Data Quality in Artificial Intelligence

Mikko Haukkala

Haaga-Helia University of Applied Sciences
Degree Programme in International Business Management
Master's Thesis
2022

Abstract

Author(s)
Mikko Haukkala

Degree
Master of Business Administration

Report/thesis title
Data Quality in Artificial Intelligence

Number of pages and appendix pages
49 + 2

This thesis is part of the AI-TIE project coordinated by Haaga-Helia University of Applied Sciences. The main goal of the project is to support SME companies in developing and growing their business in Finland by utilizing artificial intelligence solutions. The aim of the thesis, which was carried out in 2022, is to study the importance of data quality in AI development, examine the dimensions of data quality and to find out the common problems and good practices affecting data quality in companies that are already using or planning to implement artificial intelligence.

The theory section explains what is meant by artificial intelligence and what good data quality means from the perspective of artificial intelligence. In addition, the study explores what data is and how data quality can be measured and evaluated. By examining and comparing methods, the body of the interview and survey conducted in the research is selected.

The research part of the thesis utilizes the means of concurrent mixed method research. Based on interviews and surveys, the research section examines the views of professionals in the field on the different dimensions of data quality and the related challenges and good practices from the perspective of AI development.

Based on the results of the study, relevancy was considered the most challenging dimension of data quality in AI development. This dimension was selected as one of the most challenging data quality dimensions six times out of seven surveys. The reasons given for the challenging dimension included the difficulty of predicting what kind of data should be collected for future needs and a sufficient contextual understanding of the business and its needs. A comprehensive understanding of business problems from a technical and business perspective was considered important to be able to start collecting relevant data. In addition, the study revealed dimension-specific development suggestions and good practices for improving each data quality dimension.

The results of the thesis can be used to improve and evaluate the quality of existing data and to support the planning of future data needs from the perspective of artificial intelligence. In addition, the results can be utilized in the development of the maturity model of data quality on the way to the implementation of a production-ready AI application.

Keywords
Artificial Intelligence, Data Quality, Data Assessment

Table of contents

1 Introduction

Artificial Intelligence (AI) is one of the biggest transformations during our lifetime. The benefits of AI are undeniable in many different business areas. The new technology introduced early may create a decisive competitive advantage over competitors, which can be challenging to catch up later (Alho et al. 2018, 1-4).

AI's main benefits today are to assist people in tasks and to streamline and automate processes. The focus of digitalization is strongly on data and its utilization for business needs. As data fuels AI applications, organization-wide understanding of the importance of data and adoption data management practices as part of daily work can at best lead to innovation and a clear competitive advantage (TEM 2017, 19-21).

AI has rapidly conquered the IT industry in the 21st century and it has grown in many other industries as well. Examples like Siri - a personal virtual assistant on Apple’s mobile phones, analytics software like Google Analytics and Search Engine Optimization provided by Facebook and Google are all good examples of the use of AI (Rouhiainen 2018, 2-18).

There are already thousands of companies in Finland that develop and utilize AI in their business operations (FAIA 2018). The number is even higher if we include companies that have carried out only proof of concept type of experiments to utilize AI. According to a survey conducted by Microsoft for Finnish companies, 80% of the companies said the data used was not mature enough for AI applications (Microsoft 2018, 30-48).

This thesis concentrates on data quality from the perspective of AI development. Good understanding of data quality and data management play a key role when considering the readiness of organizations to apply AI and to scale the use of AI more widely. Possible risks caused by the poor data quality in AI development can be better managed through more mature data management processes. Organizations should be able to assess what level their data quality is and how data quality should be developed both before and during AI development. An adequate level of data quality helps to ensure the development of a production-ready AI solution and the value it brings to the business.

The purpose of this thesis is to examine the challenges of data quality in the development of AI and to increase understanding of what good data quality is from the perspective of AI. The purpose is also to increase understanding on how Finnish companies can ensure better data quality for the use of AI solutions. The research aims to produce development proposals that enable

organizations to improve data quality and advance in the development of AI towards a production-capable AI solution.

The thesis is a part of the AI-TIE project coordinated by Haaga-Helia University of Applied Sciences. The main objective of the AI-TIE project is to support small and medium-sized companies in developing and growing their business by utilizing AI solutions in innovation work and service development phase as well as in selling and delivering products and services to customers (Haaga-Helia 2018).

1.1 Objectives and scope

The purpose of this thesis is to find out what is good data quality, how data quality is measured, and which data quality dimensions most often lead to weak quality data in AI development. Based on the results obtained, development measures will be presented to ensure better data quality and therefore better opportunities for successful development of the AI application.

Research questions will be answered both in the empirical section and in the theoretical section of the thesis. Chapters 2.2, 2.3, 2.4 and 2.5 in the theoretical section present the definitions of data quality, how data quality is measured, how data quality is assessed and how data quality can be improved. The empirical section clarifies the recurring problems of data quality in AI development and the means to avoid them by interviewing professionals in the field. Research question 1 will be answered in the theoretical part of the work, while research questions 2 and 3 will be answered by interviewing professionals.

Artificial Intelligence is a very broad topic, and it is not possible to cover all aspects in one thesis. This work focuses on general data quality aspects in AI development and what areas should be considered in data quality to ensure expected outcomes. The study does not carefully address other aspects of data management in AI even if it touches data management on a high level.

The thesis examines the topic with the following research questions:

Q1: What is good data quality in AI development?

Q2: What are the most recurring problems within data quality in AI?

Q3: How to avoid the most frequent problems in data quality?

1.2 Research Methodology

The methodological choices of the research are from philosophies of science. The research philosophy used is pragmatism, which combines different perspectives in the data collection and in data interpretation (Saunders 2019, 130). In pragmatism, data can be collected through combined or multi-methodical ways, and it is characterized by practical orientation and the production of new information in the conduct of research (Saunders 2019, 148-149). The research approach is inductive, i.e., data-driven, as the results of the study are formed based on the collected data (Tuomi et al. 2018, 56-76). The research strategy is formed using the substance theories of data quality and the theoretical frameworks developed for its assessment. The used research approach of the thesis is case study, where the goal is to gain more in-depth understanding of a studied subject. Case study aims to provide detailed and intensive information on the chosen topic, but it does not necessarily aim to develop anything concrete other than provide ideas and development suggestions. The time horizon of the research is cross sectional, which is not primarily interested in change, but in phenomena and situations at a given time. For collecting data, concurrent mixed methods are used which involve the separate use of quantitative and qualitative methods within a single phase of data collection. The analysis of the data utilizes classification of content and benchmarking, which can be measured quantitatively and using generalization (Saunders 2019, 149). The theory of the research is formed from key scientific articles and books on data quality, its measurement and assessment. The material used for the theoretical sections has mainly been retrieved from FINNA.fi -library system and Google Scholar search service. The semi-structured interview conducted in this thesis has been built using well-known data quality assessment frameworks.

Table 1. Research choices

Research perspectives	*Research choises*
Philosophy of science	Pragmatism
Research approach	Inductive
Research strategy	Case study
Research method	Concurrent mixed method research
Time horizon	Cross sectional
Data collection	Concurrent mixed method

Data analysis	Content classification, benchmarking, generalization

1.3 Terminology

Data quality and data quality assessment have been studied for over 30 years. The terminology used in the literature varies slightly depending on the sources. Below are the most important terms and their explanations in context of this work.

The term *measurement* is used for the process which measures the quality of data along relevant data quality dimensions. In other words, measuring is used to measure the value of defined data quality dimensions (Batini et al. 2016, 50-52).

Data quality dimensions are measured using individual metrics and indicators. E.g., the dimension *security* can be measured using a metric *number of failed login attempts* or a survey conducted to the data users (Batini et al. 2016, 8-13).

The term *assessment* is used when measured values, answers or other types of outputs are compared to reference data or other values to examine the data quality. Assessment methodologies set the steps used in the assessment (Batini et al. 2016, 18).

In this thesis, the quality of data is assessed using qualitative and quantitative methods. The thesis examines the quality of data through quantitative survey and refines research through qualitative interviews, based on models and reference frameworks selected from literature. More information on research methods is given in later chapters of this thesis.

2 Theoretical Framework

To be able to answer the research questions of this thesis, it is necessary to understand what artificial intelligence is and how data and its quality are defined in this context. In addition, it is necessary to understand how data quality can be measured and evaluated, what are the tools and methods used to assess data quality and how data quality can be improved. The theoretical part of the thesis proceeds as described below.

First, it is important to understand what artificial intelligence is. As AI is a broad concept, it is necessary to understand the essentials to be able to investigate what is needed for a good and beneficial AI solution. A wide range of useful literature is available on the subject, which can be used to understand what is meant by the term.

Secondly, since data is a key topic of this thesis, it is worth clarifying what data is and how the quality of data can be measured. The thesis introduces the existing measuring methods of data quality and explains how data quality can be measured.

Next, the thesis introduces how data quality can be evaluated and presents the typical evaluation methods and the differences between them.

Finally, the theoretical part of the thesis explains how data quality can be improved by clarifying the methods used for improving data quality.

2.1 What is Artificial Intelligence

Artificial intelligence is an area of computer science where machines operate and learn independently without constant human involvement. AI is sometimes referred to as machine intelligence, machine learning and deep learning. However, these terms are components of AI, but not direct synonyms for artificial intelligence. Terms and names are sometimes misused in the general debate, which can cause confusion among different stakeholders. However, AI consists mainly of statistics, programming, and mathematics, but at its core is data (Marr 2020, 13).

According to another definition, AI includes technologies like decision support systems, expert systems, knowledge-based systems, agent-based systems, machine learning, neural networks, deep learning, natural language processing, robotics, autonomous systems, machine vision and other human intelligence applications such as cognitive science and biology (Håkansson & Hartung 2020, 14).

In any case, compared to humans, AI can solve complex and multidimensional mathematical problems quickly but is unable to understand and decide in the same way humans do. AI is not aware

of its action. It does not understand or think what it is doing, and the machines do not realize the connections or consequences of things (Kananen et al., 2019, 20-23).

Sometimes, a statistical or analytics solution is also called artificial intelligence. AI researchers also debate the definition. The definition of AI is also constantly changing, as some aspects are no longer considered to be AI as they become more common. Also new aspects are added under the definition. E.g., route optimization systems have been considered as AI in the past but are now so common that they are considered as part of the fundamentals of computer science (MinnaLearn & University of Helsinki 2018).

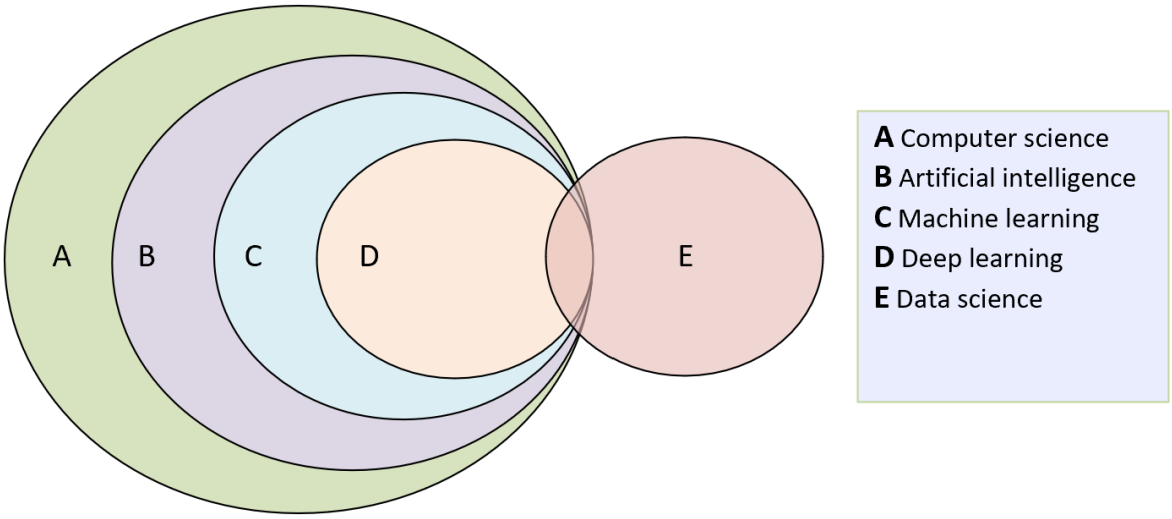

Figure 1. Classification of AI (MinnaLearn & University of Helsinki 2019)

The AI used today is often classified as weak AI, meaning that it is only able to operate effectively in a well-defined environment. Another weak point in today's AI-solutions is the large amount of training data it requires. Learning is based on statistical data analysis, which requires a large amount of data (Alho et al. 2018, 6-10).

