. The occurrences of particular dimensions in selected frameworks (Cichy & Rass 2019)

It can be seen in FIGURE 1 that Accuracy, Completeness, Consistency, and Timeliness are dimensions utilized most in several methodologies. Since investigating supplementary dimensions is beyond the scope of this thesis, only basic dimensions are elaborated, which are four dimensions stated in previous sentence. These dimensions are analyzed in the majority of scientist articles.

2.3.1 Completeness

Insufficient data can hinder the effective execution of a task. The degree to which the data adequately covers all attributes of assigned entities is referred to as completeness. In instances where a dataset is missing a particular value, a designated null value will be utilized to substitute for that specific position. The aforementioned element is referred to as "null." A null value denotes the absence of a value within a dataset, despite the possibility of its existence in the corresponding real-world scenario. There exist three potential scenarios that dictate the significance of the null value. There are three possible scenarios regarding the existence and knowledge of an attribute:

- The value of an applicable attribute exists but it is unknown (1)
- The attribute does not exist (2)
- The existence of the attribute is impossible to define (3)

An example scenario could be a survey with incomplete data, precisely the absence of one phone number. In scenario (1), wherein the individual possesses a mobile device, albeit the contact number is undisclosed. In this instance, there is a state of incompleteness. In the event that the scenario falls under case (2), the individual in question does not possess a mobile device. The adequacy of the data in this instance stems from the absence of an attribute that would facilitate the display of its value. In scenario (3), the individual's possession of a mobile device is indeterminate. The classification of data can be determined based on its completeness or incompleteness, as the status of the attribute remains unknown (Fox, Levitin & Redman 1994 [Lee 1988]; Batini, Cappiello, Francalanci & Maurino 2009 [Atzeni & Antonellis 1993].)

The presence of incompleteness in data can result in various issues, which can be classified into two categories: the absence of triples <entity, attribute, value> and the existence of additional triples <entity, attribute, value>. The phenomenon of missing triples occurs when a value that is expected to be present in a given dataset is absent, and conversely, extra triples appear when in the same attribute of an entity, there is more than one value (Fox, Levitin & Redman 1994.)

2.3.2 Timeliness

Data alters throughout time. The utility of data collection is contingent upon a specific temporal context and may become obsolete as circumstances evolve. Hence, the dimension of timeliness significantly contributes to data quality evaluation. The concept of timeliness is commonly associated with being sufficiently current, and it comprises two components, namely currency and volatility (Batini, Cappiello, Francalanci & Maurino 2009 [Bovee et al. 2001].)

The age of information can be quantified through the expression of currency. The statement pertains to the identification of outdated data values. The term "out-of-data" refers to a situation where the data is imprecise at a specific time t, yet it remains valid to a certain degree prior to t. The antonym of out-of-date is up-to-date, denoting a situation where the information is accurate at a specific time t (Fox, Levitin & Redman 1994; Batini, Cappiello, Francalanci & Maurino 2009 [Bovee et al. 2001].) For instance, the price of products ten years ago was considered out-of-date because it has no value applying to the present time.

Volatility can be defined as a metric that quantifies the degree of variability and frequency of fluctuations in the value of a particular attribute of an entity. Data can be considered extremely sensitive and susceptible to alterations in certain aspects. The foreign exchange rate exhibits volatility and is subject to fluctuations based on various external factors. The determination of an entity's permanence is a challenging task owing to its mutable nature, albeit with a low possibility, for example, nationality and gender (Scannapieco, Missier & Batini 2005; Fox, Levitin & Redman 1994; Batini, Cappiello, Francalanci & Maurino 2009 [Bovee et al. 2001].)

2.3.3 Accuracy

The existence of inaccurate data may lead to substantial consequences. The accuracy dimension is utilized to evaluate the exactness of the information. The metric of accuracy is utilized to measure the level of concurrence between a given value v and the value v' of a particular attribute a, which is considered precise for a given entity e. Fox, Levitin, and Redman (1994) posited that a data value, v, is deemed accurate when it corresponds with the precise value v'.

The assessment of data involves two distinct types of accuracy, namely syntactic accuracy and semantic accuracy. The measure of syntactic accuracy evaluates the dissimilarity of the comparison function between the values v and v'. The concept of semantic accuracy pertains to the distinction in meaning between two values, v and v', wherein v is deemed syntactically correct (Scannapieco, Missier & Batini 2005.) Consider a scenario where an individual has composed two statements: "A dog has for legs" and "A dog has one leg." The first sentence has a syntactic accuracy error because it is supposed to be "four," not "for." The second sentence has a semantic accuracy error because, in fact, a dog has four legs, not one leg.

The value v' supplied by the facts is used to determine the correctness of the value v. Yet, identifying the proper value in practice is challenging, even whether it can be stated or cannot be specifically described (for example, some nations have two kinds of spelling names). As a result, quantifying data accuracy is complicated because error arises not only from the entity and attribute but also from the application (Fox, Levitin & Redman 1994.)

2.3.4 Consistency

Consistency and accuracy are different, although it is difficult to distinguish in some contexts. The consistency dimension indicates the infringement of semantic rules delimited over data collection. The data gathered should be persistent in the realistic world, and this is the integrity constraint, with reference to the relational theory (Batini, Cappiello, Francalanci & Maurino 2009; Scannapieco, Missier & Batini 2005.)

Two primary categories of integrity constraints can be distinguished, namely intra-relation constraints and inter-relation constraints. Intra-relation constraints refer to the limitations imposed on one or more attributes within a relation, which determine the permissible values within the domain of an attribute. Interrelation constraints pertain to attributes that are present in multiple relations. Consider the case of a Movie relationship. This set has the Title, Director, Year, and LastRemakeYear attributes. Moreover, Oscar Awards is another set that provides information on which movies win Oscar awards in which Year. The intra-relation constraint states that LastRemakeYear must be higher than the Year, and the inter-relation constraint states that OscarAwards.Year must be equal with Year (Batini, Cappiello, Francalanci & Maurino 2009; Scannapieco, Missier & Batini 2005.)

2.4 Data quality assessment

Data management is a vital activity since it relates to the majority of business operations, and an essential part of this activity is the data quality assessment. In this process, selecting and determining methods is a priority because this will affect the results and upcoming stages. By choosing a suitable data quality measurement method, the efficient level of each dimension is evaluated precisely. Furthermore, the assessment process must consider data-influencing aspects such as data source, data-producing process, and data aggregation. An enormous scope of obligatory investigating objects engages in evaluating the data quality effectively. Hence, the complexity of the process may vary based on the objectives and breadth of the organization (Cichy & Rass 2019.)

The measurement type can be subjective or objective, and several frameworks were built objectively. In methodology, various data quality metrics, which are used to indicate the quality level, are combined so as to gain a comprehensive view since, in some circumstances, one metric needs to be measured sufficiently for the data quality dimension. Moreover, it is shown that using metrics to measure data quality

may cause errors, so subjective measures are used in some dimensions instead of objective measures. Subjective measures are survey questionnaires, interviews, et cetera. Nevertheless, research indicated that the number of methodologies using objective metrics is remarkable. FIGURE 2 illustrates the summary of measurement types used in distinct methodologies (Cichy & Rass 2019.)

Framework	Main Components
AIMQ	Subjective assessment: Survey questionnaire
CDQ	User Interviews and definition of data quality metrics for accuracy and currency
COLDQ	Consumer surveys and definition of various data quality metrics
DQA	Stakeholder expectations and definition of quantitative metrics (functional forms)
DQPA	Definition of primary data source and derived data quality metrics
DQAF	Definition of set of data quality metrics for different types of measurement
HDQM	Definition of data quality metrics for accuracy and currency
HIQM	Objective assessment through measurement algorithm suggested
OODA DQ	Not specified
TBDQ	Survey questionnaire and simple ratio
TDQM	Consideration of business rules and definition of data quality metrics
TIQM	User expectations and definition of data quality metrics

FIGURE 2. Type of measurement used in each framework (Cichy & Rass 2019)

There are several data quality assessment frameworks serving different purposes for different organizations. Each framework has different steps within the assessment process. However, basically, they can be divided into two main types: a framework with no formal steps and a framework with detailed steps. It can be seen that frameworks with no formal steps tend to combine subjective measurements and objective metrics, while in frameworks with formal steps, objective metrics are plurally used (Cichy & Rass 2019.)

For a comprehensive and meticulous understanding of the difference between the two main types, two frameworks are selected: DQA (no official steps) and DQPA (has official steps). DQA is the short name for Data Quality Assessment, created by Pipino in 2002 (Batini, Cappiello, Francalanci & Maurino 2009.) In this framework, although metrics should be utilized, no formal process is built. To support the quality assessment, subjective measurements are combined with objective measurements. The outcomes are classified into four categories: low subjective and objective evaluations, low subjective and high

objective evaluations, high subjective and objective evaluations, and high subjective and low objective evaluations. This outcome will be the basis for the consequent analysis and improvement stage (Cichy & Rass 2019.)

As opposed to DQA, DQPA contains seven steps for the assessment process. DQPA is the short name for Data Quality Practical Approach, created by Angeles and Garcia-Ugalde (2009.) The initial phase involves the establishment of crucial data quality characteristics to construct an impartial collection of measurement criteria. The second stage involves a discussion of pre-existing metrics to furnish objective Quality Metadata, comprising an impartial evaluation that is not contingent on user input. The third phase entails the establishment of the requisite methodologies for the representation, interpretation, and evaluation of data quality indicators. The inclusion of objective criteria and process criteria in methods is emphasized as a means of generating scores that are both meaningful and valuable. Furthermore, the assessment of data through the utilization of data lineage is a significant aspect of this stage. During the fourth step of the process, an estimation of the quality scores of primary data sources is conducted and subsequently recorded in the Quality Metadata that was generated in the second step. Subsequent to this procedure, quality evaluations are employed to appraise the acquired information. The results are additionally stored in the Quality Metadata. The sixth step presents two analytical alternatives based on the user's specifications and corporate data. The assessment of data quality can be performed through two methods: firstly, by selecting the most optimal data sources based on their quality scores prior to query execution; and secondly, by comparing aggregated scores associated with different query plans for the identical business problem. Ultimately, the seventh and final step entails the prioritization of data sources through a comprehensive evaluation of their respective data quality scores and the preferences of the users (Angeles & Garcia-Ugalde 2009.)

Despite their minor similarities, the two frameworks have separate architectures. If this is the case for two methodologies belonging to different main types, then frameworks under the same types involve more resemblance. Although they might share some features if looking from a few angles, delving into the details will reveal the discrepancy considerably. To conclude, each framework possesses unique characteristics when the intricacy level is considered (Cichy & Rass 2019.)

3 DATA HARMONIZATION

In light of the swift pace of global development, companies are required to manage a greater volume of information at an accelerated rate relative to historical norms. The global situation is subject to constant change, and organizations that fail to respond promptly may find themselves at a disadvantage. As the economy has grown, the organization's scope has expanded, resulting in an increase in the amount of data that needs to be collected and processed. The term "Big Data" emerged in the early 21st century, representing a significant milestone in the advancement of technology and AI. It is widely recognized that the processing of large-scale data sets exceeds human capabilities. De Mauro, Greco, and Grimaldi (2015) assert that Big Data is a vast and heterogeneous information resource that requires the utilization of innovative technologies to analyze and convert it into valuable insights. In order to fully leverage the potential of big data, it is necessary to implement a variety of processing procedures.

Effective handling procedures optimize data management efficiency, yield favorable results, and prevent unfavorable issues. Data management refers to the systematic process of transforming raw data into refined and organized information, which is achieved through various stages, such as data input, collection, and filtration, culminating in the production of final results. Data management is a comprehensive procedure that involves the ingestion, accumulation, organization, and preservation of data. Effective data management is crucial for obtaining valuable information that can be used for data analysis. This, in turn, facilitates efficient operation and decision-making that aligns with the organization's goals and strategies. The development of a comprehensive data management system necessitates the application of essential disciplines, namely data modeling, data harmonization, data governance, data quality management, and master data management (Stedman 2022.)

In order to obtain a comprehensive understanding of the present state of the organization, as well as a detailed perspective of each subsidiary and headquarter, leaders must establish a systematic approach for gathering data from diverse sources and consolidating them into a cohesive dataset. This measure enables the management to conduct a comprehensive and thorough analysis of the collected data. Data harmonization plays a critical role in the data management process of organizations, providing numerous benefits to firms. The implementation of this system enables effective management of the organization, providing enhanced oversight and facilitating informed decision-making that aligns with the company's

objectives and long-term plans. Thus, maintaining a reliable data warehouse serves as a robust foundation for a company to achieve multiple accomplishments.

For the purposes of acquiring profound knowledge of data harmonization, this chapter scrutinizes the definition of data harmonization, the stage involved in the process, the element influencing the process, and other relevant concepts.

3.1 Definition

There are various approaches for a single problem, and there are two main types of approaches which are through primary sources and secondary sources. With primary sources, researchers must investigate by themselves. Thus, primary sources can also be understood as those directly and closely related to the research topic. Secondary sources are documents that interpret related topics with primary sources, for instance, books, articles, et cetera (Cowart 2017.)

The consolidation of information from various secondary sources into a single database is crucial in order to prevent the omission of details or missed opportunities. The platform enables individuals to efficiently and conveniently access information from diverse sources. In addition, researchers can enhance the efficiency of their inquiry by avoiding interruptions caused by the need to incorporate primary sources into their methodologies (Adhikari et al. 2021.)

Nevertheless, it is worth noting that disparate sources may have distinct methodologies for addressing a comparable issue. The occurrence of conflict during the synchronized process leads to disturbances and impedes the advancement of the task. To address this particular concern, the concept of data harmonization is introduced. Data harmonization refers to the systematic procedure of unifying and merging data obtained from diverse sources that exhibit heterogeneity (Adhikari et al. 2021 [Fortier, Doiron, Burton & Raina 2011].)

While integrating various data sources with standard components, the comparability of collected data across separate studies is enhanced, generating the possibility of answering a question that a single source is incapable of solving because of the deficient sample scope (Gurugubelli et al. 2022.) In addition, there is another document that explains the term “data harmonization” in a more perspicuous way.

With an approach that uses a daily situation as an example, the perception of “data harmonization” is illustrated understandably, as well as providing an explicit definition from other relevant terminologies such as data integration and data synchronization.

In Data Harmonization and Modelling Guide for Single Window Environment (UNNexT, ESCAP & UNECE 2012), an example of delivery date from the perspective of three parties (sender, deliverer, and recipient) was illustrated. For the sender, the delivery date is when goods are expected to be delivered. For a deliverer, the delivery date is when goods are set to be delivered. For the recipient, the delivery date is when goods are demanded to be delivered. Because the three parties do not share any standard, and they may use different systems, the data format can be different. FIGURE 3 demonstrates some types of date formats which are commonly used.

DD/MM/YY:	20/08/10
DD/MM/YY (Buddhist year):	20/08/53
DD-MM-YYYY:	20-08-2010
MM/DD/YY:	08/20/10
MM-DD-YYYY:	08-20-2010
YYMMDD:	100820
YYYYMMDD:	20100820
YYYY-MM-DD:	2010-08-20
DD MONTH YYYY:	20 August 2010
MONTH DD, YYYY:	August 20, 2010

FIGURE 3. Diverse types of date format (UNNexT, ESCAP & UNECE 2012)

The system used by the sender may employ the DD/MM/YY format. In contrast, the system used by the recipient utilizes the MM/DD/YY format, and the system used by the deliverer employs the YYMMDD format. This leads to incompatibility when information is shared across organizations, makes the process more complicated, and may cause misapprehension. However, data harmonization can tackle this hurdle. Data harmonization is the reconciliatory process of the nature and configuration of the collected data. Data harmonization improves the consistency of a dataset's description, which includes the meaning and illustrating format, by formalizing the description of a data item (UNNexT, ESCAP & UNECE 2012.)

It is crucial that the information-sharing process streamlines among involved parties so as to ensure the efficiency, agility, and transparency of the cooperation. Thus, data harmonization is an essential component of the data management process. It reconciles data from heterogeneous sources and coheres the dataset by modifying it into a consolidated form.

3.2 Stages of the process

The optimal aim of the data harmonization process is to create a standardized dataset that allows cross-border information exchange interchange via a single platform. Thanks to this convenience, data redundancy, as well as information exchange costs, are minimized, and the information interoperability among stakeholders is increased, leading to the development of cooperation between parties (UNNexT, ESCAP & UNECE 2012.)

For the purpose of building an efficient data harmonization process, each step included is deliberately examined and applied in practice. According to UNNexT, ESCAP, and UNECE (2012), there is a process that is acknowledged as the best practice and is suggested to use in various data harmonization guides. However, in a journal article written by Rolland et al. (2015), a stepwise approach is generalized and is proven to be engaged in several data harmonization projects. Therefore, a cohesive methodology for data harmonization is formulated following an evaluation of two separate sources. FIGURE 4 manifests the synopsis of the structure data harmonization process, which is adopted from two mentioned sources, and an elaborate description of each step will be discussed in this section.

FIGURE 4. A step-by-step summary of the data harmonization process (adopted from Rolland et al. 2015 and UNNexT, ESCAP & UNECE 2012)

3.2.1 Capture

The initial stage of data harmonization necessitates the establishment of clear objectives and inquiries that delineate the scope of the endeavor. It is imperative that all members of the team reach a consensus on these matters, as the primary aim of data harmonization is to collect pertinent data that can be utilized to address current issues. Furthermore, posing additional inquiries pertaining to the identical subject matter that can be addressed would enhance the efficacy of the procedure. As a result, data is collected from multiple related sources depending on the planned goal (Rolland et al. 2015.)

In order to implement a clear vision of a desired outcome and import data, a formal approach is recommended. This approach is called Business Process Analysis (BPA), and there are three phases included in the BPA: Scope setting, Data collection, and Organizational analysis. In the first phase (Scope setting), the scope of the project is identified so as to establish the limitation and work as a beacon guiding the project toward the desirable goal. In the next phase (Data collection), background information is acquired by conducting interviews and investigating relevant documents. This phase supplies an insight into the stakeholders' requirements for final outcomes, which complete the desired picture mentioned earlier. In the final phase (Organizational analysis), after collecting subjective information (interviews) and objective information (documents), it is time to analyze the current situation of the organization and identify the bottlenecks. Its purpose is to find out the present impediment as well as the hidden factors which might cause congestion in the future. Afterward, a detailed plan is built, and essential resources, such as labor and technological resources, are well-prepared. It is notable that the data harmonization plan should be realistic and suitable for the project's conditions (UNNexT, ESCAP & UNECE 2012.)

3.2.2 Define

Upon delineating the inquiries necessitating the utilization of a harmonized data set, the subsequent procedure delineates the overarching data concepts. The team responsible for harmonization is required to engage in conceptual deliberation regarding their overarching objectives and reach a consensus on the fundamental concepts that are indispensable for meeting the needs of users. The ultimate formulation of the project's inquiries may differ depending on the arrangement of the data concepts. Following this, a comparison is made between the concepts related to the data and the sources of data to ascertain whether the data sources are adequate for the optimal objectives of the teams. Upon conducting an evaluation of the accessibility of data sources in relation to data concepts, the harmonization team must engage in

deliberation regarding the identification of common data elements that most effectively depict the corresponding data item across multiple data sources (Rolland et al. 2015.)

Since each data source may demonstrate the exact semantics of the data but with different terminologies and formats, it is vital to obtain a rigorous description of the indispensable data. The outcome of this stage is a common data element (CDE), which stakeholders agree upon. CDEs for the logistics field can include consignee, freight forwarder, bill of lading, et cetera. The CDE serves as a foundation for a substantial perception of the data's linguistic, type, depiction, composition, and restraint. In addition, a good set of CDE facilitates the afterward process. Transparent, coherent definitions of data items' semantics reinforce the harmonization process and evade the disparity in the mapping (Rolland et al. 2015; UNNexT, ESCAP & UNECE 2012.)

3.2.3 Analyze

The set of CDE created in the Define step is used in step 3: Analyze. Depending on the meaning of the data items in divergent sources, CDEs, which semantically relate to others, are collected into the same document category. For instance, CDEs for documents relevant to commercial transactions can be listed in one category, and CDEs for documents that are pertinent to legal procedures can be accumulated under one category. This step encourages the comparability of distinctive documents as well as the consistent mapping of the data items to the correlative CDE in a data model (UNNexT, ESCAP & UNECE 2012.)

Notably, the data items in various sources may have equivalent names, but their interpretations and demonstrations differ. For example, two documents have a list with similar names, "Document type," but their components differ. The first Document type list includes the initial and final documents, while the second list consists of the contract and receipt documents. It is transparent that they are incongruous with others. Therefore, it is vital to do a conceptual analysis of each data item in the relevant sources to summarize the information gathered and to shape the reconciling process, as these data items will be linked to the corresponding structures in the data model in the next step (UNNexT, ESCAP & UNECE 2012.)

3.2.4 Reconcile

Once all sources have undergone semantic analysis and their respective data dictionaries have been adequately prepared, a data model is built or derived from an existent data model contingent on the illustrative purpose. In the harmonization process, the data model serves as a basis and an intermediary for the association of data items in distinctive data dictionaries. Since its function has a considerable impact on the process, it must be agreed upon and verified by the stakeholders in order that no controversy will happen among participating parties (World Customs Organization 2017.)

Upon completion of the data dictionaries, a process is initiated to map the data items from diverse sources in order to generate a unified data element in the data model, provided that they possess semantic equivalence. The result is a complete data model with a subset including data items compiled from various relevant documents. The data model will be used in the ultimate step for creating a particular code for each subject in specific aspects for the purpose of streamlining the documentation process and avoiding mistakes in delivering information among related objects. Depending on the outcome requirement, the reconciliation can be simple or complicated. Simple mapping consists of making minor adjustments, while complicated mapping involves data conversion so as to provide necessary information for the placed questions. It is apparent that decisions and rules which address inconsistencies in the data must be included when executing the data conversion to prevent turmoil in the common data elements (Rolland et al. 2015; UNNexT, ESCAP & UNECE 2012.)

In order to offer an intelligible approach to this step, FIGURE 5 depicts a procedure for selecting a common name for a data item from several data dictionaries. The outcome of this procedure will be utilized for international commerce. Therefore, the result utilized the World Customs Organization (WCO) data model as the basis for its reference data model and the United Nations Trade Data Elements Directory (UNTDED) as the globally recognized standard. More specifically, The UNTEDED is a compilation of data elements that find application in global commerce (World Customs Organization 2017; The United Nations Trade Data Element Directory 2005.)

FIGURE 5. The process of reconciliation (World Customs Organization 2017)

As it is illustrated in FIGURE 5, following the linguistic analysis, it is possible to infer that "unloading" and "unlading" have the same meaning, and so "unlading" is used in both situations. Furthermore, the words "foreign" and "domestic" are recognized by their aim (export or import) and can be eliminated. Finally, the word "port of unlading" is used. Yet, in the UNTEDED, there is no phrase named "port of unlading" however there is "location of Discharge". Following a comparison of the semantics of both terms, it is agreed that "site of discharge" will be used (World Customs Organization 2017.)