2.1.1 The Importance of Data and Data Quality in Artificial Intelligence

Data is an asset that can be, e.g., numbers, text, images, or videos. Training AI usually requires a large amount of good quality, unambiguous, and consistent data. The amount of data required to train the AI solution depends on the complexity of the problem. The data type should be selected

according to the problem, such as images for image recognition (Kananen & Puolitaival 2019, 146).

Because the function of AI-solution is based on the data it receives as an input, AI development strongly involves the management of data quality. The underlying data of the solution must be fit for purpose, both in terms of content and quality, so ensuring data quality is a critical part of AI development (Combs 2021).

2.1.2 Artificial Intelligence in Finland

In 2017, Risto Siilasmaa, the founder of Finnish IT-security company F-secure and the former chairman of Nokia, urged companies in Finland to start using and developing AI capabilities. He justified this by saying that AI would completely change the business and therefore the journey developing AI capabilities should be started. In addition, he said that companies should start to think about what kind of data they will need in the future to be able to teach their AI solutions better than their competitors. In the same article, Siilasmaa states that in 5–10 years, AI may be crucial for Finland's global business and export industry. How we succeed depends a lot on what kind of opportunities Finland builds by utilizing AI capabilities within in its companies (Lähteenmäki 2017, 24-31).

Organizations in Finland still use AI capabilities mainly for experiments and learning. The responsibility of AI projects mostly lies within the small groups of experts, and only a few organizations have a clear structured approach on how successful solutions are developed and deployed more widely within the organization. However, the most successful projects already provide real business value to organizations of Finland (Microsoft 2018, 26-40).

AI is forecast to increase Finland's GDP by several billion in the next few years. However, the realization will require substantial investments in AI capabilities and the widespread use of AI in all industries. According to a study, Finnish organizations are still in the early stages of this journey (Microsoft 2018, 2-6).

2.2 Data Quality

The base of competition between companies has changed from tangible physical products to intangible data and information. The data represents the collective information used to produce and deliver products and services to consumers. The quality of data is increasingly recognized as the company’s most valuable asset (McGilvray 2008, 352). However, data is a very ambiguous concept because it can mean data, information, and knowledge.

When an organization wants business value out of its data, the data must be accurate. There is no generally accepted definition of data quality, but the prevailing view is that data is good quality when it fits for use (Sebastian-Coleman 2013, 40). Good quality data can be invalid for another use (Tayi & Ballou 1998, 54-56). This can make it difficult to assess data quality, as different use cases may have very different data requirements.

The quality of the data depends on the processes involved in creating the data. To get better quality data, one must first understand what quality means and how it is measured (Wand & Wang 1996, 86-95). There are several approaches in the literature that can be applied to understand the concept of data quality. One of them is the data lifecycle, which looks at the functions from data generation to its endpoint (Wang et al. 1998, 58-65). Data quality can also be controlled from the perspectives of different functions of information systems (Boyadzhieva & Kolev 2010, 386-395).

2.2.1 Data Classifications

Data can be classified by examining its different definitions. Data is considered as unstructured facts; information refers to structured data utilized in analyses and knowledge is human knowledge based on experience. Data can be refined as information by creating a structure to it, and information is obtained into knowledge when the information is examined (Laihonen et al. 2013, 84). Data and information terms are often used interchangeably. However, they differ, as information refers to processed data (Pipino et al. 2002, 114-116).

Data is often divided into three different types:

- **Structured data** is typically categorized as highly organized data. Database tables and statistics are an example of the most common type of structural data (Batini et al. 2009, 14-16)
- **Unstructured data** is considered as qualitative data which cannot be processed and analyzed with conventional tools. This type of data cannot be stored in rows and columns in a relational database. (Batini et al. 2009, 16). For example, images and videos are unstructured data (Aljumaili et al. 2016, 232).
- **Semi-structured data** has some degree of flexibility. Semi-structured data is also called unscheduled or self-descriptive data (Batini et al. 2009, 16-17). They can be considered as a bridge between structured and unstructured data as they represent partially structural data, but they do not have an exact structure of the data model (Aljumaili et al. 2016, 232).

Often, AI solutions tend to use structured data because it is easier for the machine to read. Unstructured data may require a lot of work before it can be used for AI applications. The literature about data quality focuses also mainly on structural data. The reason being that this type of data is mostly utilized in most organizations (Batini et al. 2009, 18-20).

There are similarities between the manufacturing of physical products and the manufacturing of data products. With products the system utilizes raw materials to produce physical products. Similarly, an information system can be seen as a data production system that utilizes raw data (e.g., individual numbers, records, files, spreadsheets, or reports) to produce data or data products. The created data product can then be processed as raw data in another data manufacturing system (Wang et al. 1998, 58-65; Ballou et al. 1998, 462).

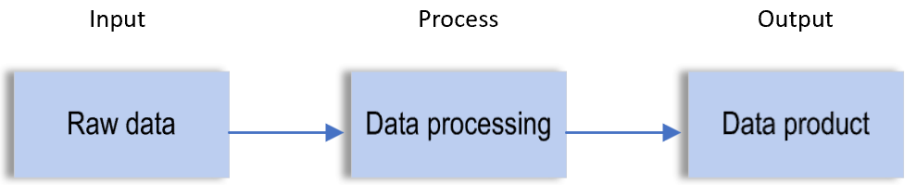

Figure 2. Manufacturing of data products (Wang et al. 1998)

Data can be categorized based on its common characteristics. Categories are useful from a data management perspective as certain data can be treated differently based on its category. Understanding the dependencies between different categories can be helpful to improve data quality (McGilvray 2008, 341). Table 2 explains the common categories of data.

Table 2. Data categories (McGilvray 2008)

Category	*Description*	*Examples*
Master data	Describes the people, places and things involved in an organization's business.	Examples like people, including customers, employees, vendors, email addresses, URLs, IP addresses are considered as master data. Also, things like accounts, products, assets, and device IDs are included in the category.
Transactional data	Describes an internal or external event or transaction that is related to the organization's business.	Examples like sales order, purchase order, invoice, medical visits, and a shipping document are considered as transactional data.

Reference data	Sets of values referred to by systems. data stores, applications, dashboards and so on.	Examples include code lists, lists of valid values, status codes, product types and social media hashtags.
Metadata	Metadata is information about data. It describes, labels, and categorizes data which makes it easier to filter, retrieve and interpret. Metadata has subcategories such as • technical metadata • business metadata • label metadata • catalog metadata • audit metadata	Examples like field names, data type and tags are usually considered as metadata.

Data can also be categorized differently than described in Table 2. For example, it can be difficult to decide whether a list of valid values is only reference data or also metadata. Reference data may be needed to create Master data and Master data is needed to create transaction data. Metadata, in turn, is needed to understand other categories of data (McGilvray 2008, 340-348).

2.2.2 Data Quality Dimensions

Data quality can be assessed through a variety of dimensions. Data dimensions usually measure e.g., completeness, accuracy, uniqueness, validity, consistency, and timeliness of the data, although there are several other data dimensions in literature about data quality. Data quality is ensured by designing and implementing techniques to measure, evaluate and develop data quality (Sebastian-Coleman 2013, 40).

Data quality dimension refers to data characteristics that represent a single perspective on quality. The literature does not recognize a single list of data quality dimensions. Instead, there are different perspectives based on intuition, previous literature, and empirical research (Wand & Wang 1996, 86-95).

Individual dimensions of data quality can be viewed objectively by comparing the number of deviations contained in a single data set. (Ballou & Pazer 1985, 126) Deviations can be calculated by comparing a data set with historical data or by comparing the data with reference values (Sebastian-Coleman 2013, 67). However, such a method does not consider the need of a user (Wang & Strong 1996, 8-14).

Dimensions can be explored based on their categories. According to (Wang & Strong 1996, 5-33) the dimensions can be divided into four categories: *intrinsic*, *contextua*l, *representative*, and *accessibility* as seen in figure 3. Haug et al. (2009) notes that *reputation* and *credibility*, the dimensions introduced by Wang & Strong (1996), are not natural characteristics of data, but subjective experiences of the user. Wand & Wang (1996) identify four natural data quality dimensions in their model which are *complete, unambiguous, meaningful,* and *correct*. Reputation and credibility are excluded from this model.

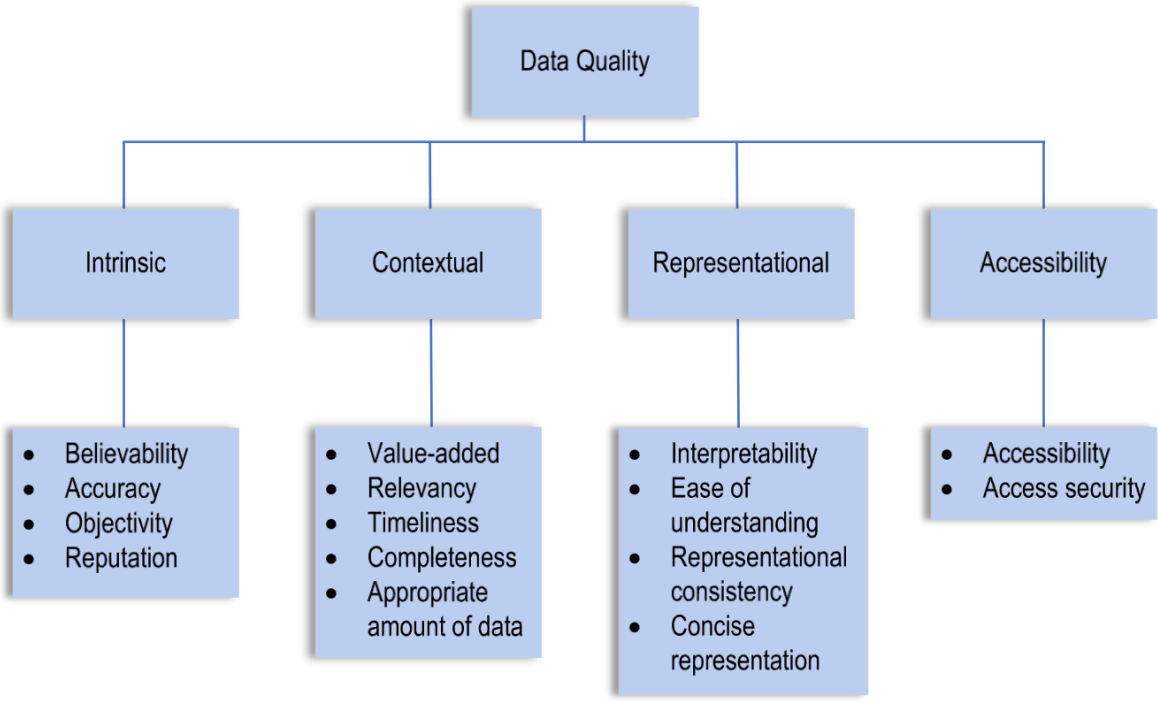

Figure 3. A Conceptual Framework of Data Quality (Wang & Strong 1996)

For many, data quality means data accuracy. However, the quality of the data is more widely measured when more qualitative characteristics are considered. The choice of measurable quality dimensions depends on the requirements of the use case (Boyadzhieva & Kolev 2010, 386-395). As mentioned earlier, there is no common consensus on which dimensions define data quality. Differences in the definitions are due to the contextual nature of quality (Batini et al. 2009, 17-20) E.g., consistency can be viewed in terms of presentation, rules, standards, or other data. Table 3 presents the main dimensions of data quality and their definitions found in the literature.