3.2.5 Review

Once the reconciliation process is completed, the harmonized data must be reviewed in order to guarantee that it is displayed in an accurate format and is in scheduled value ranges. At this step, the quality of the harmonized data is assessed. Issues that may arise after the mapping activity are outlier data items, inaccurate values, redundant values, missing data, duplicated values, and logical constraint violation. Moreover, eligibility logic checks are required if the harmonized data contains the eligibility criteria. In addition, since the data harmonization process occurs to optimize the data handling process, surplus data elements that do not contribute to the analyzing or decision-making process should be eliminated, deducting the burden to the data storage and processing pace. Consequently, the harmonized team must identify the data issues and elucidate the mapping logic, and a supplementary process may be executed until the result satisfies the stakeholders (Rolland et al. 2015; UNNexT, ESCAP & UNECE 2012.)

3.2.6 Illustrate

The data harmonization process will be worthless if it is neither recognizable nor utilized. The outcome of the final stage, as well as the harmonization process, is to build a structure for the data model. The framework can be constructed using software programs with particular syntaxes and associated technological techniques. Therefore, this step requires sufficient technical capacity to satisfy the demands and ensure no errors emerge in the final product. In addition, the final system should be both machine-readable and human-readable as the human being is the entity who needs the information for other activities, and the software program is the entity that executes the data harmonization process, providing compiled information to the human being (UNNexT, ESCAP & UNECE 2012.)

3.3 Data quality problem and its impact on the harmonization process

The operational efficiency of the data harmonization process is significantly influenced by the quality of the data. The procedure of data harmonization encompasses the integration of data from diverse origins. Even a slight discrepancy in a single source can have a substantial impact on the overall process, resulting in disruptions during the integration and reconciliation of data dictionaries with the data model.

Upon conducting a thorough examination of pertinent literature, it has been determined that two distinct methods of analysis exist for identifying data quality issues and assessing their impact on the harmonization procedure. It is noteworthy that the impact significantly disrupts the intermediary phases of the procedure, encompassing Define, Analyze, and Reconcile. The impact of inadequate data quality varies depending on the situation and can affect either of two subsequent stages. In terms of methodology, the approaches can be categorized into two distinct types: hierarchical analysis and dimensional analysis. In regard to the dimensional analysis, it was stated in section 2.3, Quality dimension, that data dimension is the foundation for assessing data quality. Therefore, it can be used to define the data quality problems and perform as the basis for evaluating the impact of poor data quality. The effect is only discussed within the mentioned dimensions since they are key criteria for the judging process.

In the context of hierarchical analysis, it is imperative to take into account that the outcome of the data harmonization process involves the amalgamation of diverse data sources. Consequently, the organization of data based on its granularity becomes an unavoidable aspect of this process. The categorization of data units into four levels is contingent upon their value capacity and interrelation: attribute/tuple,

single relation, multiple relations, and multiple data sources. The data organization hierarchy is illustrated in FIGURE 6 (Oliveira, Rodrigues & Henriques 2005.)

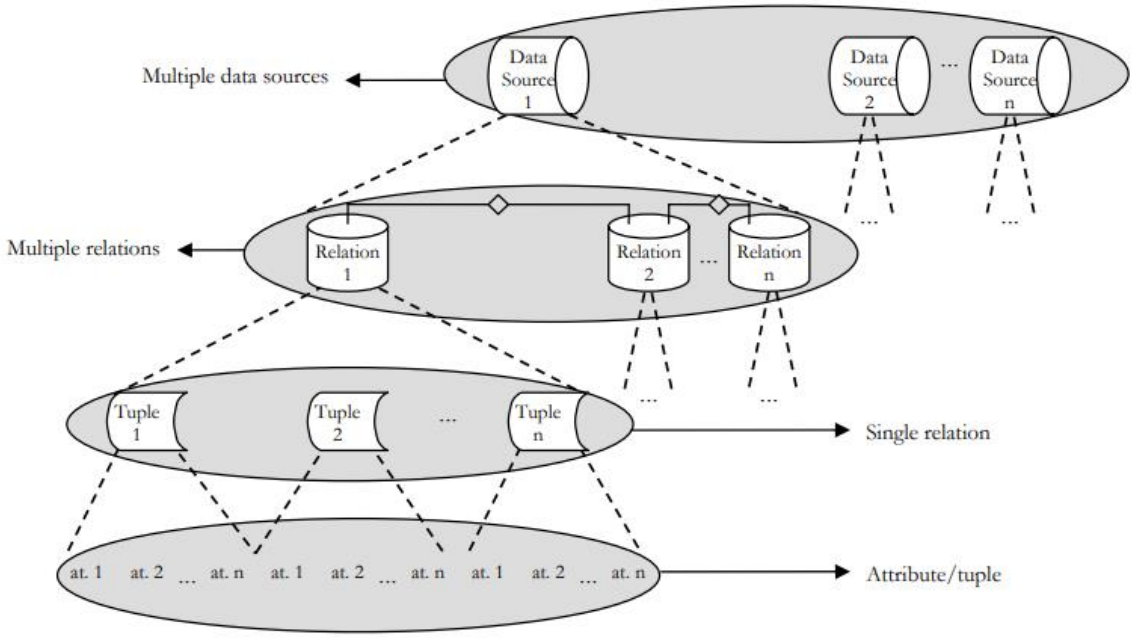

FIGURE 6. The data organization (Oliveira, Rodrigues & Henriques 2005)

3.3.1 Hierarchical analysis

In the context of data organization, a tuple comprises multiple attributes, a relation comprises multiple tuples, and a data source comprises multiple relations. Challenges are present at every level of granularity. However, a mistake in a specific attribute can result in inaccuracies at higher levels. Thus, employing a bottom-up approach is a suitable methodology for analysis. However, a comprehensive analysis has not been conducted as it falls outside the purview of the thesis.

Within the context of attributes and tuples, three distinct categories of errors may arise. These categories include errors of a single attribute within a single tuple, errors of a single attribute across multiple tuples, and errors of multiple attributes within a single tuple. In the scenario where there is a solitary attribute belonging to a singular tuple, the issues that can be identified are various. Incorrect value means a situation where a value is semantically wrong. For example, the expired date is 28/02/2023 but it is displayed 28/03/2023. Misspelling value means a situation where a value is syntactically wrong. For instance, data harmonization is found instead of data harmonization. Imprecise value is a value that does not provide

enough information. For example, for the type of management, the list contains an acronym PM but it is not transparent that PM means Project Management or Product Management. Domain violation: a value is illustrated as illogical information when applied to the required aspect. For example, negative value in sales quantity. Syntax violation refers to a value displayed in a different format compared to other values in the same type of attribute. For example, the unified format for date is dd/mm/yyyy but there is one value displayed in mm/dd/yyyy. Invalid substring identifies a piece of prepend information appears in an attribute that does not require that information. For instance, in a number-only attribute, there is a value "14 tables" while others are "13," "15". Domain constraint violation is a value fails to respect the constraint. For example, the customer's name attribute must have at least two words, but some entities do not satisfy this restraint. Missing value refers to a situation where a value cannot be found in the attribute which is not allowed to have a null value. For example, the customer's name attribute has a null value (Oliveira, Rodrigues & Henriques 2005.)

Insufficient data may arise due to an attribute having an inadequate value. Hence, during the capture and define stages of the data harmonization process, implementers may encounter difficulties as they are unable to obtain comprehensive information for data analysis, decision-making, and the establishment of suitable terminologies for insufficient data elements. Furthermore, the process can be time-intensive as the individuals responsible for carrying out the task may be required to input missing information in order to finalize the data sources, leading to a prolonged timeline for completion. Values that contain error formats may pose a challenge to synchronization and integration processes and may be excluded from the final output due to the inability of the operating platform to recognize their format syntax. Consequently, these values will transform into null values during subsequent stages, ultimately rendering the final output flawed.

In the second circumstance, which consists of the same attribute in multiple tuples, the problems which can be detected are: Unique value violation, Existence of synonyms, and Domain constraint violation. Unique value violation means a scenario that distinctive entities own identical unique attributes. For example, two different products have similar identification codes. Existence of synonyms refers to the situation of two or more words that are semantically identical in an attribute referring to the same subject. For example, in the Cargo situation of two products, one product is "Unlading," and one product is "Unloading" although they refer to the same action which is removing goods from a means of transport. Domain constraint violation is the set of values in a same attribute failing to respect the constraint. For

example, the customer's name must be organized in descending order, but they are not (Oliveira, Rodrigues & Henriques 2005.)

The consistency and coherency among several tuples are vital for a clean data relation. The duplicate of the peculiar value available in more than two tuples will lead to ambiguity and perplexity, mystifying the analyzing and defining process. Moreover, because of the existence of the synonyms, an excess operation is required to determine the unified term for these synonyms, which takes more time and effort from the team.

In the third circumstance, which involves various attributes in a single tuple, the problems which can be detected are: Semi-empty tuple, Functionally dependent violation, and Domain constraint violation. Semi-empty tuple is the circumstance when more than half of a tuple's attributes are void. Functionally dependent violation arises when the values of closely relevant attributes conflict with each other. For example, Helsinki and Kokkola (belong to attribute 'City') have similar zip code 67000 (belong to attribute 'Zip code'). Domain constraint violation refers to the situation when an attribute does not belong to the tuple appearing in that tuple. For example, the tuple for a customer includes the quantity of the product (Oliveira, Rodrigues & Henriques 2005.) The deficient and erroneous issue in a tuple result in ambiguous and perplexing circumstances, as lacking over half of the value in a tuple discard its contribution to the relation, and functionally dependent violation leads to confusion since it is hard to distinguish which is the correct value.

In the level of single relation, several tuples are engaged in, so the scope of deficiency is also increased. Consequently, the majority of the errors derived from the correlation among components in the relation. Three detectable problems are: Approximate duplicate tuples, Inconsistent duplicate tuples, and Domain constraint violation. Approximate duplicate tuples is nearly similar copies of a tuple in the relation. Example for this problem can be viewed in TABLE 1, second row. The tuple of the participant Anna Heavens is approximately duplicate, but her name is not fully written. Regarding Inconsistent duplicate tuples, it is the situation that the copies of a tuple in the relation but the values under the same attribute are dissimilar. Example for this problem can be viewed in TABLE 1, third row. The tuple of the participant Anna Heavens is duplicated, but her last name is mistaken with another participant whose first name is similar to hers. The relation consists of tuples which are not appropriate to the relation's domain is Domain constraint violation. For example, the minimum of product types in a relation is 3, but there are only 2 different product types in that relation (Oliveira, Rodrigues & Henriques 2005.)

TABLE 1. Examples of false duplicate tuples

List of participants			
Numerical order	Participant ID	Name	Date of birth
1	221100AH	Anna Heavens	22/11/2000
2	221100AH	A. Heavens	22/10/2000
3	221100AH	Anna Johnson	22/11/2000
4	240299AJ	Anna Johnson	24/02/1999

Whether or not the duplicate tuples are correct, they are sizable problems in the harmonization process, which can be acknowledged through TABLE 1. Containing numerous copies of a tuple not only disturbs the capture and define step since approximately similar tuples create disconcertment but also causes the discrepancy in the reconcile step as in the single relation A has duplicate tuples while other relations B, C, D, et cetera may not face the same situation.

After investigating problems emerging at the basic level, it can be considered that those problems directly influence the systematic upper level. In the level of multiple relations, some problems emanated from defects in the attribute/tuple level and the single relation level. Firstly, Referential integrity violation refers to a situation where a value v of an attribute a exists in relation r but not in relation r'. For example, in Recipient relation, the Recipient.Ordered_product contains a value 'table' but in the Product relation, there is no such value in the Product.Type. In the next error, assuming the value v is the correct value and the value v' is the incorrect value for the tuple t in the relation r and both v and v' are existing values in relation r' which are linked to the relation r. Tuple t contains the incorrect value v' while it is supposed to have the correct value v. This is Incorrect reference. For example, customer H bought tables, but he actually bought chairs. Both tables and chairs are listed in the product category of the company. There is an incorrect reference in the product that customer H really bought. Thirdly, Heterogeneity of syntaxes is different syntaxes for the same attribute in divergent relations. For example, in the Purchase relation, the purchasing date format is dd/mm/yyyy while in the Receipt relation, the invoice date format is mm/dd/yyyy. Both attributes refer to the same date. Circularity among tuples in a self-relationship is a paradoxical relationship between entities in different relations. For example, product X is a sub-product for product Y in the relation r but at the same time in the relation r', product Y is a sub-product for product X. Domain constraint violation is the situation where the data provided by relevant relations is inconsistent logically. For example, the total number of product types in the relation r is not equal to the

total number of invoices (for each product type) in the relation r' (Oliveira, Rodrigues & Henriques 2005.)

Upon identifying quality issues across various levels of relationships, it is evident that errors occurring at the lower echelons of the hierarchy exert a significant impact on this level. The impact of poor data quality has been discussed in previous parts of the single relation level. In cases of multiple relationships, the impact thereof is often magnified and can have repercussions on other relationships if not properly managed. Inadequate data quality has the potential to result in the non-viability of a complex system that is comprised of numerous interdependent relationships.

At the pinnacle of data organization lies the data sources, which concurrently serve as the origin of the data harmonization procedure. It is apparent that a massive number of issues encountered in source integration can be attributed to complications within the individual components of the sources, in addition to those arising from their interaction. In order to identify issues with data quality, it is assumed that the data sources pertain to the same subject, while the relational data schemas belong to distinct sources. The quality problems that can be detected are various. Firstly, Heterogeneity of syntaxes means that each schema uses disparate syntaxes for a similar attribute. For example, the invoice date in data source DS1 is illustrated in dd/mm/yyyy while in the data source DS2, it is mm/dd/yyyy. Secondly, Heterogeneity of measure units means various measure units because of the discrepancy in operating countries. For example, the data source DS1 is retrieved from a subsidiary located in Germany so it uses the euro while the data source DS2 is retrieved from a subsidiary located in the U.S., so it uses the US dollar. Thirdly, Heterogeneity of representation refers to a scenario where each schema may have divergent illustrations method to indicate the value of the same attribute. For example, in the Free ship delivery section, to specify a Yes or No value, the data source DS1 uses the alphabetical value Y and N while the data source DS2 use the numerical value 0 and 1. Depending on the creator of the data sources, synonyms are used in schemas and this error is Existence of synonyms. For example, the data source DS1 uses the word "Unlading" while the data source DS2 uses the word "Unloading" and both words indicate an identical condition of the goods. In contrast with the synonymous problem, homonymous problem refers to a term that has several meanings and this is Existence of homonyms. For example, a mouse in the data source DS1 refers to a technical device while in the data source DS2, it refers to an animal. Lastly, Domain constraint violation refers to the situation when data sources are integrated, and the result is inappropriate with the pre-established rules. For example, there are 10 types of products manufactured

but the accumulation of different types of products of the data source DS1 and the data source DS2 is over 10 (Oliveira, Rodrigues & Henriques 2005.)

At this stage, it is evident that the methodology prioritized the challenges that emerge from the interplay between data sources. The dissimilarity in data structure and lexical usage underscores the insignificance of diverse data origins and the significance of the data standardization procedure. In the event that the root of the data source, specifically the attribute, and tuple, is afflicted with the issues mentioned above, it is highly probable that the data source will be similarly affected and encounter equivalent repercussions. These may include delays in progress, increased expenditure of time, effort, and resources, as well as ineffective data management. Consequently, organizations face the challenge of managing macro-level consequences that require significant time investment for decision-making while potentially risking missed opportunities.

Upon identifying quality issues across various levels of relationships, it is evident that errors at the hierarchy's lower echelons significantly impact the higher levels. The impact of poor data quality has been discussed in previous parts of the single relation level and multiple relations, the impact is simply exaggerated and affects other relations if it is not well handled. Insufficient data quality may lead to the unsustainability of the whole system constructed by multiple relations.

In summary, the utilization of hierarchical analysis facilitates the dissection of the impact of substandard data quality at each level. It reveals the interrelation between the small constituent units and the larger units comprising them. It is worth noting that issues with data quality are not necessarily absolute but rather contingent upon the perspective of the evaluator. Considering in a single relation, the date format is dd/mm/yyyy and it is still correct if all tuples are displayed in this format. Nevertheless, this format will be a quality problem if it is assessed in the multiple relations nexus, and the other relation has a different date format. The same principle can be extended to more complex data levels.

3.3.2 Dimensional analysis

Hierarchical analysis is concerned with examining a process from a microscopic to a macroscopic level. In contrast, dimensional analysis involves identifying problems on a comprehensive scale and catego-

rizing them into appropriate dimensions. The four dimensions outlined in section 2.3, pertaining to quality, are utilized to examine issues related to data quality and their effects on the process of harmonization in a targeted manner.

The identification of poor data quality traits and their impact on data harmonization can be determined based on the definition of data quality dimensions. The four dimensions under consideration are Completeness, Timeliness, Accuracy, and Consistency. As previously discussed in section 2.3, the aforementioned dimensions served as the primary foundations for both evaluating and enhancing the quality of the data. These dimensions are fraught with significant and substantial data quality issues as a result of the aforementioned argument.

The completeness dimension accentuates the perfection of the dataset. As a result, the problems belonging to this dimension are factors that cause data inadequacy. The most common problem is missing values in the attribute, missing attributes in the tuple, and insufficient information. Three possibilities instigate incompleteness: the anonymously existent status of the value, the anonymously existent status of the attribute, and the non-existed attribute. Furthermore, the significance of an incomplete dataset is determined by the efficiency of the empty attributes in proportion to the maximum possible informativeness of the tuple and the relation. Relying on the applied condition that the null value does not encumber the accumulated content of the relation (Scannapieco, Missier & Batini 2005.)

Nonetheless, incompleteness impedes data harmonization as it does not provide thorough information for the capture and define steps. Without sufficient data sources, obtaining a comprehensive view of the subject is tough. Moreover, if the lacking value bears a vital role, the period for the capture step and define step is lengthened so as to fulfill that empty attribute.

In the timeliness dimension, problems are linked with the data decay and the variable frequency of the data. Take the current milestone as the measure of the profitable level of the data to the user. The degree of the data is determined. It can also be computed through the latest update of the dataset. If the information provided is no longer valuable to the business's current situation, the data quality problem is identified as not current. Regarding the variable frequency of the data, it is highlighted that the statistical data is sensitive and subject to change. If the frequency change interval is protracted, the data source can not deliver accurate and up-to-date information (Scannapieco, Missier & Batini 2005.)

Out-of-date data is a strain on data storage as well as the main element causing the moratorium of the data processing pace of the software. It is useless if the data source contains data that is no longer suitable for the aim of the users, and including these data items will stress the system and put an impediment to the operation. Moreover, suppose one of the data sources has elongated variability frequency. In that case, it can cause information asymmetry, which leads to inconsistency among data sources and the inaccurate final product of the process.

When it comes to data quality problems, the most general issue is inaccurate data. In the accuracy dimension, there are three types of inaccurate issues: syntactic error, semantic error, and duplication. The root of syntactic errors and semantic errors are from the input proceeding. The value has grammatical flaws due to typographical error, and it cannot convey the information correctly. In regard to the semantic error, the incorrect value entered in the attribute provokes this error. While the value falls within the identical domain as the attribute, it is erroneous for the entity in practice. As opposed to the deficient data problem in the completeness dimension, data redundancy because of duplication is categorized as a problem in the accuracy dimension. This issue happens by dint of accidentally duplicating values in a data source (Scannapieco, Missier & Batini 2005.)
The impact of poor data quality on the data process in general and data harmonization, in particular, is evidently immense. In the first two phases of the data harmonization process, imprecise data items generate ambiguity and misinterpretation. As a result of the abnormal values in many data sources, the following steps may encounter a contradicting status. An additional period is required for the purpose of figuring out the origin of the conflicting circumstance, which delays the completion milestone and influences the judgment and formation of the company's strategy, resulting in excessive expenses.

Regarding the consistency dimension, data quality problems have a tight relation to illogicality and contradiction, which include integrity constraints violation when assessing from the semantic perspective and syntax violation when assessing from the format perspective. The provenance of these problems is to record data in a disorganized manner. The data items are posted without considering constraints or unifying the displaying format, leading to the circumstance that the values can be antagonistic to each other and infringe the logic or the regulation (Scannapieco, Missier & Batini 2005.)

The consequence of the contradiction may be viewed transparently in the analysis and reconcile steps. The subsequences of the inconsistent data quality problem include not just irreconciliation but also information asymmetry. Consequently, the process will be unsustainable, and the operating software may

be unable to implement the harmonization. Furthermore, the executors must spend significant time and effort identifying problems across several data sources.

To conclude, dimensional analysis categorizes the data quality problems correspondingly to the appropriate dimensions, which facilitates the assessment of the poor data quality's influence on the harmonization process. Once the dimension of the problem is determined, it will be more straightforward to remedy that problem, accelerating data harmonization in particular and data management in general.

Although it depends on the type of quality problem, its ultimate impact on the harmonization process is causing the process to be slowed down, extra time, effort, and cost consumed, and potentially planned project will be disrupted, leading to the inability to complete the data harmonization process on time and leading to larger consequences for the organization, such as missing opportunities, spending unnecessary costs, or missed opportunities to change tactics to align with the goals and direction of the organization.

3.4 Data quality improvement

After assessing the data quality, it is vital to enhance the data quality of data sources so as to ensure the data harmonization process operates fluently and is expedited. Although during the harmonization process, the aggregated data is completed and improved with a view to delivering an accessible, agile, appropriate data system that meets the requirements of the users, it is critical to provide as clean a source of data as possible and as standardized as possible. In this section, the criteria for data quality improvement are discussed, and these criteria are accumulated from existing methodologies. Additionally, a data architecture for improving data quality is also covered so as to provide an extensive perspective on this process. Furthermore, since data quality improvement requires an intensive workload, the role of AI in this aspect is examined. Back to the first mentioned topic, TABLE 2 depicts the phases for quality improvement compiled from several techniques.

TABLE 2. Comparison of improvement steps among several methodologies (Batini et al. 2009)

Method Acronym/Step	TDQM	DWQ	TIQM	DQA	ISTAT	AMEQ	COLDQ	DaQuinCIS	CDQ
Evaluation of Costs	+	+	+				+		+
Assignment of Process Responsibilities	+		+						+
Assignment of Data Responsibilities	+	+	+						+
Selection Strategies and Techniques	+	+	+		+		+	+	+
Identification the Causes of Errors	+	+	+	+	+	+	+	+	+
Process Control							+		+
Design of data Improvement Solutions		+	+		+		+		+
Process Redesign	+		+		+		+		+
Improvement Management	+	+							
Improvement Monitoring	+		+			+	+		

It appears that all reference frameworks include the step of identifying the causes of errors in order to facilitate improvement. Undoubtedly, in order to improve a particular subject, it is imperative to identify its fundamental impediment. Subsequently, the identification of appropriate methodologies for enhancing data quality through the implementation of effective strategies and techniques is a commonly adopted approach. Diverse approaches and perspectives are adopted by entities based on their respective goals and capacities when addressing the issue at hand. Cost evaluation is a crucial step for most structures. By assessing the cost and benefits of the improvement process, organizations can gain insight into their current capabilities and identify opportunities for practicalizing the process. Furthermore, the evaluation procedure of the methodology can be replicated to scrutinize the outcomes of the improvement phase (Batini et al. 2009.)