Table 3. Data Quality Dimensions

Dimension	*Definitions*	*Ballou & Pazer 1985*	*Wang & Strong 1996*	*Lee et al. 2002*	*Sebastian-Coleman 2013*	*Batini et al. 2009*
Accessibility	Measures data availability, or how easily and quickly the data is retrievable.		x	x		
Accuracy	Measures how accurate the data is for the purpose.	x	x	x		x
Appropriate Amount of Data	Measures the volume of data and the appropriateness of the data volume.		x	x		
Believability	Measures how true and credible the data is.		x	x		
Completeness	Measures the sufficiency and depth of data and if any data is missing.	x	x	x	x	x
Concise Representation	Measures how compactly the data is represented.		x	x		
Consistent Representation	Measures if the data is presented consistently in the same format.	x	x	x	x	x
Ease of Manipulation	Measures how easily the data can be manipulated and applied to different tasks.		x	x		
Free-of-Error	Measures how correct and reliable the data is.				x	
Interpretability	Measures how appropriate the languages, symbols, units, and definitions are.		x	x		
Objectivity	Measures how unbiased, unprejudiced, and impartial the data is.		x	x		

Relevancy	Measures how applicable and helpful the data is for the task.		x	x		
Reputation	Measures how the data is regarded in terms of its source or content.		x	x		
Security	Measures if access to the data is appropriately restricted.		x	x		
Timeliness	Measures how sufficiently up-to-date the data is for the task at hand.	x	x	x	x	x
Understandability	Measures how easily the data is comprehended.		x	x		
Value-Added	Measures how beneficial the data is and how much advantage is added from its use.		x			

2.2.3 Weak Data Quality

Problems with data quality can be caused by many factors. Often the cause can be categorized into two groups: practical factors from the collection or processing of incomplete data in an information system or structural factors caused by the inconsistencies between user requirements and the functionality of the actual data system. Practical factors can be mitigated with thorough data management methods, while correcting structural problems requires fundamental changes to the data architecture (Maydanchik 2007, 28-32).

A significant portion of errors in data are caused by human error during the entry phase of the data (Mahanti 2014, 22; Umar et al. 1999, 280-281). Other reasons for weak data quality may be the unclear definition of the data or the inconsistent data model, which leads to errors in the data, especially when using multiple information systems. Integrations between systems can also weaken data quality when some of the data is not transferred correctly or the values are in the wrong place (Silvola et al. 2011, 146-162). Data quality problems may not only be due to inaccurate data, but also due to unclear responsibilities of the people or poorly managed information systems.

2.2.4 Data Management in AI Development

The development of an AI solution starts with an idea of the product after which the stages of the project can be defined. The collected data is then curated, which includes filtering and organizing the data. After that the prototype can be created and trained with the data. After prototyping and testing, the product can be taken into piloting and production use. The lifecycle management of the

product and its data should go on throughout and after the development process (Anderson & Coveyduc 2020, 145-162).

The data lifecycle management involves creating, using, modifying, sharing, and transferring the data. The activities that form the basis of data management must be considered as part of the planning phase of the data management and governance structure in AI development (Sebastian-Coleman 2013, 118). Activities during the data lifecycle are supported by good data security, metadata management and adequate data quality. All three aspects of data management must be developed throughout the data lifecycle to ensure the reliability of the data in the organization. Data security, metadata, and data quality management are the cornerstones of data management that must be integrated into organizational processes (Sebastian-Coleman 2013, 35-45).

Data security ensures both data protection and data confidentiality, as well as proper access rights to the data. The first step is to identify the data that requires protection and to identify the systems that contain secure data. The level of security is then determined and the business processes that need the data are identified. Based on the identified data, the criteria, and conditions about which data can be used are determined. Data security is important to ensure that confidential data does not end up in the AI solution (Sebastian-Coleman 2013, 46-52).

It is obvious that higher quality data gives more relevant results. Proper quality data leads to more accurate AI solutions, up-to-date data leads to more current results, and more comprehensive data teaches AI better in its operations. Among other things, good quality data can improve the customer experience, increase productivity, enable rapid response to business opportunities, and provide a competitive advantage through insights from the data (Sebastian-Coleman 2013, 14-18).

2.3 Data Quality Measurement

The quality of the data cannot be measured without common dimensions, metrics, thresholds, and indicators based on the use case. When considering data quality, it is important to understand that quality can mean different things in different contexts, organizations, and industries. It is therefore essential to be able to choose the right dimensions of data quality according to the application and industry (Korpela 2018). Understanding the right dimensions has led to several approaches to measuring and evaluating data quality dimensions (Bronselaer et al. 2018, 36-38).

The metrics used for data quality dimensions must be understandable. If the metrics cannot be understood the measurement is not useful, even the measured subject would be very important. Measurement is a communication tool and at the same time an analysis tool. In addition to the data being measured, consumers of the data must understand what the measurement represents and

the context of the measurement. Measurement must also be reproducible to be able to compare the measurements (Batini et al. 2009, 20-22)

2.3.1 Measurement Types

Measurement types answer the question how to measure? Data quality can be measured e.g., from the perspectives of data models, data values, data domain, data presentation, and data policies. In practice, there are two options to measure data quality: real-world test and evaluation. The real-world test confirms whether the data corresponds to reality or not. The real-world test can be carried out using reference data or a team of experts. However, the differences in expert's opinions may lead to uncertainty (Sebastian-Coleman 2013, 44-48).

Measuring data quality can be objective or subjective. An objective measurement is based on quantitative metrics (Batini et al. 2009, 23-24). The objective metrics measure independent characteristics, and the metrics can be used without contextual information of the data. The objective measuring involves at least one of two basic comparisons: data can be compared to a clearly defined standard or to itself over time (Sebastian-Coleman 2013, 44). The objective metrics can be divided into task-independent and into task-dependent metrics. The task-independent metrics describe the state of data without contextual understanding of the application. The metrics can be utilized in any data set, regardless of the task in question. The task-dependent metrics, which include an organization's business information and specific regulations, are developed in specific contexts (Pipino et al. 2002, 115).

In subjective measuring qualitative metrics are used to gain opinions of data users and managers (Batini et al. 2016, 42). Measuring subjective dimensions like credibility and relevance, information from the data consumers is collected through surveys and interviews. Subjective data measurement reflects the experiences of the data consumers (Sebastian-Coleman 2013, 69).

2.3.2 Data Profiling

Determining data quality metrics can be difficult because the metrics are often application dependent. A common way of determining data quality is data profiling (Andreescu et al. 2014, 3) Data profiling is a type of data analysis used to characterize the properties of a data set. Profiling provides an overview of data structure, content, rules, and relationships using statistical methods. The result is information about the properties of the data, such as data types, field lengths, value sets, format and content models, and indirect rules (Sebastian-Coleman 2013, 78). The main methods of data profiling can be divided into three groups, which are structure, content, and relationship analysis (Dorr & Murnane 2011, 12; Mahanti 2014, 30; Azeroua et al. 2018).

Profiling techniques can be divided into two categories: manual and automated. Manual techniques require people to look at the data through queries. This approach is suitable for smaller and simple data sets. Automated technologies utilize software tools to summarize and analyze data. Automated technologies are more suitable for big data with multiple fields and sources. Once the data profiling process is complete and problems are identified the source data should be “cleaned”. Data cleaning eliminates errors and inconsistencies in source data and improves the data quality (Andreescu et al. 2014, 5).

2.3.3 Metrics

Data quality metrics define what is being measured. The dimensions provide certain perspectives on data quality. There are several different metrics for quantifying these dimensions (Heinrich et al. 2018, 72). There is flexibility in the methods for measuring data quality, as each dimension can be measured in several different ways (Aljumaili et al. 2016, 242). Often the most difficult task in measuring is to define the data quality dimensions. The formation of the meter for the dimensions can be considered more straightforward (Pipino et al. 2002, 124)

To measure data quality dimensions a few most suited metrics should be selected. Several different factors can be considered when selecting metrics, such as meter priority, measurement method, measurement frequency and cost-benefit ratio (Umar et al. 1999, 298-300). Table 4 presents the most common dimensions and possible metrics to examine the dimensions.

Table 4 Dimensions of data quality and their metrics (Batini et al. 2009)

Dimension	*Metrics*
Accessibility	• Request time – Delivery time • A survey / interview
Accuracy	• The distance between the values stored and the correct values • Number of exact values provided • A survey / interview
Completeness	• A survey / interview
Consistent Representation	• Number of values violating the used format • A survey / interview
Interpretability	• The amount of data to be examined • Documentation of key values • A survey / interview

Objectivity	• A survey / interview
Relevancy	• A survey / interview
Reputation	• A survey / interview
Security	• The number of failed login attempts • A survey / interview
Timeliness	• The time when the data is stored in the system - the time when the data is updated in the real world • Time since the last update • Update time • A survey / interview

2.4 Data Quality Assessment

The purpose of data quality assessment is to recognize errors in the data and to understand the possible impact of the errors. Both identifying errors and understanding their implications are critical. Data quality assessment can be implemented e.g., by using a simple qualitative assessment or a detailed quantitative assessment. Evaluation of the assessment can be done either by general knowledge, guiding principles, or using certain standards.

The goal of data quality assessment should be to understand the state of the data in relation to expectations and to draw conclusions whether the result meets the expectations for a particular use. This process includes the need to understand how effectively the data represents the objects, events, and concepts it is designed to represent (Sebastian-Coleman 2013, 113).

Batini et al. (2009) points out in a review about data quality assessment that many of the existing methodologies are theoretical and have not been applied extensively in practice. This is important to take into consideration when planning and applying the methods in different use cases and in different organizations. According to the review, the assessment methods can be divided based on 5 recurring elements (Batini et al. 2009, 20-24):

- **Data analysis** where an overall picture of the current situation is created based on the data and related rules.
- **Analysis of quality requirements** where data users and administrators investigate current problems and set new quality targets.
- **Identifying critical areas** where key databases and data flows are selected for quantitative viewing.
- **Process modeling** for modeling data production and update processes.
- **Quality measurement** where the quality dimensions related to the problems are identified and metrics for these dimensions are selected.

The different dimensions of data quality are examined in data quality assessment. Several frameworks have been developed to assess data quality using both objective numerical indicators and more subjective evaluation methods depending on the nature of the data. The techniques and objectives used by different methods vary and may not involve the same steps as other methods (Batini et al. 2016, 46-48).

To help understand the differences between the methods and the data the method can be used, the methods are classified into four different categories based on their content data: operational, financial, general picture and auditing (Batini et al. 2016, 50-53).

E.g., auditing methods focus on assessing the current state of data quality and do not provide visibility for the development of processes, while operational methods examine evaluation and development activities from a technical perspective (Lee et al. 2002, 45). Financial methods focus on assessing the costs related to data quality (Wang 1998, 60).

2.4.1 Assessment Methods

The purpose of this thesis is to examine the quality of data used on training AI applications in Finnish industrial companies. Therefore, the method chosen to assess the data must be able to audit the current state of the data. The assessment method types mentioned in this thesis are divided into two categories: standardized and modular. Standardized methods are intended to be used as is and modular methods can be used by selecting a suitable module depending on the assessment project (Wang 1998, 62). Table 5 lists all data quality assessment methods examined in the thesis.

Table 5. Data Quality Assessment Methods

Method	Source	Type
Data processing quality control model	Ballou & Pazer (1985)	Standardized
Total Data Quality Management (TDQM)	Wang (1998)	Standardized
A Methodology for Information Quality Assessment (AIMQ)	Lee et al. (2002)	Standardized
Data Quality Assessment Framework (DQAF)	Sebastian-Coleman (2013)	Standardized
Data Quality Assessment (DQA)	Pipino et al. (2002)	Standardized
10 Step process	McGilvray (2008)	Modular
Hybrid approach model	Woodall et al. (2013)	Modular

2.4.2 Standardized Methods

Methods for assessing data quality have been developed for decades. As the earliest method, Ballou & Pazer (1985) presented four dimensions (accuracy, completeness, timeliness, and consistency) in their model to measure data quality that can only be applied to data with numeric values. The model provides information on the magnitude of deviations and monitors errors at different points in the data stream. One of the earliest data quality frameworks is Wang's (1998) total data quality management (TDQM) in figure 4, which is based on the quality dimensions of Wang & Strong (1996). According to the TDQM model, the organization must think of information as a product passing through the manufacturing line in the same way as in the traditional manufacturing industry. Where the physical product is made from raw material, the data product is made similarly from raw data in the information system. The purpose of the TDQM model is to provide high-quality information to information consumers.