Aside from collecting improvement steps from various techniques, gaining an inquiry on a data improvement structure offers an alternate viewpoint on the same subject. Data architecture is a structure that is built monolithically from numerous data handling tasks, such as data collection, data standardization, data integration, et cetera. The purpose of the data architecture is to provide a transparent view of data sources, accessible and efficient data management, ensure data security, detect data quality problems, increase the centralization, and value uniqueness, and data profiling. There are six steps for designing a functional data architecture, which are listed below:

- Delineate the desired goal to systematize stages aligning with the goal.
- Categorize current data sources and filter the necessary datasets.

- Create a homogenous phraseology set.
- Modify the reference architecture to suit the demands.
- Establish a KPI set to evaluate the efficiency of the data architecture.
- Build a plan for enforcing data architecture.

(Przybycień 2023.)

The assessment of the scope of AI assistance is a crucial factor in enhancing data quality. AI is progressively gaining prominence in the daily routines of individuals. In an industry that is heavily reliant on computer science, AI has become an indispensable element. The process of automating data collection, identifying and resolving data quality issues, cross-referencing data with pre-existing sources, filling data gaps, enhancing existing data, and eliminating duplicate data elements through the evaluation of underlying information is commonly employed. Additionally, AI is utilized to expand the scope of data quality operations in correlation with the magnitude of the data repository. Conclusively, the process of improving the quality of data cannot be accomplished seamlessly without the incorporation of AI (Reno 2022.)

4 RESEARCH METHODS

In this chapter, the framework, the argued topic, and the methodology for the selected type of interview are introduced in order to provide a comprehensive perspective on the empirical part. The empirical part will conduct an interview to find an answer for each domain lying under three key questions placed in the introduction chapter.

4.1 Research design

With the purpose of profoundly investigating the interrelationship between data quality and the data harmonization process, the combination of secondary sources and primary sources is optimized. Primary sources are those that have a direct and intimate association with the study issue, whereas secondary sources are those that interpret comparable themes to primary sources, such as journal articles, books, and other pieces of literature (Cowart 2017). After outlining the scope of the study, collecting data sources, and scrutinizing the contents, the fundamental knowledge is constructed. As a result, an insight into the theme is brought up, and inquisitiveness is promoted.

However, several books, journal articles, and website documents are published at various timestamps, and some of them do not accurately concentrate on the thesis domain. Since acquiring information from secondary sources can be insufficient to contribute to an up-to-date, accurate, realistic viewpoint, the finding from primary sources should be accounted for so as to fulfill the empty details and correct the out-of-date information.

Regarding the empirical component, there exist three viable approaches for conducting the procedure: qualitative, quantitative, and a hybrid methodology that incorporates both qualitative and quantitative techniques. According to Mcleod (2023), the qualitative methodology is focused on obtaining non-numerical information, such as written text, audio recordings, and other forms of data, followed by the process of analyzing and interpreting them. This method of investigation elicits an individual's subjective perception. Ethnography, interviews, and case study research are typical qualitative research methods. Regarding the quantitative method, Mcleod (2023) stated that it concentrated on compiling numerical data, later analyzing, and visualizing them so as to manage, depict, and anticipate interested metrics. Through quantitative research, the interaction between variables is examined, and the outcome confirms

or denies a tested theory. Questionnaires and controlled observations are two common quantitative research tools. Furthermore, in some specific cases, the combination of qualitative and quantitative methods is used for the most efficient outcome.

In this thesis, the qualitative method is chosen as a form of an interview. Due to the nature of the investigated topic, an intensive singular interview is required. The interviewee is deeply associated with the data processing field in general and the data harmonization aspect in specific. To conclude, the research structure consists of two phases. In the first phase, academic knowledge is gathered and systematized in an appropriate manner with the predominant theme. Proceeding to the subsequent stage, an interview is administered, and the outcome of the interview will be juxtaposed with the theoretical component in order to conclude the responses to the enigmas that have been presented. Section 4.2 provides an illustration of the definition of interview types, and the rationale for selecting the semi-structured interview is revealed.

4.2 Semi-structured interview

In qualitative research, an interview is one of the regular methods, and there are three types of conducting an interview, which include a structured interview, unstructured interview, and semi-structured interview. The structured interview consists of a planned table of questions with no adaptive questions added throughout the interview. In contrast to a structured interview, which is rigid and makes it hard to excavate deeply into the subject, an unstructured interview contains unprepared questions, which will be generated in response to the contextual evolution of the conversation. In the final structure of an interview, which is a semi-structured interview, a set of questions is formed, and additional questions are spontaneously developed based on the interviewee's answers (Pollock 2019.)

The semi-structured interview type was chosen for this thesis's empirical part due to the fact that having a basis of questions to follow would consolidate the interview process, and supplementary questions could be derived to delve deeper into the theme. The interview took place via Microsoft Teams, and it was recorded for the research of the thesis. The interviewee was informed about this at the beginning of the interview, and the interviewer received acceptance from him. The result was analyzed and interpreted in Chapter 5.

4.2.1 The interview subject

As it is stated in section 4.1, Research design, the interviewee should be a person who has experience in working in the data process in order to satisfy the requirement of the thesis's goal. The interviewee for the empirical part is a teacher from Centria University of Applied Sciences. In addition to being a teacher, he has rich experience working with data in the business industry, and he has a unique perspective on the thesis's topic, which contributes a contrasting point of view to the thesis**.** Most of his sharing in the interview comes from his occupation before being the teacher at Centria University of Applied Sciences, which is the Chief Information Officer of a company operating in construction industry.

4.2.2 The interview content

Section 1.3 delineates research problems and identifies three fundamental inquiries that form the foundation of the interview questionnaire. Two questions have been formulated to investigate the methodology for assessing the quality of data sources and the significance of data quality in the process of harmonization. Furthermore, the observation of the advancement of AI across various sectors prompts consideration of the facilitation of AI in this specific domain.

The interview's subject matter comprises three discrete themes. The primary focus of the theme is to explore the adverse effects of inadequate data quality on the data harmonization process, emphasizing the correlation between data quality and the harmonization process. The aim of this investigation is to evaluate the obstacles that arise during the process of data harmonization and the consequent costs that organizations incur when dealing with data sources that are not in an optimal state.

The second thematic area concerns the formulation and examination of a series of standards that are essential in appraising the excellence of data sources, as a component of the yardsticks for evaluating input data. The aim is to detect erroneous data elements, minimize the occurrence of inadequate data quality, and validate the suitability of the input data for the organization's intended purposes.

The third theme of this study centers on exploring the potential of AI in enhancing data quality. Specifically, the study aims to investigate the extent to which AI can aid organizations in improving data accuracy. This theme is particularly relevant given recent advancements in the field of computer science,

which have led to significant breakthroughs in the application of AI across various domains. The objective of this thematic area is to present speculative situations and forecast the patterns in the workforce as AI progressively supplants human beings in diverse undertakings.

Each theme is accompanied by three to five questions. The primary objective of the first theme is to uncover the fundamental reasons behind data quality concerns, their impact on operational procedures, and the viewpoint of the respondent regarding the level of importance accorded to data quality issues by organizations. The second theme of the study aimed to investigate the assessment criteria that companies prioritize when constructing their data evaluation frameworks for information decision making. Additionally, the study explored the reasoning behind the selection of these criteria and the preferred data management techniques among business leaders. The responses also demonstrate the strategies employed by enterprises in addressing challenges arising from inadequate management of data quality control. The final examination queries pertaining to the subject matter are comparatively more weighted towards AI, a subject that is presently garnering significant attention. This theme pertains to the disclosure of a former Chief Information Officer's perspective on the role of artificial intelligence in facilitating business operations. TABLE 3 presents the questionnaire utilized during the interview.

TABLE 3. List of details questions used in the interview

#	**Question**
1	Can you please describe your main role when you were a CIO?
2	In your opinion, what are the main causes of the poor data quality?
3	Have you ever encountered the situation in which you do not have enough data to perform your work?
4	What are the common problems of poor data sources, and which has the greatest impact?
5	Can you please describe the cost of poor data quality on the data analysis in general and in the data harmonization?
6	Do you think that organizations nowadays are fully aware of the importance of the data quality?
7	Which circumstances that you think are more popular: a company has its own data management team, or they use the services of professional companies? And why?
8	In your opinion, which criteria should a good data source meet?

(continues)

TABLE 3. List of details questions used in the interview (continues)

9	If you have the data that is very old, and you can not use it anymore, will you get rid of it or just leave it there?
10	Do you think that current organizations are taking more care of assessing data quality?
11	I used to encounter data quality problems when I was in a course. Since the volume of the data source was small, I was able to detect them manually by using exclusion method but when it comes to firms with billions of data values, what are their solutions?
12	How much do organizations depend on AI in the data quality improvement process?
13	Do you think that most of the company nowadays are at the low level of applying technology to their business or many of them are already at the high level?
14	Have you seen any errors that cannot be detected by AI and required humans to investigate?
15	Do you think that AI will totally replace human in the data analysis process in the future?
16	For example, we have a set of requirements and we input that to the computer, can the computer return a unified table of data that we are looking for?
17	What do you think about the potential of ChatGPT replacing people in the decision-making process?

5 RESULTS

In order to give a deep and nuanced grasp of the thesis issue, this chapter presents a thorough analysis and illustration of the interview findings. So as to provide empirical support for the theoretical framework and to illuminate the interviewee's viewpoint on the subject matter, the identified problems are methodically examined.

A question concerning the interviewee's occupation was asked before moving on to the specialized themes in order to learn more about their background. The interviewee held a number of jobs prior to beginning his career in 2023 at Centria University of Applied Sciences. He began his career at a business whose key offerings were the enterprise resource planning system SAP as well as other technology solutions. After that, he launched a new business before beginning to teach at Centria. Thereafter, he started his new page as a Chief Information Officer in a company that manufactures steel. He was in charge of that company's data management and information system. Since there were only two employees in the Information Technology & Information System office when he joined the company, despite the fact that it was a sizable enterprise, the respondent openly said that the burden was heavy, particularly when it came to the responsibilities he had to assume. He boosted the number of employees in the office to six after a year of staying there. Therefore, his experience is valuable to the author for the purpose of enriching the knowledge.

5.1 The impact of poor data quality on the data harmonization process

The process of managing data involves a crucial component known as data harmonization. The process guarantees the consolidation of data from diverse origins into a standardized format and its alignment with a consistent semantic framework. The efficacy of data harmonization is contingent upon the presence of data that is of superior quality, devoid of errors, inconsistencies, and other related issues. The process of harmonization can be significantly hindered by substandard data quality, which can have adverse effects on the overall management of data. This can result in erroneous conclusions, missed prospects, and potential business setbacks. The identification of underlying factors responsible for inadequate data quality is imperative in effectively resolving such concerns. The interviewee offered significant perspectives on the matter at hand.

The interviewee posits that the factors contributing to inadequate data quality vary across multiple scenarios. Diverse entities may exhibit varying factors contributing to suboptimal data quality. In general, suboptimal data quality management can often be attributed to outdated technology, which suggests that the organization's technological infrastructure may be insufficient to carry out necessary tasks. Although the impact of outdated technologies cannot be denied, the lack of accountability is ultimately a more worrisome issue. The primary factor that poses significant challenges is the absence of accountable personnel for the management of data quality.

To be precise, within the industry of the interviewee's former employment, the CIO was not deemed accountable for ensuring contextual data quality. In the context of sales data, the responsibility for ensuring the quality of information lies with the sales manager, who may receive assistance from the CIO in addressing any technological issues that may arise. Nonetheless, the absence of a specifically assigned personnel accountable for the management of data quality and the lack of technological measures for data quality management are apparent. Consequently, although the data is used on a daily basis, its integrity cannot be guaranteed and is beyond scrutiny.