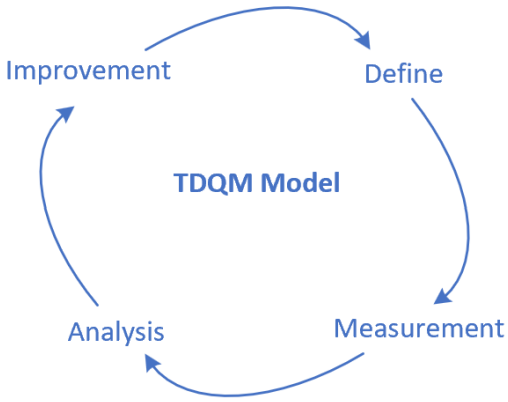

Figure 4. TDQM Model (Wang 1998).

The TDQM model is an iterative process that consists of 4 different steps: define, measurement, analysis, and development. During the definition phase, the characteristics of the data product, the data quality dimensions, and the quality requirements of the information system are identified. The measurement phase identifies suitable quality indicators for the system. The analysis phase aims to understand the root causes of quality-related problems and the costs of fixing the possible problems. Finally, in the improvement phase, methods for developing quality will be produced through appropriate dimensions (Wang 1998, 63).

Lee et al. (2002) developed the Methodology for Information Quality Assessment (AIMQ) which examines data quality subjectively. The method consists of three parts that can also be used

independently. The first part is the 2x2 matrix, which describes the importance of data quality to its users and administrators. The fields in the matrix divide the dimensions of data quality into four categories: stable, reliable, useful, and usable information. The second part is a questionnaire based on the given data quality dimensions that can be used to assess an organization’s data quality by scoring answers of the survey based on the used dimensions. The third part consists of two alternative methods of analysis. The methods compare the results of the gap analysis and the questionnaire of the same organization or to a selected well-established reference organization (Lee et al. 2002, 56). Although Batini et al. (2009) points out that there is no such database about reference organizations known in the literature about data quality. The AIMQ method also differs from others with its subjectivity. The method also focuses purely on data quality assessment and does not provide tools for quality development. Some of the frameworks concentrate more on the objective type of measuring providing tools to assess numerical content.

The Data Quality Assessment Framework (DQAF) contains only objective indicators that continuously monitor data quality. Objective metrics are preferred here because the data should still meet certain basic requirements to be usable, even though the quality of the data is determined by the needs of its users. The DQAF framework provides a general model for continuous assessment of data timeliness, completeness, accuracy, consistency, and integrity. The model includes a total of 48 different metrics for these dimensions (Sebastian-Coleman's 2013). Cappiello et al. (2004) writes that continuous data quality measurement based on algorithms can ignore different data requirements of users and present a model in which the automated measurement process can be tailored to users' requirements.

Different subjective and objective assessment methods can also be combined within the same framework. The Data Quality Assessment (DQA) framework assesses the current state of data quality using both subjective and objective methods, after which their results are compared. If either a subjective or objective review reveals problems or there are differences between the results, the process proceeds to investigate the root causes of the problems. Based on an analysis of root causes the development proposals will continue case-by-case (Pipino et al. 2002, 139). The DQA framework guides organizations to formulate the appropriate indicators for their purposes but provides three base categories for them: simple ratio of the desired values, calculation of the minimum or maximum, and weighted average. The DQA model is the most informal of standard models and does not directly provide strict guidelines or concrete tools for assessing data quality. This can make it suitable in cases where the indicators and possible development measures are intended to be determined separately (Pipino et al. 2002, 140).

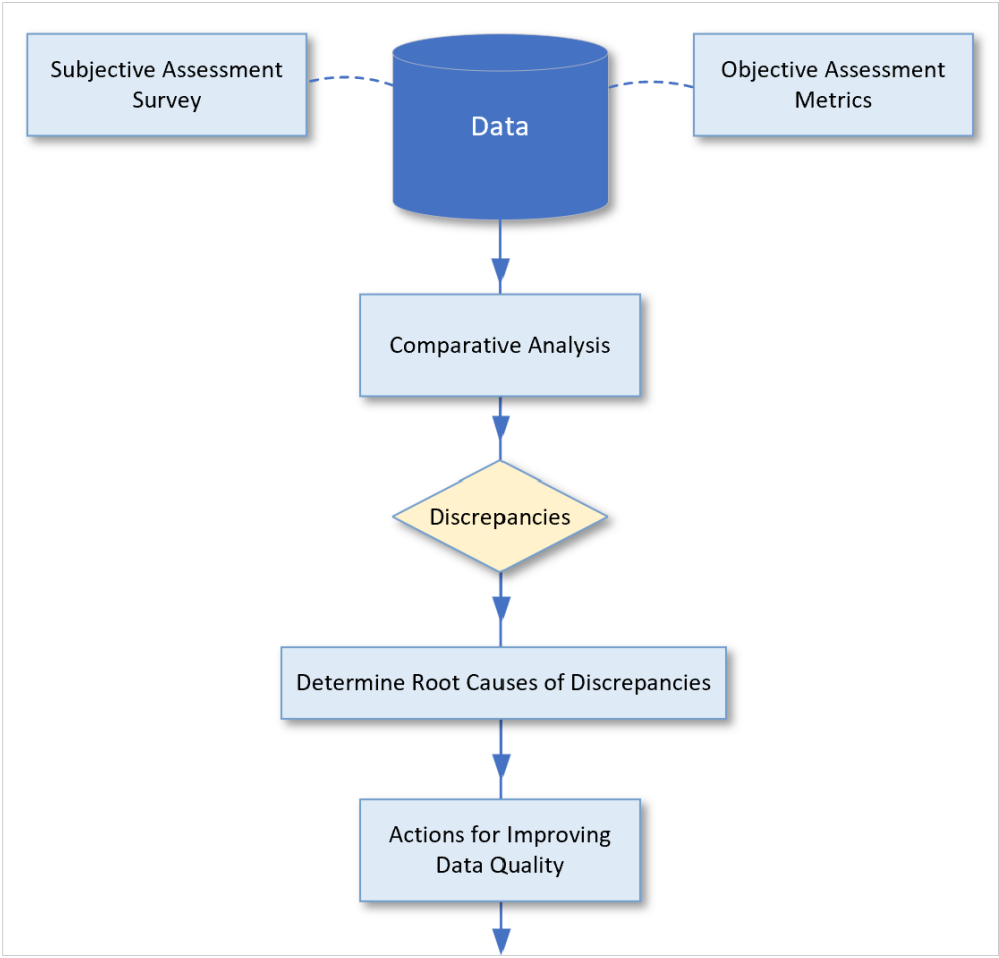

Figure 5. DQM (Pipino et al. 2002).

2.4.3 Modular Methods

In some methods, the measures are selected to meet the individual needs of each case. McGilvroy (2008) presents a 10-step iterative model where the steps used are selected according to project requirements. The model is based on the Plan-Do-Check-Act (PDCA) cycle including three high level sections: evaluation, understanding and operation. The evaluation section includes the first four steps: identifying business needs, analyzing the data environment, evaluating data quality, and evaluating business impacts. In the understanding section root causes are identified behind the problems and a plan of development measures is developed. Eventually, during the operation phase, future errors in the data will be prevented, existing errors will be corrected, and monitoring methods will be applied. The tenth and the last step is communication on actions and results which crosses all 3 high-level sections as continuous operations.

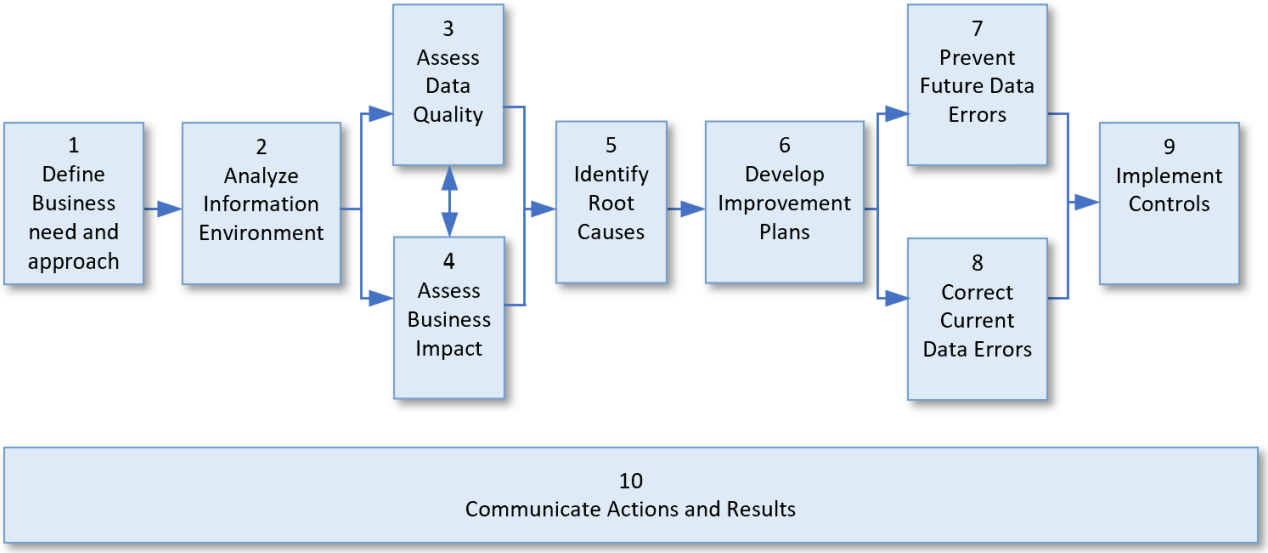

Figure 6. 10 Step Plan-Do-Check-Act (PDCA) (McGilvray 2008).

Woodall et al. (2013) underlines the formulation of assessment methods according to the needs of each organization. The Hybrid model approach is not necessarily a ready-made operating model, but it provides four steps for the development of an organization-specific data quality assessment method. The first step defines the purpose of the assessment, e.g., to examine a previously discovered data quality problem or to assess the current state of the organization's data quality. The second step identifies the organization's requirements, which must be in line with the objective of the first phase. The requirements set may include calculating the costs of weak data quality or modelling data flows. The third step is to select the functions of the assessment methods that meet the organization's requirements. The second and third phases also support each other and can be carried out iteratively, as it can be difficult to understand the requirements without knowledge of the assessment methods. Finally, in step 4, a functional order is defined, considering dependencies between functions.

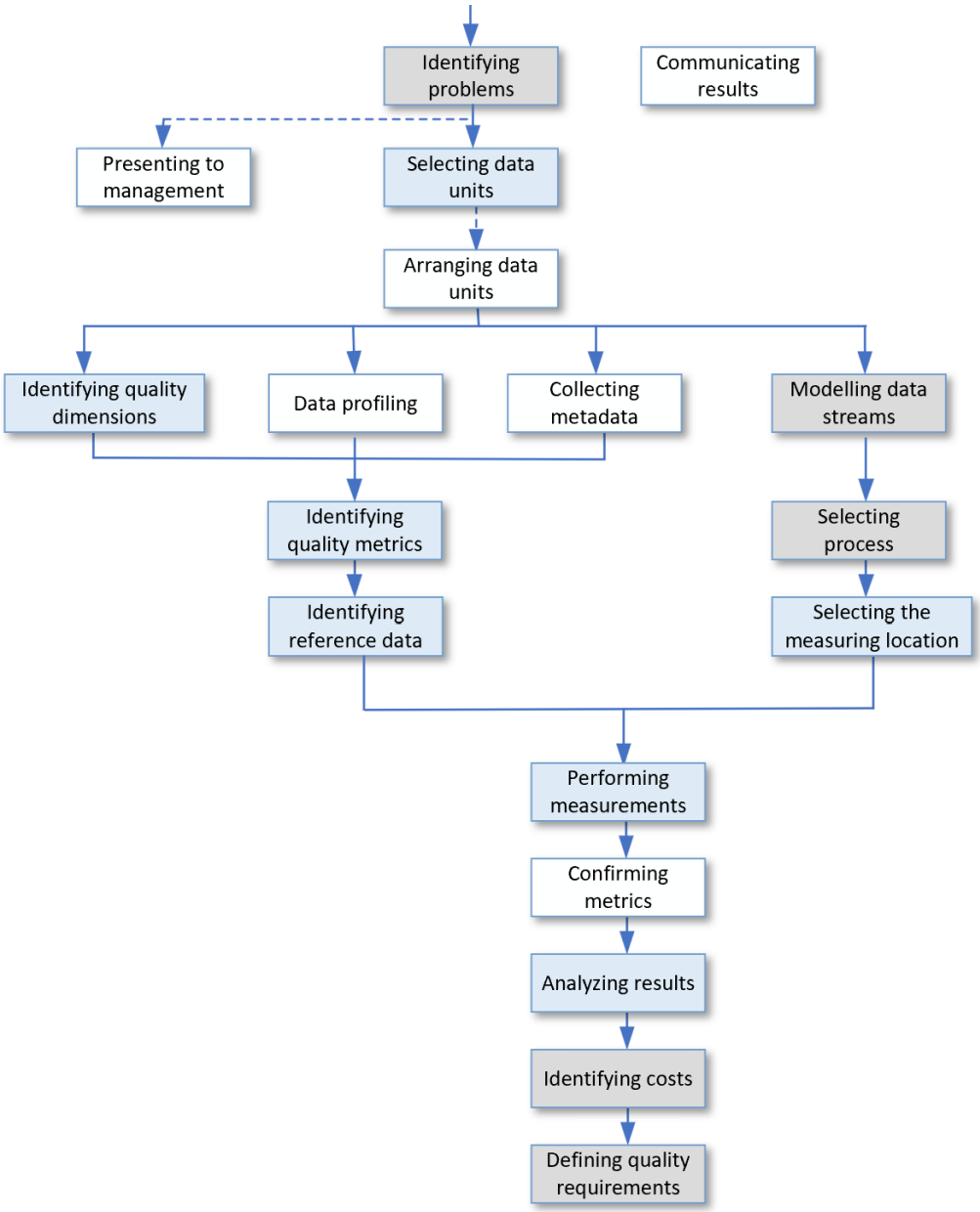

Figure 7. The Hybrid model approach (Woodall et al. 2013)

Figure 7 shows in general terms all the functions of the assessment methods in the hybrid model approach. The blue boxes describe the recommended functions that can be found in all the methods examined in this thesis. These steps include selecting data units, identifying quality dimensions, identifying quality metrics, identifying control data, selecting the measuring location, performing measurements, and analyzing results. Gray boxes are adjustable functions that can be performed in several different places depending on the actions selected. White boxes are more fragmented functions that can be utilized in appropriate situations. Dashes between boxes reflect the dependence of the function e.g., one cannot give a presentation about the problems without identifying them (Woodall et al. 2013, 370).

2.5 Improving Data Quality

Batini et al. (2009) classify improvement methods into data-driven and process-driven strategies. Data-driven methods modify data values directly by updating the data source values. Process-led methods are aimed at redesigning the data processing processes like adding functions or modifying the existing functions of the process. Lee et al. (2002) also mentions similar process-oriented methods for data quality improvement. Umar et al. (1999) divides methods similarly for cleaning data and cleaning processes. Silvola et al. (2011) further divides improvement methods into four categories: passive, reactive, active, and proactive.

There are several ways to improve data quality. Mainly, the improvements can be divided into two categories: problems can be reactively corrected when they are noticed, or they can be prevented proactively at their source (Mahanti 2014, 29). Often it is better to do the corrective measures proactively, as individual errors in the data can recur quickly, which can be costly to correct. In some cases, the improvement methods are tied together with data assessment methods as knowledge of the current state is necessary before implementing development measures (Woodall et al. 2013, 372).

2.5.1 Proactive Methods

The proactive approach aims to address the possible data quality issues at data source. According to Batini et al. (2009) identifying the causes of errors is the most common step in data quality improving methods. There can be many factors behind an individual problem, so therefore figuring out the root causes can be challenging. There is no one-size-fits-all method for identifying the root causes, but a thorough investigation is essential. According to Lee et al. (2002) identifying root causes usually requires the collaboration of technology and business experts. One concrete way to find the root causes of problems is to track the data flow from the data creation to its various use cases (Loshin 2001, 44; Silvola et al. 2011, 98). After detecting a possible error, the data flow is monitored backwards until the source of the error is found. Possible problem areas can be at the production phase of data or modification and transferring of data between systems (Loshin 2001, 76). Data flow modeling can be done e.g., with the Information Production Map (IP-MAP) tool using its eight elements: data sources, processing, data warehousing, decision points, quality control, information system boundaries, organizational or business process boundaries, and Information Products (Loshin 2001, 82; Shankaranarayan et al. 2003, 14). IP-MAP is based on the same perspective as TDQM framework where data is considered as a product (Shankaranarayan et al. 2003, 14). In addition to identifying the root causes, IP-MAP visualization helps to understand the data flows at a general level (Silvola et al. 2011, 124). Batini et al. (2009) points out that the process modeling required by IP-MAP can be very expensive and, in some cases, impossible to

implement in practice. Active development of data quality requires continuous monitoring. Continuous monitoring is mentioned in the literature as part of several data quality development models. The model presented by Loshin (2001) identifies the effects of data problems, data quality objectives, data designs, and implements quality improvement measures, and finally monitors data quality by comparing the current state with defined objectives. If monitoring reveals problems, the cycle starts again. In McGilvray's (2008) model, ongoing monitoring and metrics are designed and implemented to monitor the impact of the actions and prevent the organization from returning to the old model with problems. Sebastian-Coleman's (2013) DQAF model uses continuous development actions based on measurement which is often used in manufacturing industry's quality control processes. Another advantage of continuous measurement is the faster response time to changes in data that may arise as technical or business processes change.

2.5.2 Reactive Methods

In case the source of the problem cannot be removed, the faulty data can be corrected directly by modifying it or replacing it completely. Prior to corrective action, problems should be prioritized to ensure efficient use of resources (Loshin 2001, 103). The goal in developing data quality should not be to solve all problems, but to have good enough data (Silvola et al. 2011, 129). Data-driven development methods include replacing data with higher-quality data, standardization of data and linking data in a way that there is no overlapping data in different sources (Batini et al. 2016, 82). Reactive data quality development also requires detecting the problem at least once. One method that facilitates this is data profiling which adds a metadata tag of different data sets to use different analysis methods (Abedjan et al. 2015). In practice, profiling utilizes various algorithms that provide information about possible quality deficiencies in a data set (Loshin 2001, 108), such as inconsistent formats, missing values, or apparent deviations (Abedjan et al. 2015). The result of profiling is a more accurate picture of the data structure, content, internal rules, and relationship between data (Sebastian-Coleman 2013, 142). Data profiling can be targeted to a single column, a comparison of columns, or a comparison of entire database tables (Loshin 2001, 113). Many ready-made information system solutions are available for data profiling, although they are not capable of continuous data quality control (Ehrlinger et al. 2022, 4).

3 Research Design

The empirical part of the research is implemented as case study. This method was chosen because the focus of the study is to understand a phenomenon in-depth and create ideas and suggestions for solving problems. The starting point of the study is problems and questions arising from work life that guide the study in the operating environment (Ojasalo et al. 2015, 8). This chapter describes the progress of the study and reviews the methods used to collect and analyze the data.

The study examines data quality problems subjectively with a semi structured interview. The aim was to emphasize the subjective approach in data collection by collecting the opinions of data quality and AI development professionals about the dimensions of data quality and their weaknesses. Another key objective was to identify the most challenging dimensions that weaken data quality as well as possible development proposals for improvements. The following requirements were identified to help achieve these goals:

- Identifying and prioritizing data quality dimensions
- Utilizing subjective measurements to develop the dimensions
- Analysis of results and creation of development proposals

The research is a part of the AI-TIE project, which aims to support small and medium sized companies in developing and growing their business by utilizing AI solutions. The research material has been collected by using qualitative and quantitative methods to collect information on the topic a from Finnish AI and data quality professionals. The study is not intended to gather data on only one specific case but to increase understanding of data quality and its potential problems in AI development at a more general level. The results obtained in this way can be applied to companies that need support in AI development or are just starting their AI journey.

The time horizon of the study is crosscutting where the study examines the current situation of data quality. The current state is explored by interviewing data quality and AI development professionals who have a comprehensive understanding of AI development and data quality through the experience of multiple clients, projects and several years in the field. Data quality was assessed using the AIMQ framework presented by Lee et al. (2002). The framework was used to create an interview frame for individuals interviewed in the study based on the data quality dimensions presented in AIMQ. According to Lee et al, the method is well suited for solving data quality problems. The aim was to assess the quality of the data mainly subjectively to obtain more in-depth information on the phenomenon. Addressing the current problems of the matter is the first step in improving data quality (Lee et al. 2002, 98; Batini et al. 2009, 48; Woodall 2013, 369).

3.1 Collecting Research Data

The collection of research data was carried out using a concurrent mixed research method that involves the separate use of qualitative and quantitative methods within a single phase of data collection (Ojasalo et al. 2015, 16). The method allows both sets of results to be interpreted together to provide richer responses to the research question. The questions in the semi-structured interview were formed based on the AIMQ method's IQA questionnaire. A semi-structured interview was chosen so that the interviewer could clarify questions about the quality dimensions and ask for examples of situations where a quality problem may occur. The dimension-specific questions of the IQA form allow a comprehensive analysis of the current state of the most important dimensions of data quality. At the end of the interview the interviewees were asked to choose 3 of the dimensions they considered most challenging from the perspective of data quality to find out if several interviewees raised the same dimensions. All dimensions in the IQA-questionnaire were included in the interview but the dimensions of comprehensibility and interpretability were combined into the same question due in part to their overlapping nature. The IQA-questionnaire consists of statements that are evaluated with numerical grades, but in the interview the dimensions of the questionnaire were treated as separate topics so that the interviewer could ask follow-up questions and get more subjective information about the background of the answers. The frame of the interview used in the interview can be seen in Appendix 1.

Individuals with long experience in the field and proven understanding of the topic were selected for the interview. All the interviewees have worked as consultants during their careers or are currently working as consultants, which allowed the interviewees to have experience of several projects, in several different industries, over several years. The average work experience of the interviewees was 13 years. Four of the interviewees had a PhD degree in the subject area. In total, seven professionals in AI development, data scientists and data quality experts were selected for the interview. They were emailed a brief introduction to the study, the data quality dimensions covered and their definitions. In addition, interviewees were asked to familiarize themselves with the dimensions prior to the interview to increase credibility of the interview and to sharpen the research delimitation for interviewees to handle correct data (Saunders et al. 2019, 64). The interviews were conducted using Microsoft Teams software. The interviews lasted about an hour per interview.

The interview started by asking the interviewee's role in their job, work experience and work history. This was followed by questions about the quality dimensions one at a time, asking about the most common weaknesses in the dimension as well as possible good practices and experiences to improve the data quality dimension. After reviewing the dimensions, the interviewee was asked to select the 3 dimensions of data quality that they believe affect the quality of the data the most. The

aim was to keep the interview body broad so that possible data quality problems could be examined as comprehensively as possible. The frame provided by the IQA-questionnaire enabled a broad view of the topic and at the same time the use of an established method increases the reliability of the study by reducing the impact of the researcher's own preconceptions (Saunders et al. 2019, 122; Eskola et al. 2014, 16). On the other hand, a wide range of questions may impair the interactivity of the interview (Eskola et al. 2014, 17), so the interviews focused specifically on the answers of the interviews by asking more specific questions on the dimension in question.