Identification of the underlying cause of substandard data quality necessitates a thorough analysis of the consequential effects on the data sources. As per the statements made by the interviewee, inadequate data sources can result in suboptimal decision-making, which can have significant ramifications in the production of the interviewee's previous organization's merchandise, specifically complete constructions. Decisions pertaining to construction are predicated upon the extant data, and the workforce depends upon said information to direct their labor. Inadequate data quality can compromise the accuracy of decisions, potentially causing errors in production processes and necessitating the re-manufacture of parts, thereby impacting the integrity of the building structure, and incurring additional expenses.

In the construction industry, inadequate data quality can result in significant costs, primarily due to the nature of the final products being buildings. The precision of data plays a pivotal role in ascertaining the necessary materials for the production process. For instance, the wrong measurement of steel bars required for the construction of a building could result in the production of an incorrect length, which could ultimately affect the entire building's structure. If the data indicates that an 8cm steel bar is required for a particular building part, but the actual required length is 10cm, then the construction of the building will not be correct. The resultant errors will necessitate the production of new parts and halt the installation process, leading to additional expenses. The interviewee mentioned another example of the cost

of poor data quality. His former company had machines that generated 10,000 pieces of quartz daily using input data. In the event of an error, the machines would produce the wrong type of quartz, incurring significant expenses and potential losses for the company in a single day. While decision-making is undoubtedly important in the manufacturing process, the accuracy of data cannot be understated. As such, it is vital to ensure the accuracy of data to prevent potential problems in the supply chain, production, and installation of products.

Ineffectual decision-making can result in a myriad of consequences that are often adverse. Upon soliciting input from interviewees, the author asked for a comprehensive depiction of the repercussions of poor data quality with regard to both general data analysis and data harmonization from the interviewee. However, due to the fact that the interviewee is more familiar with the business aspect than the technology aspect, the result tends to be more economical. He continued to use the provision of a specific illustration involving steel bars which were mentioned in the paragraph above. He pointed out several costs associated with poor data quality, including expenses linked to the production of new goods, the cost of transferring materials, the protraction of procedures, and customer reimbursement for tardiness in project delivery. The interviewees reported that when delays occurred due to poor data quality, their former company was required to compensate its customer base based on the stipulations outlined in the contract, highlighting the enormity of the cost of bad decision-making caused by poor data analysis.

The inclusion of an unstructured question was motivated by curiosity regarding the interviewee's ability to manage situations involving incomplete data in their work. As per the interviewee's perspective, the insufficiency of data is a frequent occurrence, necessitating the adoption of assumptions regarding numerical and conditional parameters in order to perform necessary calculations and mathematical operations.

One of the main objectives of this thesis is to explore the pragmatic implementation of data management and to scrutinize the viewpoints of professionals operating within this domain. To achieve this goal, two research questions have been developed to evaluate the extent to which organizations are aware of the importance of data quality and to investigate the prevalent practices in managing data quality. The two inquiries are designed to examine whether entities possess an internal data management division or contract the expertise of technology firms.

Regarding the initial issue, the interviewee noted that numerous contemporary organizations lack a comprehensive understanding of the importance of data quality. He contended that the significance of data quality has escalated as a result of the widespread adoption of digitalization, globalization, and automation. Historically, humans used to manufacture products manually, but in contemporary times, machines perform such tasks. However, machines lack human intelligence and are unable to detect errors in data the way humans can. For instance, if a human-produced a 10cm steel bar but data showed that it needed to be 8cm, they would recognize the problem and rectify it. However, machines follow input data and are unable to differentiate between 10cm and 8cm or recognize errors in data. Thus, it is imperative that data quality monitoring be given more attention. However, the awareness of data quality's importance is low, especially among small enterprises that lack sufficient knowledge about this field and have limited resources to invest in technological upgrades.

As for the second problem, the interviewee agreed that large firms often have their own data management departments, recognizing the importance of such a unit. For instance, in his former company, which was small, only two individuals worked in data management, while a large Finnish pharmaceutical company, Orion, had a data management team comprising 150 individuals. The interviewee opined that having a data quality management department was crucial, given that errors in data quality can occur in both physical products and machines. However, his former company did not have a data quality management department. He expressed dissatisfaction with this state of affairs, noting that industries such as healthcare and education, where data plays a critical role in activities, typically have good data quality management practices.

In conclusion, the insights provided by the interviewee were crucial in shedding light on the first theme, which was instrumental in addressing existing problems. The two primary reasons for poor data quality management were identified as outdated technology and inadequate data quality control. Inadequate technical capabilities and a lack of clear responsibility for data quality management have resulted in an unreliable data source, which can lead to erroneous decision-making in certain scenarios. Consequently, companies may incur significant losses, have to rectify errors, and compensate customers for contractual delays. Thus, the significance of data quality management cannot be overstated, especially in the current era of digitalization, globalization, and automation. Unfortunately, many firms, particularly small enterprises, are not prioritizing data quality management despite its vital importance. In contrast, larger enterprises tend to be more aware of the issue and may have dedicated data quality management teams.

Ultimately, the extent to which a company prioritizes data quality management is contingent upon its financial and technological capabilities.

5.2 Criteria for evaluating input data

For the purpose of ensuring the integrity of input data, a specific set of criteria is established, tailored to the particular business activities of an organization. As distinct departments within an organization require differing information types, data quality standards also vary. The purpose of a data assessment process is to assess the quality of data prior to its advancement to subsequent stages, to identify and rectify errors present in the data source, to minimize the likelihood of errors emerging in subsequent stages, and to optimize the utilization of final data analysis results. Given the close relationship between this theme and an organization's operating field, the criteria utilized in the data assessment process may vary based on the specific nature of the organization's activities, potentially limiting the scope of interviewee responses to a certain degree.

During an interview, the interviewee provided insights into the desirable characteristics of a good data source. According to the interviewee, a good data source should furnish users with information that meets their specific requirements and purposes. However, in the interviewee's case, data that was either extraneous or not vital to his needs was frequently collected. The interviewee underscored that the most crucial criterion for a good data source is that it provides the "right data" needed for a given purpose. The second important criterion is the ability to regulate the flow of data. The interviewee emphasized the importance of monitoring the data that enters the system and the ability to evaluate its quality. In essence, a good data source should not only provide the right data but also ensure that the data quality is maintained throughout its use.

In addition, considering the content in Section 2.3.2, currency is identified as a vital element when assessing data quality. Nonetheless, the interviewee held a different opinion on this matter. Drawing on his experience, he indicated that timeliness may not always be the most important factor. He recounted situations where he had invested considerable resources into attempting to leverage data for business intelligence or automation, yet the resulting output did not meet his expectations and was, therefore, unusable. Thus, he considered timeliness to be a relatively unimportant factor in assessing data quality.

The interviewee also expressed his viewpoint regarding out-of-date data, which is commonly viewed as obsolete and irrelevant. Contrary to widely accepted theories, he argued that old data still holds value. He based his argument on the premise that old information serves as a foundation for new knowledge. He further argued that the attributes of humans, such as personality, knowledge, and skills, are considered old data, and yet they are invaluable. Humans are an integral part of history; therefore, old data may prove useful in certain situations. The interviewee emphasized the importance of learning from history and how old data can provide useful insights that may inform current decisions.

Additionally, the interviewee indicated that, in general, data quality assessment presents a significant challenge for small companies operating in the manufacturing industry, such as his former company. This poses a question about the future viability of such companies, particularly whether they can effectively manage and assess data quality in the face of budgetary constraints, especially as automation and robotization continue to shape the industry. Companies that fail to prioritize data quality assessment will inevitably experience reduced efficiency and competitiveness, ultimately leading to their downfall. Therefore, organizations must take this issue more seriously and prioritize data quality assessment in their operations.

At the outset of the thesis, the rationale for selecting the significance of data quality in the data harmonization process as the topic was introduced. Consequently, during the interview, the problem was presented to solicit insights from an individual with relevant expertise. In his former occupation as the CIO, he shared that he had to process a substantial amount of data ranging from 10,000 to 20,000 rows daily. Given that these data served as input information for the manufacturing process, any errors in the data source could lead to significant challenges. However, considering the vast volume of data, evaluating the quality of all data points with human eyes is impractical. Therefore, automation support and mathematical tools are necessary to ensure data quality. Moreover, the interviewee acknowledged that even small companies handle substantial amounts of data daily, implying that larger enterprises are likely to require supercomputers to facilitate data quality assessment.

In summary, a reliable data source must furnish pertinent information suitable for the desired objectives while affording users the ability to regulate data flow. These criteria are vital because situations can arise where the collected data is inadequate, and uncontrolled data flow can lead to unfavorable consequences and result in substantial expenses for firms. Contrary to the popular notion, the interviewee did not regard Currency as a dispensable attribute. According to his perspective, the future is founded on the past;

hence, outdated data still holds value to some degree. Nevertheless, the interviewee noted that data quality assessment is an arduous task for companies and one that is frequently neglected due to financial constraints. Despite this challenge, companies must acknowledge the significance of data quality assessment and prioritize it in their operations. While small firms may struggle with their technical level issues, larger firms have gained a competitive edge by leveraging technology and supercomputers to manage vast amounts of data and maximize profits from data sources. In the long run, it depends a lot on the company's technological capabilities so that the company can gain comparative advantages.

5.3 The potential of AI in providing data immaculacy in the future

In the past, humans had to perform numerous tasks manually. However, with the advent of modern technology, machines have become capable of accomplishing a multitude of tasks, thereby saving time and effort that humans can redirect toward other pressing issues. This enhances efficiency and enables businesses to generate greater profits. Notably, the emergence of AI has played a significant role in supporting humans in several tasks, including those related to data management, which corporations have widely adopted globally. Nonetheless, the extent to which AI can assist organizations in improving data quality remains an open question that has piqued the curiosity of several individuals and organizations.

The implementation of AI can facilitate the automation of various procedures, including but not limited to data mapping, data cleaning, and data validation. The process of data mapping is widely acknowledged as a challenging task due to the presence of data elements that exhibit similar patterns across multiple sources, thereby leading to potential confusion among individuals involved in the mapping process. The utilization of machine learning can facilitate the automation of the aforementioned process through the identification and reconciliation of patterns and relationships. In addition, AI technologies have the capability to identify errors and inconsistencies within the data, thereby mitigating the need for extensive data cleansing and validation processes, ultimately resulting in time savings. Consequently, the assurance and enhancement of the quality, efficiency, and consistency of the final product are achieved.

The future of the labor market is a growing concern as AI gradually supplants humans in various tasks. Despite some knowledge barriers, the interviewee shared valuable insights on this subject. In evaluating

the impact of AI on organizations' data quality improvement, it is imperative to concentrate on the potential effectiveness of AI in this process. However, based on the interviewee's experience, organizations do not fully utilize AI's capabilities in this regard. Although AI is a valuable tool for assessing data quality, it may not be the optimal choice for some particular industries, such as the construction industry of the interviewee's former company, where each product is unique and independent from others, even if they have the same shape. Even when products have the same shape, they may still contain distinct details. The uniqueness of each product presents a significant challenge for the interviewee's former company, as each new order introduces entirely new data, making it difficult for AI to determine the quality of the information.

However, the interviewee's limited knowledge of this subject and the low level of technology in their former company prevented a thorough discussion. The application of AI to organizations differs depending on the product type and business background. In the interviewee's field of expertise, construction materials, and buildings, technological concepts such as Enterprise Resource Planning (ERP) systems and product data management are new, and the company is just beginning to learn and apply them. Additionally, the interviewee observed that his former company had a low level of automation in activities that could be done by machines, such as steel welding. Although he successfully made some changes before leaving the company, the process could not be entirely replaced by machines.

Even though the interviewee advocated for the automation of business activities, he emphasized that machines cannot think like humans and cannot identify abnormal situations without human intervention. Therefore, the effective and smooth operation of machines relies on the utilization of AI and similar tools for data quality management. Failure to adopt these tools may have detrimental consequences for the organization.