3.2 Analyzing Research Data

The interview material was analyzed using the content analysis methods. Essential points from the perspective of the research questions were extracted from the material and then unified and themed into entities relevant to the research question (Tuomi et al. 2018, 56-76). The analysis phase was started by transcribing the recorded interviews into a text document. Interviewees' comments on each dimension were carefully documented, leaving out only irrelevant and out-of-context issues. The transcribed material was summarized and important points for the study were selected, such as mentions of possible causes for problems with data quality dimensions as well as identified good practices in the data quality dimension discussed. In addition, the three dimensions of data quality selected by the interviewees that they think pose the most challenges to data quality were placed in a table for later prioritization.

Each dimension was then placed in the horizontal row of the table. Positive and negative comments from respondents on each dimension came to the verticals. The comments were differentiated so that the challenges were recorded in blue and the positive experiences in red. The table was analyzed one dimension at a time and the data were themed into similar types of observations for each dimension and placed in columns below each other to calculate the number of times the same observations were repeated in the dimensions. The themes were formed with the aim of identifying common background factors and root causes behind the problems mentioned by the interviewees as some of the interviewees could combine some dimension with different issues than other interviewees. At this stage, each dimension had been selected by the interviewees as the most challenging which could be used to prioritize the dimensions that were most likely to cause poor data quality.

After transcribing and analyzing the material, the most challenging dimensions, the possible causes underlying them and things that were perceived to improve the quality of the dimension could be identified. The aim was to prioritize the dimensions, problems, and good practices, which were repeated in at least three of the interviews.

Based on the themes, a summary of the results was written in Chapter 4 of the report. The thematic presentation was chosen so that the report clearly shows the priority of the dimensions and the issues attached to it. The aim was to open the content of the themes with direct quotations and to summarize the results in a section.

4 Results

The results of the study are presented by dimension. The results review each dimension and the challenges and possible development proposals that have emerged during the interviews. At the end of each interview, respondents were asked to select three dimensions that often degrade data quality. The distribution of selections is shown in Figure 8. The dimensions in the results are arranged in a way that the dimension chosen most often by the interviewees is presented first. The results consist of dimensional challenges and development suggestions in the interviews according to frequency. In addition, it is mentioned how often this dimension was chosen among those that degrade data quality.

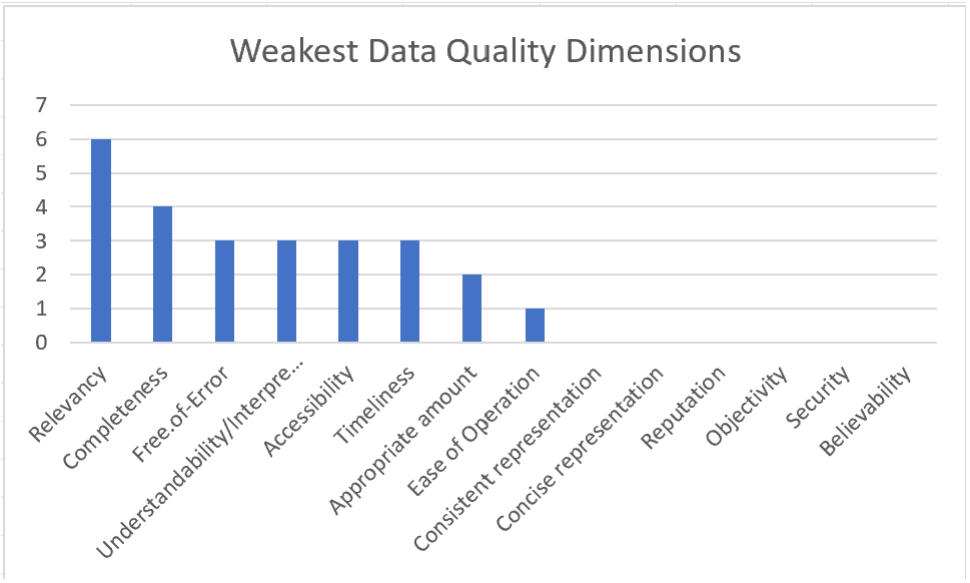

Figure 8. The distribution of selected data quality dimensions

4.1 Relevancy

The most frequently selected dimension of data quality in the interviews which was perceived to have a detrimental effect on data quality in AI development was relevancy. This dimension was chosen by six of the seven interviewees. The biggest problems with relevancy were perceived to be that it was not always known what kind of data should be collected and whether the data collected could be used to solve a business problem. To improve relevancy, the relationship between business and technical people and the importance of understanding to identify and collect the right kind of data were often highlighted.

> “When it comes to the needs of the business, the technical development team may not be able to think about the business need to decide what data to collect”

> “It would be helpful for the business to understand what is technically possible as a lot of data can be collected.”

> “The gap between IT and business has a major impact on this problem”

Concerns were raised about not knowing what data should be collected, which may lead to the lack of necessary data when trying to solve a business problem with AI.

> “One must always make sure that there are things in the data that can be used to predict something”

> “It is difficult to know in advance what data is needed”

Interviewees also raised other common problems with the relevancy in data quality, such as:

> “Data is there but it cannot do what is wanted”

> “There is data, but it does not tell what it should”

To improve the relevancy, the communication between business and technical people should be enhanced to develop an understanding of possibilities of AI and to develop the domain understanding of technical people.

> “Domain expertise helps a lot here”

> “Domain understanding and the availability of people with an understanding of data helps”

> “Communication between business and IT helps so that the IT can prepare for what data to collect in the future

> “Business needs to know what AI can be used for”

To collect and verify the relevant data, it was recommended to start from the perspective of the business problem.

> "Through domain information, it's a good idea to think about the use case and depending on the variables involved collect the data."

> “According to business case, it is important to first outline the metrics to be selected and use them to consider whether we have relevant data.”

In addition, the importance of documentation in the collection and management of relevant data was highlighted to make it easier for individuals to find out what data is being collected in to get a better picture of what data is relevant.

> “Data catalogs and metadata systems help to find relevant data”

> “Documentation of data across the organization helps to identify relevant data”

4.2 Completeness

The second most frequently selected dimension of data quality in the interviews was the completeness of the data. This dimension was chosen by four of the seven interviewees. On the completeness of the data the importance of contextual understanding in identifying complete data and an understanding of what the data is being used for was highlighted.

> "Domain expertise is also important here to understand if we are missing some data"
>
> “The gap between business and IT is also a problem here”
>
> “One must understand what data is necessary at all because complete data can be very incomplete elsewhere”

Gaps in data and responding to them were also perceived as challenging and time consuming.

> “Often we start by looking for nulls and zeros in the data and think how we should react to them”
>
> “Problems come when there are missing values in the data and even rows in the database and one should think about how this affects the overall picture. These also take a lot of time”

The lifecycle of the systems and their evolution was also felt to affect the completeness of the data. In addition, data generation and lack of anomalies were perceived to impair data completeness.

> “Systems evolution and changes in data accuracy also affect data completeness. For example, if we take data that has been collected for 10 years and the data to be collected has been refined somewhere in between. Then the history data may be quite incomplete”
>
> “It may be that a lot of normal data has been collected but there is not enough data on the exceptions to which AI should react”

To improve the completeness of the data, the importance of increasing communication between business and technical people and the importance of domain expertise were again suggested.

> “Reducing the gap between business and IT helps to know what the necessary values are in that data”
>
> “Business people’s interest in the data that is collected helps”
>
> “Domain expertise is important. We are better able to say whether we are missing something”
>
> “It is good to understand what data is necessary from the end user's point of view”

4.3 Free of Error

The third most frequently selected dimension of data quality in the interviews was free-of-error. This dimension was chosen by three of the seven interviewees. The following challenges were mentioned.

> “Errors in the data produced by IoT devices can be identified and corrected quite easily, but other types of data always require more clarification”

> “Errors in IoT sensor data can be captured quite easily but for other types of data you need to know the context”

Issues related to data production were also perceived as a challenge.

> “Data may be entered into systems in quite different ways, affecting completeness”

> “The problem is often the inability to define and produce information accurately and in a controlled way”

To improve the accuracy of the data quality, it was proposed to enhance the control of data production and the importance of defining the data was emphasized.

> “In my experience, modeling and defining data can improve the accuracy of the data”

> “To improve accuracy, it is possible to enforce the form data entered into a database”

4.4 Understandability / Interpretability

In the interview, three out of seven people also chose the dimension of understandability as one of the dimensions affecting debilitatingly to the quality of data. The following things were mentioned about the understandability of the data.

> “Lack of documentation is often an issue”

> “The understandability of data is often affected by the lack of documentation, which is often seen mainly as a cost”

Domain knowledge also emerged from the interviewee’s responses.

> “Domain understanding is often in the head of people and requires knowledge transfer”

> “It is important to understand the context and how information is generated”

To improve the understandability of the data quality, the understanding of the context and the importance of documentation was reminded.

> “It is important to understand what the data really means”

> “A data catalog is necessary that is often missed”

> “It is good to have one place where you can see as comprehensively as possible what data can be found in the company”

> “Documentation in general, if something is started by a person and another person will continue later, then it is really important to document the data”

4.5 Accessibility

Three of the seven interviewees chose accessibility as one of the three weakest dimensions. Accessibility was felt to be affected by issues such as the fragmentation of systems and information on who can be asked to use and access the data.

> “Access to different systems may be hampered by different policies and systems fragmentation”
>
> “Data is fragmented into different systems with separate user controls”
>
> ““Human contact is often required to know where any data is”

In addition, obtaining data seems to be sometimes very time consuming, which affects the project schedule.

> “It must be first determined whether the data can be used for this purpose, and it can take a lot of calendar time”
>
> “It can sometimes take up to weeks to get the data, and this sometimes affects the overall schedules a lot”

Increasing use of cloud services was seen to improve accessibility, although there were also differences of opinion.

> “Many organizations already have data in the cloud that helps with data availability”
>
> “Cloud services have improved availability”
>
> “Cloud services may even fragment data more when the cloud storage is cheap and quickly available but then availability may suffer when it is not known where certain data is in cloud”

In addition, it was felt that documentation often helps with data availability.

> “Comprehensive documentation helps”
>
> “Metadata systems can help here”

4.6 Timeliness

Three interviewees selected timeliness as one of the three most challenging dimensions. Problems with timeliness included issues about the data used to train the AI application, data lifecycle and the timeliness between the systems.

> “AI may be trained with old data that results in an outdated AI application”
>
> “If data collection is not automated in any way, then the data will not stay up to date”
>
> “One data dump may be taken that is fresh for a while but expires over time if not updated”

In addition, problems were seen in the timeliness of data between systems.

> “Synchronizing data can cause problems in operational systems. Failure to update the data quickly enough can lead to operational errors”

> “Timeliness between systems can often be problematic. If the data has not come at the right time, it is the same than it is not available”

Timeliness was also seen as a larger entity as the overall time dimension is important to understand on a larger scale.

> “It would be good to understand the complete time dimension and how it works”

> “It is important to understand, for example, whether data is up-to-date throughout the processing pipe on time”

In addition, timeliness was seen to depend a lot on the context and the data lifecycle required.

> "Timeliness can be challenging, for example, if the patient first receives the first laboratory results and additional results are obtained from Labs 2 and 3 during the day, the information in Lab 1 may be misdiagnosed because the results in Labs 2 and 3 have not yet come"

> “The information is getting better all the time. Up-to-date information does not always give the best results”

> “Real-time is different from timeliness. Much depends on the context.”

Timeliness was also seen to be affected by the history of data, storing of data, collection of data, and the concept of time associated with them.

> “There may not be enough historical data either, so no trends may emerge from the data”

> “Timeliness can be partly misleading or even poor data quality dimension. The time dimensions are much more essential. Historizing, versioning, understanding the passage of time, understanding the data as a chain of events, and the data lifecycle are more important”

Things that had a positive effect on timeliness were seen as e.g., automation of data collection and regular training of the AI model.

> “Automating data collection helps”

> “The more data collection can be automated the better”

> “Ideally, the AI model is trained with new data from time to time”

In addition, a good understanding of the context was also seen to help ensure up-to-date data.