Since each organization has a different technological level, it is very hard to tell the overall picture. The interviewee observed various cases but could not give a general comment. However, if it comes to a specific case, taking his former company and the construction industry as an example, he could tell that his company just started with a low level of applying technology in some activities, which was a drawback if the competitors had already been at a higher level, using machines for tasks which a human did in his former company. It is obvious that his former company would lose the battle in such a situation, so they must develop themselves, or else they would be eliminated.

Determining the level of technological development across different organizations can be challenging due to their varying degrees of advancement. Although the interviewee has observed various cases, a general comment on this matter is difficult to provide. However, when considering a specific case, such as the interviewee's former company in the construction industry, he noted that it had only recently begun to adopt technology in certain activities, indicating a disadvantage compared to competitors who have already implemented machine-based solutions to tasks that were carried out by humans in his former company at the same time. This disadvantage in technology adoption brought significant consequences to his former company, put it in a drawback situation, and potentially led it to be eliminated. It is, therefore, imperative that companies should continuously develop themselves to keep pace with technological advancements in the industry.

The interviewee acknowledged that while AI is useful in detecting errors, there are certain types of errors that it cannot identify, which necessitate human intervention. This position was underscored by an example provided by the interviewee with the assumed background was the sales department in his former company. The sales team had weekly meetings in their former company to assess progress and strategize for upcoming activities. To aid in these efforts, a Customer Relationship Management system was utilized to store all sales data and track the sales process. The interviewee assumed that he was the team leader and had set a target of 1.5 million euros. During one of the meetings, it was reported that the target had not been met. However, one of the salesmen said that he had an upcoming deal in the works but had forgotten to enter the information into the system. Such an error originated from human activity and could not have been detected by AI. Given that humans are an integral part of such systems, managing human-created errors can be a daunting task. While the interviewee believed that AI might have the potential to address such problems, he could not confirm this possibility owing to his limited experience.

From this basis, the interviewee opined that AI could not replace humans completely, particularly in the realm of the data improvement process, where a human must take a leadership role. There must always be someone responsible for overseeing business-related activities. While some aspects of the data improvement process can be fully automated, AI cannot replace humans in the entire process. Humans play a crucial role in ensuring the quality of data management. Regarding data quality management, the interviewee recognized that AI could handle a significant portion of the process, although the degree of automation depends on how much humans can automate the data cleaning process.

Additionally, the interviewee expressed his positive thoughts toward the future of AI assisting humans in data-handling processes. He thinks that with current technology and its potential, it is difficult to predict how far it will develop in the future. All scenarios are possible, and we cannot predict anything. Most of the time, people think they cannot do it, but a few years later, they look back and realize that they succeeded.

At the end of the interview, an additional question was placed with the aim of obtaining the interviewee's perspective on a current, popular AI tool, namely ChatGPT. Developed by OpenAI, ChatGPT is a language model system that uses deep learning techniques to generate textual responses and has been utilized in various situations, such as chatbots and other content-creation tasks (Ortiz 2023.) The question was predicated on the potential for ChatGPT to supplant humans in decision-making. However, the interviewee expressed a contrary view as he maintained that ChatGPT could only calculate predictions based on the data it had been fed and was limited in its capacity to generate new ideas beyond its collected data. In contrast, the interviewee highlighted the unlimited creativity of the human mind, that humans could innovate and are capable of producing new and diverse possibilities, whereas AI is limited to a set of pre-existing data. He contended that AI could only execute a small portion of what humans were capable of doing. If ChatGPT were to replace humans in decision-making, the interviewee posited that everyone would produce identical results, leading to a scenario without any competition or winners, which was deemed to be undesirable. Therefore, the interviewee concluded that ChatGPT was not a suitable substitute for human decision-making.

The topic of AI has gained significant traction in recent years, with numerous discussions surrounding its potential applications. In summary, the interviewee held a positive view regarding AI's potential to assist humans in improving data quality. While firms are still in the early stages of implementing AI, their utilization of this technology has yet to reach maximum efficiency. However, AI may not be a suitable solution for every industry, such as the construction industry, which deals with unique and ever-changing products. Constructing buildings to meet varying customer requirements makes it difficult to the application of AI in the manufacturing process. It is merely impossible to have different customers ordering buildings with the same requirements about buildings and their specifications. Nonetheless, AI can replace humans to a certain extent in the data cleaning process, detecting various issues except the ones from humans. Thus, human intervention remains necessary for this process. Regarding data harmonization, there is a positive outlook that AI can perform various steps in different cases, albeit not in all scenarios. At the end of the interview, the possibility of AI making decisions instead of humans was

discussed, and it was noted that AI lacked the ability to generate new ideas or inventions, illustrating the importance of human creativity.

5.4 Findings

Once information and perspectives have been gathered from an interviewee, it is crucial to compare the findings from both primary and secondary sources to differentiate between sources and determine the disparities between theoretical and practical aspects of data quality. The first and second themes primarily contain differing viewpoints, while the third theme seeks to expand knowledge about the role of AI in improving the data-cleaning process.

Section 2.3 of the thesis discussed four crucial quality dimensions: Completeness, Timeliness, Accuracy, and Consistency. Upon conducting the interview, it became apparent that the interviewee's responses indirectly encompassed only the dimensions of completeness, accuracy, and consistency. It is widely acknowledged that data sources must be adequate, accurate, and consistent to ensure reliable information delivery to users. Even a single error within these dimensions can lead to the dissemination of misleading information, potentially resulting in adverse consequences for decision-making processes. During the interview, the interviewee provided examples of situations where inaccuracies led to unintended consequences.

In contrast, it is noteworthy that issues related to the quality dimensions may occur frequently, as exemplified by the interviewee's situation. Due to the specific nature of his work, he frequently encounters situations with insufficient information, and in order to mitigate this problem, he must make assumptions about every missing piece of information. This action necessitates extensive calculations and proficiency in mathematics to minimize the occurrence of false numbers and to ensure the practical application of the results. Data from previous products and related sources must be carefully studied to achieve this. It is believed that this is one of the fundamental reasons for the interviewee's opposing viewpoint regarding the Currency element of the Timeliness dimension. In contrast to the majority of frameworks on data quality, the interviewee asserts that old data still holds value to some extent, as humans continue to learn from history every day. While this view may not be aligned with most theories, it can be applicable in certain cases.

Moreover, the interviewee's criterion for a reliable data quality source, which is the ability to regulate the data flow, is also noteworthy. This criterion holds significant importance, as being unable to control the data quality may result in potential errors in the data source, which may go undetected in the early stages and continue to flow through the data into later stages, leading to more troublesome issues. Additionally, having control over the data flow enables the identification of areas where the process can be improved, thereby enhancing the performance of the data quality management process and facilitating consistent monitoring of data quality.

The adverse impact of inadequate data quality on organizations should not be underestimated, as it can lead to escalated expenses and harm to their reputation. The production of defective products by a company results in supplementary costs associated with remanufacturing, re-delivery, and replacement, ultimately leading to a reduction in revenue. Moreover, delayed delivery or inaccurate product delivery adversely affects the reputation of the company, leading to a detrimental impact on business relationships and a reduction in credibility.

Organizations appear to have an inadequate grasp of the crucial role that data quality plays in data management, despite the substantial financial consequences that result from subpar data quality. Consequently, the evaluation of data quality is frequently disregarded and fails to receive the requisite level of consideration. The main determinants that are responsible for this situation are the sphere of execution, financial constraints, and the technological proficiencies of the entity. Differences in technological proficiency are observable across firms operating in diverse sectors. IT companies demonstrate a higher degree of technological sophistication compared to construction companies.

Similarly, small and medium-sized enterprises face limitations in accessing technological resources when compared to their larger counterparts. Financial limitations pose a challenge as the implementation of technology and automation incurs significant expenses for companies. The protracted nature of the process and its adaptive challenges instil fear of change among corporate leaders. Acquiring new knowledge and skills can pose a formidable obstacle for certain organizations, as they may harbor reservations regarding the efficacy of the outcome. Nevertheless, we currently exist within a period characterized by the proliferation of technology, globalization, and automation. A company that relies solely on manual and traditional operations is likely to be outcompeted by a modern enterprise that leverages high-tech machinery across its various processes. Moreover, the foremost issue is that the enterprise does not perceive the inadequate data quality as a crucial challenge to its business operations. The action

taken can be classified as a significant error. It is imperative for companies to prioritize this issue in order to remain competitive in the long term.

Despite the significant financial and reputational risks associated with poor data quality, many companies do not recognize its importance and fail to prioritize data quality assessments. This can be attributed to several factors, including industry-specific challenges, financial constraints, and technological limitations. Companies operating in different industries may have varying levels of technological sophistication, with construction companies generally operating at a lower-tech level than IT companies. At the same time, small and medium enterprises may face more limited access to technological resources compared to large enterprises. Furthermore, financial constraints are an issue as technological solutions and automation can be expensive. Companies may hesitate to embrace change due to concerns about time, adaptation, and satisfaction with the results.

However, in an era of increasing technological, global, and robotic advancements, companies that continue to operate manually and in a traditional way will inevitably fall behind competitors with more advanced technological capabilities. Therefore, it is crucial that companies view poor data quality as a critical problem in their business operations and take appropriate steps to address it, or else they will fall behind their competitors in the long run.

In terms of the potential applications of AI in the data cleaning process, it is evident that some firms have already integrated AI into their business operations, while others are still in the early stages of exploring and implementing such tools. Generally speaking, large enterprises with dedicated data management teams are more likely to have the capacity to optimize these technologies. However, it is important to consider the specific context of each company when assessing the potential usefulness of AI for their business needs, as not all situations may benefit from AI integration.

Take the interviewee's former company as an example. It dealt with unique building projects that required calculations and decisions to be made on a case-by-case basis. In such a scenario, the complexity of the task may limit the usefulness of AI integration, at least for the time being. Despite this, the interviewee expressed optimism about the future of AI and its ability to handle increasingly complex tasks. As the limits of human invention and imagination are unknown, the possibilities for AI are likewise limitless.

The author concurs with the interviewee's perspective on this matter. While AI can replace humans to a certain extent in the data cleaning process and can identify various issues except those caused by human error, human interference is still necessary. In terms of data harmonization, there is potential for AI to perform various steps in different scenarios, although not all situations may benefit from such integration. It is impossible to predict with certainty whether AI will completely replace human intervention, as there are always unforeseen possibilities that may arise. Therefore, it is important to approach AI in data cleaning with an open mind and consider each situation on a case-by-case basis. It can not be said that AI will totally replace humans because everything is possible and impossible at the same time, and it is impossible to ensure an occasion that will not have happened yet.

6 CONCLUSIONS

The aims of this research endeavor are to examine the standard of data, establish uniformity in data harmonization, and assess the importance of data quality in the comprehensive data management process, as well as in the particular data harmonization process. In modern business practices, it is imperative for organizations to acknowledge the importance of managing data quality. The author of this thesis aims to present a thorough and inclusive examination of the subject matter and act as a reference for individuals who hold a vested interest in the domains of data quality and data harmonization. AI has gained considerable attention in recent times owing to its capacity to aid humans in diverse undertakings. Therefore, the current research investigates the potential of employing AI to streamline the data cleansing procedure, with the objective of achieving effective data standardization.

Undoubtedly, data stands out as one of the most crucial assets for any enterprise. The provision of valuable information enables companies to effectively manage the situation, formulate strategies, and capitalize on opportunities. Possessing a comprehensive comprehension of a company's circumstances and being privy to more information regarding the current market than one's competitors will invariably confer comparative advantages upon the company. In the event that the data acquired by a company is erroneous, it may result in a disadvantageous outcome rather than a beneficial one. Consequently, ensuring the quality of data holds significant importance for the organization.

Ensuring high-quality data sources is a critical aspect of data collection for subsequent stages. When dealing with a single data source, the process of error identification is relatively simple and the resulting impact of such errors is minimal. When dealing with multiple data sources that originate from distinct origins, any errors that occur within one of the sources can lead to significant complications. Moreover, it is possible for data sources to comprise of errors that could pose a challenge in terms of identifying them before proceeding with the data integration procedure. Possible challenges that could emerge encompass syntactic and semantic inaccuracies, which might hinder the harmonization procedure or generate imprecise outcomes. Inadequate data quality at this level can pose a significant obstacle to the entire process, given the extensive volume of data that requires scrutiny, leading to a laborious and costly undertaking. Moreover, there exists the possibility to nullify the result and initiate the task again.

The thesis is composed of three discrete phases. The initial phase entails the gathering, examination, and amalgamation of diverse secondary information sources pertaining to the dimension of data quality, the assessment of data quality, and the influence of data quality on the harmonization of data. During the ensuing phase, an interview is carried out with a former CIO of an organization to acquire supplementary understanding regarding the subject matter. Ultimately, the results obtained from two distinct sources are amalgamated to deduce overarching conclusions that encapsulate the research findings. Ultimately, the investigation undertaken effectively responds to all of the queries presented in the introductory segment of the dissertation.

Data quality has a notable influence on the overall data management procedure and the specific data harmonization process, leading to suboptimal decision-making as the primary consequence. The existence of various inaccuracies such as redundancies, omissions, syntax discrepancies, and logical incongruities in low-quality data presents considerable obstacles to the process of harmonization and may lead to unfavorable outcomes in decision-making. Obsolete technology and inadequate personnel supervision can give rise to additional factors. Hence, in situations where complications arise, a dearth of responsibility and inadequate technological proficiency exacerbate the prevalence of errors in the quality of data sources.

The quality of data is often compromised by incomplete, inaccurate, and inconsistent information. To evaluate the input data, the assessment criteria are classified into three dimensions, namely Completeness, Accuracy, and Consistency. Then, these dimensions are combined into a singular criterion, namely, the attainment of "accurate data." The information contained in the data resulting from the assessment process must be sufficient. For the purpose of enhancing quality control, data intended for the data harmonization stage must exhibit characteristics that enable users to monitor its progression. Additionally, the aforementioned criteria serve the purpose of ensuring the cleanliness of input data and streamlining subsequent processes in order to prevent unforeseen and expensive scenarios.

The integration of AI is imperative in enhancing the efficacy of the data cleansing procedure. The potential for AI assistance in ensuring data purity is poised for growth, contingent upon human ingenuity in the realm of innovation. Nevertheless, it is important to note that the complete replacement of humans in the data harmonization process by AI is unlikely, given the indispensable role that humans play in this process. It is noteworthy that certain scenarios present difficulties in implementing AI for data cleaning, owing to the intricate nature of the prerequisites. In conclusion, the utilization of AI in the data

cleansing procedure has the potential to exceed human performance, albeit its efficacy is contingent upon the specific circumstances. Therefore, human intervention is indispensable to assess the feasibility of AI's optimal utilization in each case.

Throughout the process of formulating this dissertation, the author acquired perspectives that transcend the confines of scholarly understanding. Her expertise in the domains of data quality and data harmonization is undoubtedly a result of her keen interest in these areas. Moreover, the author's acquired insights exceed the mere acquisition of knowledge. The application of qualitative research methodology allows a researcher to examine a subject matter from multiple angles, going beyond superficial observations to reveal underlying contextual elements. This methodology fosters the development of critical thinking abilities through the requirement of evaluating alternative perspectives and potential constraints of suggested resolutions. Through the process of scaling the problem and examining contrasting possibilities, the author can acquire a more all-encompassing comprehension of the matter at hand and pinpoint potential areas for enhancement. The objective of the thesis is to furnish significant insights not only to the author but also to other individuals who are in search of pertinent resources concerning data quality. In its entirety, the thesis has not only facilitated the author's individual development but also endeavors to serve as a resource for individuals pursuing knowledge in this particular domain.

REFERENCES

Adhikari, K., Patten, S. B., Patel, A. B., Premji, S., Tough, S., Letourneau, N., Giesbrecht, G. & Metcalfe, A. 2021. Data Harmonization and Data Pooling from Cohort Studies: A Practical Approach for Data Management. *International Journal of Population Data Science* 6(1), 1680. Available at: https://doi.org/https://doi.org/10.23889/ijpds.v6i1.1680. Accessed 28 February 2023.

Alizamini, F. G., Pedram, M. M., Alishahi, M. & Badie, K. 2010. Data quality improvement using fuzzy association rules. *2010 International Conference on Electronics and Information Engineering* 1, V1-468–V1-472. Available at: https://doi.org/10.1109/ICEIE.2010.5559676. Accessed 8 March 2023.

Anodot. 2019. *The Price You Pay for Poor Data Quality*. Available at: https://www.anodot.com/blog/price-pay-poor-data-quality/. Accessed 26 February 2023.

Atzeni, P. & Antonellis, V. D. 1993. *Relational Database Theory*. Massachusetts: Benjamin-Cummings Pub Co.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys* 41(3), 1–52. Available at: https://doi.org/10.1145/1541880.1541883. Accessed 26 February 26 2023.

Cambridge. No date. *Meaning of data in English*. Available at: https://dictionary.cambridge.org/dictionary/english/data. Accessed 25 February 2023.

Chapman, D. 2022. *Qualitative research: What is it and why should you use it?* Available at: https://mr.onepoll.com/qualitative-research-what-is-it-and-why-should-you-use-it. Accessed 12 February 2023.

Cichy, C. & Rass, S. 2019. An Overview of Data Quality Frameworks. *IEEE Access 7*, 24634–24648. Available at doi: https://doi.org/10.1109/ACCESS.2019.2899751. Accessed 25 February 2023.

Cowart, J. A. 2017. *Information For Students: What is the difference between Primary and Secondary sources?* Available at: https://libguides.furman.edu/special-collections/for-students/primary-secondary-sources#:~:text=Primary%20sources%20can%20be%20described,sources%20and%20often%20interpret%20them. Accessed 04 March 2023.

Cuesta, H. 2016. *Practical Data Analysis*. Second edition. Birmingham: Packt Publishing Ltd.

De Mauro, A., Greco, M. & Grimaldi, M. 2015. What is Big Data? A Consensual Definition and a Review of Key Research Topics. *AIP Conference Proceedings* 1644(1), 97–104. Available at: https://doi.org/10.1063/1.4907823. Accessed 04 March 2023.

Del Pilar Angeles, M. & Garcia-Ugalde, F. 2009. A Data Quality Practical Approach. *International Journal on Advances in Software* 2, 259–273. Available at: http://www.iariajournals.org/software/. Accessed 28 February 2023.

Fortier, I., Doiron, D., Burton, P. & Raina, P. 2011. Invited commentary: consolidating data harmonization–how to obtain quality and applicability? *American journal of epidemiology* 174(3), 261–264. Available at: https://doi.org/10.1093/aje/kwr194. Accessed 4 March 2023.

Fox, C., Levitin, A. & Redman, T. 1994. The notion of data and its quality dimensions. *Information Processing & Management* 30(1), 9–19. Available at: https://doi.org/10.1016/0306-4573(94)90020-5. Accessed 27 February 2023.

Gurugubelli, V. S., Fang, H., Shikany, J. M., Balkus, S. V., Rumbut, J., Ngo, H., Wang, H., Allison, J. J. & Steffen, L. M. 2022. A review of harmonization methods for studying dietary patterns. *Smart Health 23*(100263). Available at: https://doi.org/10.1016/j.smhl.2021.100263. Accessed 21 April 2023.

Mcleod, S. 2023. *Qualitative vs Quantitative Research: Differences, Examples & Methods*. Available at: https://simplypsychology.org/qualitative-quantitative.html. Accessed 11 March 2023.

Moore, S. 2018. *How to Create a Business Case for Data Quality Improvement*. Available at: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement. Accessed 25 February 2023.

Oliveira, P., Rodrigues, F. & Rangel Henriques, P. 2005. A Formal Definition of Data Quality Problems. *MIT International Conference on Information Quality.* Available at: http://mitiq.mit.edu. Accessed 8 March 2023.

Ortiz, S. 2023. *What is ChatGPT and why does it matter? Here's what you need to know*. Available at: https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know. Accessed 7 April 2023.

Pollock, T. 2019. *The Difference Between Structured, Unstructured & Semi-Structured Interviews*. Available at: https://www.oliverparks.com/blog-news/the-difference-between-structured-unstructured-amp-semi-structured-interviews. Accessed 11 March 2023.

Przybycień, G. 2023. *Data architecture strategy for data quality*. Available at: https://www.ibm.com/blogs/journey-to-ai/2023/01/data-architecture-strategy-for-data-quality. Accessed 10 March 2023.

Redman, T. C. 2016. *Bad Data Costs the U.S. $3 Trillion Per Year*. Available at: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year. Accessed 25 February 2023.

Reno, G. 2022. *The Role of ML and AI in Data Quality Management*. Available at: https://firsteigen.com/blog/the-role-of-ml-and-ai-in-data-quality-management. Accessed 10 March 2023.

Rolland, B., Reid, S., Stelling, D., Warnick, G., Thornquist, M., Feng, Z. & Potter, J. D. 2015. Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach. *American Journal of Epidemiology* 182(12), 1033–1038. Available at: https://doi.org/10.1093/aje/kwv133. Accessed 21 April 2023.

Scannapieco, M. & Catarci, T. 2002. Data quality under a computer science perspective. *Journal of The ACM - JACM* 2 1–10. Available at: https://www.researchgate.net/publication/228597426_Data_quality_under_a_computer_science_perspective. Accessed 26 February 2023.

Scannapieco, M., Missier, P. & Batini, C. 2005. Data Quality at a Glance. *Datenbank-Spektrum 14*, 6–14. Available at: https://www.researchgate.net/publication/220102773_Data_Quality_at_a_Glance. Accessed 25 February 2023.

Sidi, F., Hassany Shariat Panahy, P., Affendey, L., A. Jabar, M., Ibrahim, H. & Mustapha, A. 2012. Data quality: A survey of data quality dimensions. *2012 International Conference on Information Retrieval & Knowledge Management*, 300-304. Available at: https://doi.org/10.1109/InfRKM.2012.6204995. Accessed 26 February 2023.

Stedman, C. 2022. *What is data management and why is it important?* Available at: https://www.techtarget.com/searchdatamanagement/definition/data-management. Accessed 21 April 2023.

UNTDED-ISO7372. The United Nations Trade Data Element Directory. 2005. Available at: https://unece.org/untded-iso7372. Accessed 7 March 2023.

Truyện cổ tích. No date. *Thầy bói xem voi (Blind fortune tellers view an elephant)*. Available at: https://truyencotich.top/doc-truyen/thay-boi-xem-voi. Accessed 25 February 2023.

Tsichritzis, D. C. & Lochovsky, F. H. 1982. *Data Models.* Englewood Cliffs , N.J: Prentice-Hall.

UNNexT, ESCAP, UNECE. 2012. *Data Harmonization and Modelling Guide for Single Window Environment.* Thailand: United Nations publication. Available at: https://www.unescap.org/resources/data-harmonization-and-modelling-guide-single-windows-environment. Accessed 04 March 2023.

Wang, R. Y. 1998. A product perspective on total data quality management. *Communication of the ACM* 41(2), 58–65. Available at: https://doi.org/10.1145/269012.269022. Accessed 26 February 2023.

World Customs Organization. 2017. *WCO Single Window Compendium.* Available at: https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/tools/single-window/compendium/swcompendiumvol2partv.pdf. Accessed 7 March 2023.