> “It is important to understand the whole chain of events and its needs”

> “Domain understanding and the availability of people to help understand data and the need is helpful”

4.7 Appropriate amount

The appropriate amount was selected twice for the dimensions that most often degrades data quality. In many cases, too little data and an insufficient amount of data were reported as a problem.

“If there is not enough historical data or the data is not comprehensive enough, analytical models lack benchmarks”

> “Often the problem is that trend data may not be available because the data collection model may have changed over the years”
>
> “Limited data does not contain enough different situations”
>
> “Deviation situations might be rare in the data, which just are important for AI models”
>
> “The problem arises when there is not enough data”
>
> “AI applications need data that is different. If there is not enough data, the phenomenon is usually overshadowed by noise”

In general, the more data there is, the better.

> “The more data the better”
>
> “Often it is not a problem if there is a lot of data, but if there is not enough data”
>
> “The denser and more accurate the data the better”

The proliferation of cloud services was also felt to help with this dimension and when generating data.

> “Cloud services have helped in data generation”
>
> “"Storing data and cloud storage is now relatively cheap in cloud"
>
> “The amount of data has also increased as cloud services become more widespread”

4.8 Ease of Operation

Ease of operation was not generally perceived as a data quality problem. One interviewee chose ease of operation as one of the most problematic dimensions of data quality and commented on the dimension as follows:

“If data is difficult to access and scattered in different places, accessing data requires more communication between people, which often takes a lot of time. In large organizations, processes can also be relatively heavy and access to data therefore takes time. Better processes would increase the self-direction of data consumers.”

4.9 Consistent representation

Consistent representation sparked a lot of thought among the interviewees but was not chosen as one of the three weakest dimensions. Issues affecting consistent representation were reported to be mainly issues related to generating data.

> “In large organizations, you may see that data is processed on a team-by-team basis. For this reason, shadow data is created, and it may be that the copy begins to live its own life”

> “Often I see team-specific solutions”

> “The problem is that the creation of the data affects consistency. The same type of data may be produced in several different ways, which should be considered if the data is in some point combined”

> “Systems may be used in a team-specific way that breaks consistency”

To improve the consistency of the presentation, a data governance model that has been implemented throughout the organization as well as various data warehousing solutions were often mentioned.

> “Data governance should be improved”

> “Centralized data warehousing solution helps”

> “Centralized data warehousing can at least help as well as widely deployed data governance”

4.10 Concise representation

Concise representation also sparked debate and was identified as a challenge even though it was not chosen as one of the three weakest dimensions by any interviewee. Issues affecting the concise presentation were reported to be related to data governance models and unreliable documentation.

> “Affects a lot if data governance is deficient or if documentation is lagging”

> “Data governance is important here too”

Data engineering was generally considered the most time-consuming part of AI projects. Filtering and organizing data masses was often seen as the biggest work in an AI project.

> “Data engineering may well take 80% of the time in developing a machine learning model”

> “Organizing and making data usable still takes up most AI projects today”

To improve concise representation, in addition to data governance and up-to-date documentation, e.g., standardized data creation and common practices were mentioned.

> “It helps if standardized codecs are used, although they can be hard to remember if there are a lot of codes”

> “I would say here again that well-implemented data governance at least enhances this”

In general, improving concise representation was considered as normal work in what is done in each project and was not perceived as a major problem from a data quality perspective. The situation was perceived to be better when more data is available that could be modified afterwards.

> “The more data and the more accurate the data is, the better in principle”

> “The richer the data, the better”

4.11 Reputation

The reputation of the data was not seen as a big problem from a data quality perspective either. In general, the opinion was that the reputation of data might need to be improved if the data had been produced by a device.

> “Data produced by an industrial device is generally more reliable than data produced by a human”

> “IoT-type data often has a good reputation”

> “Reliability is largely related to the way the data is produced and relies on the process because some data producers are more reliable than others”

Better management of master data and the construction of reliability monitoring were proposed to improve the reputation of data.

4.12 Objectivity

The objectivity of the data was widely recognized as a dimension affecting the quality of the data and objectivity of the application developed. The dimension was not chosen as one of the weakest, but its importance was identified especially through biased data. In addition, its difficult observation during AI development was identified.

> “The problem is non-comprehensive data, in which AI learns selectively”

> “Difficult to detect in development”

> “Although data is based on facts, data can be very subjective”

> “For example, the layout of questions can have a big impact on the outcome”

> “Bias is a big problem that is often not identified”

> “Although the information is accurate and correct, the data may not be fair and open-minded”

> “Bias can be challenging to identify when developing the application”

To improve objectivity, the importance of understanding the context, the importance of testing and the understanding of the content of the data in making the decision were mentioned.

> “Enter the data into the application that you want to use in decision making”

> “To avoid e.g., sexual discrimination, you may not even want to include data about gender if it is not relevant from the application point of view”

> “Recognizing a data bias requires a lot of understanding of the context”

4.13 Security

The security dimension was not felt to have much of an effect on data quality. In general, data security was not seen to compromise data quality in AI development.

> “Not so much a matter affecting quality”
>
> “Often moderately ok”
>
> “Doesn’t really affect my work”

4.14 Believability

The dimension of believability was also perceived to be a part of the other dimensions and there was little discussion about it. The origin and source of the data were perceived to affect credibility. In general, the data produced by the devices was perceived as more credible from which errors are easier to detect.

> “A lot depends on where the data comes from”
>
> “Metric data is generally credible and easier to detect”
>
> “User-generated data can sometimes cause problems”
>
> “To improve the credibility of the data, it is good to be able to evaluate the data on a case-by-case basis”

5 Conclusion

The purpose of the study was to investigate data quality problems in AI development in Finland and to consider possible development measures to improve data quality. The quality of the data and its most common challenges and development proposals were investigated using both the literature and the interview material. The steps of data quality methods are generally divided into three parts: quality state reconstruction, measurement and evaluation and development proposals (Batini et al. 2016, 66). This study focuses mainly on the methods and evaluation of data quality measurement and development suggestions, as the phenomenon was intended to be studied at a general level and was not conducted for a specific organization.

The thesis and its topic were moderately challenging due to its general nature. The subject area is very broad, and it would have been easier to do a traditional and well-defined case study on a specific case. On the other hand, the purpose of the study was to gain a broad understanding of the challenges of data quality, in which case a single case would give a narrower view of the subject. Although the topic of the study seemed broad at times, the challenges in the data quality dimensions that came up again and again in the interviews strengthened the understanding of the challenges of the dimensions, which are repeated in almost every project. The amount of material in the theory part was comprehensive and high-quality, which made it easier to get an overview of the evaluation and measurement tools of data quality. Getting a clear overall picture was necessary to do the study because there is no single correct definition or way to measure data quality in the literature.

The results of the study provided subjective opinions of data quality challenges and development measures, which were repeated in several interviews. The data quality dimensions helped to approach the topic in a systematic way, which helped to form a more comprehensive overall picture of the research topic. Non-dimensional causes, which may lead to a weakness in the dimension itself, such as the gap between business and technical staff and the inability of the business to understand the benefits of AI to identify the right kind of data, became key findings. It was also possible to identify potential problems in the use of data quality dimensions in measuring data quality. Although the dimensions of quality highlight important issues, they are very open to interpretation and as such do not consider the context of the use case.

As a development proposal, the research shows the need to develop suitable metrics for measuring the dimensions of data quality in companies. Metrics that measure the quality of data and how well they fit into the context of the measured object can be considered more important than the dimension itself. Improving the quality of data can also be said to be a continuous process in which data and its quality must be evaluated and improved cyclically. Implementation and development of

data maturity models are recommended. It is also worth allocating sufficient resources for the development and evaluation of the quality of data. The results of the study also show the need for roles that can act as "interpreters" between the company and the technical implementers, clarifying the business problem to the technical staff and exploring the company's potential to solve the problems.

The research complements the data quality research field with a new perspective, in which professionals working in the field are better informed about data quality weaknesses and development proposals at a general level without focusing on one specific case.

5.1 Answers to research questions

The purpose of the study was to find out what data quality is and what weakens data quality in AI development. In addition, efforts were made to identify possible development proposals to minimize the challenges. The first research question was:

5.1.1 What is good data quality in AI development?

The first research question was answered in Chapters 2.1 and 2.2 and in Subsections 2.2.1, 2.2.2, .2.2.3 and 2.2.4. To answer the question, it was necessary to understand what good quality data is and how it is measured and evaluated. In the theoretical part, the definition of data quality and what data quality consists of was first clarified. To determine this, data quality can be divided into different subjective and objective dimensions. The theoretical part also introduced the different frameworks for measuring data quality, which define the dimensions to be used in more detail.

The definition of data quality “fitness for use”, which is widely accepted in the literature on the subject, is also well suited as a definition of data quality for AI development. Data quality measures how well data is suitable for its intended use. The quality of the data depends a lot on its use, in which case the same data may be sufficient in another system, while for another purpose it may be considered insufficient from the point of view of data quality. The key is to be able to meet the requirements of the current use and that the data represents what it is intended to represent. Data can also be evaluated from the perspective of its production as well as the end user.

The dimensions of data quality measure the quality of data from a particular perspective. What’s more important than the data quality dimensions are choosing the metrics that best serve the data quality dimension from the perspective of the current context. The selection of metrics should consider several different factors, such as the priority of the meter, the measurement method, the frequency of measurement, the cost-benefit ratio, and the risk of disregard. It is also important to be able to measure data quality objectively and subjectively to obtain the most comprehensive result.

The second research question in the study was:

5.1.2 What are the most recurring problems within data quality in AI?

Because the literature does not recognize a single established method for measuring data quality, the AIMQ method, which includes the most common dimensions of data quality, was chosen as the reference framework to support the study. The dimensions of the AIMQ method formed the frame for the interviews, which allowed the most common challenges in data quality to be explored.

The research question was answered in more detail in the results of the study in Chapter 4. The most challenging dimensions of data quality were relevancy, completeness, accuracy, understandability, and accessibility. In general, data quality problems often reflected a lack of understanding between business and technical developers and a lack of knowledge of the context of the application.

The incompleteness of the data governance model and the lack of documentation were also mentioned as common problems, because of which time is spent on finding the right kind of data and editing the data. In addition, problems presented in the literature, such as data fragmentation of systems and lack of control over data production, also emerged in the interviews.

5.1.3 How to avoid the most frequent problems in data quality?

It is important for a business to know what AI can enable while it is important for technical people to be able to understand the challenges of a business from a technical and data perspective. In addition, the data needed by AI applications should be comprehensive enough. This can be supported by considering what kind of needs will be seen in the future. In this way, the necessary data can be collected in time.

Achieving good quality data also requires the organization to have effective data governance, such as defining and adhering to data-related rules and responsibilities. The responsibility for the data should be with the business units that use the data, as they also have the best knowledge of the needs. It is important to start by identifying the business need that is being addressed. After that, a technical implementation is chosen, such as developing an AI application if it can solve the problem. Determining a business need requires an understanding of the context. Only after defining the business need can the variables of the data be defined. It is also good for business and technical people to try to think about future needs from a data perspective so that the data can be collected comprehensively enough for possible future needs.

The importance of the data governance model and documentation and the fact that the data governance model has been widely implemented throughout the organization was also highlighted. In addition, documentation about the data, such as data catalogs and metadata systems are useful and should be kept up to date. Still, most of the work of technical developers goes into data engineering, which involves finding, editing, and processing data before the work associated with the actual AI model can be started.

In general, it can be said that it is good to have a lot of data rather than too little and that the data should be as rich as possible.

5.2 Research Evaluation, Validity and Limitations

The main benefit of the research is to increase the understanding of the possible causes of data quality problems, as well as development suggestions that can improve the AI capabilities of companies. Data quality studies have been conducted in Finland before, but more specific reasons for data quality have largely been derived from a case study, which increases the value of this study by providing a broader picture of the data quality challenges compared to individual dimensions.

Examining data quality without a specific case is challenging because the assessment of data quality depends largely on the case in which the data is used. In addition, more important than the dimensions of data quality is to think about what kind of metrics should be used to measure the dimension, which also depends a lot on the purpose and context of the application. The interpretation of data quality dimensions through qualitative interviews is also very subjective and based on the user's personal experience. On the other hand, the interviewees were carefully selected and had vast experience in the field and had seen numerous projects on the topic, which increased the interviewers' perception of the topic.

The qualitative aspect of research can be assessed through its validity and reliability (Tuomi & Sarajärvi 2018, 72). Validity measures whether a promised issue has been investigated in a study, while reliability refers to the reproducibility of research results. Reliability can be used to ensure that research does not give random results (Hirsjärvi et al. 2007, 32).

To improve the validity of the study, the interview questions were made easy to understand and were also explained more extensively during the interviews. In addition, pre-information was sent to the interviewees, explaining the purpose of the study and the data quality dimensions to be reviewed, including the definitions, which in turn increases the reliability of the study (Hirsjärvi et al. 2007, 41). The literature on research theory is based on the theory of measuring and assessing data quality, so the presented evaluation and measurement methods can be assumed to be related to the subject under study. The interview frame was defined using the AIMQ framework,

which provided a sufficiently comprehensive selection of the dimensions, as some of the dimensions contributed to similar responses as other dimensions. There was a total of 7 interviewees, all of whom had worked as consultants at some point in their careers. Although the number of interviewees could have been larger, the interview provided a comprehensive picture of the data quality challenges in AI development because each interviewee had already done numerous projects, with numerous different employers or clients. This made it possible to obtain a comprehensive picture because the purpose of the study was to examine the phenomenon at a more general level than focusing on only one specific case.

The aim was to improve the internal validity of the study during the analysis phase of recurring problems and development proposals by classifying related problems as separate themes (Yin 2018, 65). The analysis phase combined the findings by forming possible cause-and-effect relationships from the collected material. In addition, the analysis allowed room for different opinions on topics, avoiding over-generalization (Yin 2018, 72). Answers to the research questions were also sought from the literature, which were compared to the information obtained from the interviews.

The reliability of the study was considered by conducting the study based on a pre-prepared research plan, which defined the data collection and the main features of the report. The interviews conducted in the study were recorded during the spelling and writing of the report. In addition, direct citations were taken from the interview material. In addition, content analysis has been illustrated with examples from interviews.

5.3 Future Research Proposals

Several research topics can be raised from the results of the study. With the expansion of artificial intelligence into many different fields and due to the rapid development of technology, similar research could be focused on one specific area of business economics. In addition, the rapid development of artificial intelligence would require the research to be repeated at regular intervals.

As the study focuses on common problems in data quality, future research could also focus on how to increase understanding between business and technology and how to better prepare for data collection for future needs in AI development. Applying a data governance model and researching its implementation would also add value to the research area.

In addition, several methods have been developed to assess and develop data quality, but they haven’t been widely applied in practice since the tests of the original studies (Batini et al. 2016, 102). It could also be interesting to compare different methods in practice to obtain more empirical results on their functionality.

References

Alho, T., Neittaanmäki, P., Hänninen, P. &Tammilehto, O. 2018. Humanoidirobotti Pepper: Mahdollisuuksia ja haasteita. University of Jyväskylä. Informaatioteknologian tiedekunnan julkaisuja 61/2018. https://www.jyu.fi/it/fi/tutkimus/julkaisut/tekes-raportteja/humanoidirobotti_pepper_mahdollisuuksia_ja_haasteita_verkkoversio.pdf Accessed: 4.4.2022.

Aljumaili, M., Karim, R. & Tretten, P. 2016. Metadata-based data quality assessment. VINE Journal of Information and Knowledge Management Systems, 46, 2, 232-250.

Andreescu, A.I., Belciu, A., Florea, A. & Diaconita, V. 2014. Measuring Data Quality in Analytical Projects. Database Systems Journal, 5, 1, 2-7.

Anderson, J. & Coveyduc, J. 2020. Artificial intelligence for business: a roadmap for getting started with AI. John Wiley & Sons, Inc. New Jersey.

Batini, C., Cappiello, C., & Francalanci, C. 2009. Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys. 41, 3, 14-24.

Batini, C., & Scannapieco, M. 2016. Data and Information Quality. Springer. Rome

Ballou, D. & Pazer, H.L. 1985. Modelling data and process quality in multi-input, multi-output information systems. The Institute for Operations Research and the Management Sciences. 31, 2, 123-248.

Ballou, D., Wang, R., Pazer, H. & Tayi, G.K. 1998. Modeling Information Manufacturing Systems to Determine Information Product Quality. The Institute for Operations Research and the Management Sciences. 44, 2, 433-594.

Boyadzhieva, D. & Kolev, B. 2010. Intuitionistic Fuzzy Data Quality Attribute Model and Aggregation of Data Quality Measurements. Studies in Computational Intelligence. Springer-Verlag. Berlin.

Bronselaer, A., Nielandt, J. & De Tré, G. 2018. An incremental approach for data quality measurement with insufficient information. International Journal of Approximate Reasoning. 96, 18, 34-46.

Cappiello. C., Francalanci. C., Pernici. B. 2004. Data Quality Assessment from the User's Perspective. Information Quality in Information Systems. 4, 7, 16-42.

Combs, V. 2021. Gartner: AI is moving fast and will be ready for prime time sooner than you think. TechRepublic. https://www.techrepublic.com/article/gartner-ai-is-moving-fast-and-will-be-ready-for-prime-time-sooner-than-you-think/. Assessed: 26.3.2022.

Dorr, B. & Murnane, R. 2011. Using Data Profiling, Data Quality, and Data Monitoring to Improve Enterprise Information. Software Quality Professional. 13, 4, 2-14.

Ehrlinger, L., Rusz, E., & Wöß, W. 2022. A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in Big Data. 5, 1, 1-5.

Eskola, J. & Suoranta, J. 2014. Johdatus laadulliseen tutkimukseen. Vastapaino. Tampere.

FAIA. 2018. FAIA Playbook – Liikkeelle Tekoälyn Hyödyntämisessä. Suomen Tekoälykiihdyttämö. Helsinki. https://faia.fi/guides/. Accessed: 12.7.2022.

Haaga-Helia. 2021. AI-TIE -Tekoälyinnovaatioekosysteemillä kilpailukykyä PK-yrityksille. https://www.haaga-helia.fi/fi/hankkeet/ai-tie-tekoalyinnovaatioekosysteemilla-kilpailukykya-pk-yrityksille. Accessed: 13.8.2022.

Haug. A, Arlbjorn. JS & Pedersen. A. 2009. A Classification Model of ERP System Data Quality. Emerald Group Publishing Limited. 109, 8, 1053-1068.

Heinrich, B., Klier, M., Schiller, A. & Wagner, G. 2018. Assessing data quality – A probability-based metric for semantic consistency. Elsevier. 110, 5, 56-92.

Hirsjärvi, S., Remes, P. & Sajavaara, P. 2007. Tutki ja kirjoita. Tammi, Helsinki.

Håkansson, A. & Hartung, R. 2020. Artificial Intelligence. Studentlitteratur. Lund.

Kananen, H. & Puolitaival H. 2019. Tekoäly – Bisneksen uudet työkalut. Alma Talent Oy. Helsinki.

Korpela, J. 2018 May. Mitä on tiedon laatu? Webinar. https://aureolis.com/bi-akatemia/kiitos-webinaaritallenne-tiedon-laatu/. Assessed 7.6.2022.

Laihonen, H., Hannula, M., Helander, N., Ilvonen, I., Jussila, J., Kukko, M., Kärkkäinen, H., Lönnqvist, A., Myllärniemi, J., Pekkola, S., Virtanen, P., Vuori, V., & Yliniemi, T. 2013. Tietojohtaminen. Tampereen teknillinen yliopisto - Tiedonhallinnan ja logistiikan laitos. Tampere. https://core.ac.uk/download/pdf/250168716.pdf. Accessed: 25.8.2022.

Lee, Y.W., Strong, D.M., Kahn, B.K. & Wang, R.Y. 2002. AIMQ: a methodology for information quality assessment. Elsevier. 40, 2, 32-102.

Loshin, D. 2001. Enterprise Knowledge Management: The Data Quality Approach. Morgan Kaufmann. United States of America

Lähteenmäki, P. 2017. Koko Suomi tekoälyoppiin. Talouselämä, 38, 2, 24–31.

Mahanti, R. 2014. Critical Success Factors for Implementing Data Profiling: The First Step Toward Data Quality. Software Quality Professional. 16, 2, 14-32.

Maydanchik. A. 2007. Data Quality Assessment. Technics Publications, LLC. United States of America.

Marr, B. 2020. The Intelligence Revolution: transforming your business with AI. Kogan Page. London.

McGilvray, D. 2008. Executing data quality projects: ten steps to quality data and trusted information. Morgan Kaufmann. United States of America.

Microsoft News Center. 2018. Artificial Intelligence in Europe – Finland. https://news.microsoft.com/fi-fi/2018/10/26/suomi-euroopan-karkea-tekoalyn-hyodyntamisessa-silti-puolet-yrityksista-vasta-pilotointivaiheessa/. Accessed: 4.6.2022.

MinnaLearn & University of Helsinki. Elements of AI 2018. https://course.elementsofai.com/fi. Accessed: 22.8.2022.

Ojasalo, K., Moilanen, T. & Ritalahti, J. 2015. Kehittämistyön menetelmät – Uudenlaista osaamista liiketoimintaan. Sanoma Pro Oy. Helsinki.

Pipino. L, Lee. W, Wang. R. 2002. Data Quality Assessment. Communications of the ACM. 45, 4, 113-145.

Rouhiainen, L. 2018. Artificial Intelligence, 101 things you must know today about our future. Amazon Books. United States of America.

Saunders, M. N. K., Thornhill, A., & Lewis, P. 2019. Research Methods for Business Students. Pearson Education Limited. United States of America.

Shankaranarayan, G., Ziad, M., & Wang, R. Y. 2003. Managing data quality in dynamic decision environments: an information product approach. Journal of Database Management. 14, 4, 1-19.

Sebastian-Coleman, L. 2013. Measuring Data Quality for Ongoing Improvement. Morgan Kaufmann. United States of America.

Silvola, R., Jaaskelainen, O., Hanna Kropsu-Vehkapera, & Haapasalo, H. 2011. Managing one master data - challenges and preconditions. Industrial Management & Data Systems. 111, 1, 123-162.

Tayi, G., & Ballou, D. 1998. Examining data quality. Communications of the ACM. 41, 2, 54-57.

TEM. 2017. Suomen tekoälyaika - Suomi tekoälyn soveltamisen kärkimaaksi: Tavoite ja toimenpidesuositukset. Työ- ja elinkeinoministeriö. http://urn.fi/URN:ISBN:978-952-327-248-4. Accessed: 6.6.2022.

Tuomi, J. & Sarajärvi, A. 2018. Laadullinen tutkimus ja sisältöanalyysi. Tammi. Tampere

Umar, A., Karabatis, G., Ness, L., Horowitz, B. & Elmagardmid, A. 1999. Enterprise Data Quality: A Pragmatic Approac. Information Systems Frontiers. 1, 279-301.

Wand, Y. & Wang, R.Y. 1996. Anchoring data quality dimensions in ontological foundations. Communications of the ACM. 39, 11, 86-95.

Wang, R. 1998. A product perspective on total data quality management. Communications of ACM, 41, 2, 58-65.

Wang, R. & Strong, D. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 12, 4, 5-33.

Woodall, P., Borek, A., & Parlikad, A. K. 2013. Data quality assessment: The Hybrid Approach. Elsevier. 50, 7, 369-382.

Yin, R.K. 2018. Case Study Research: Design and Methods. Sage Inc. https://books.google.fi/books?hl=fi&lr=&id=FzawIAdilHkC&oi=fnd&pg=PR1&dq=+Case+study+research+and+applications:+design+and+methods&ots=l_3X89hU1t&sig=jUD8ML0jOfzDBmG54EeqTSnTtNU&redir_esc=y#v=onepage&q=Case%20study%20research%20and%20applications%3A%20design%20and%20methods&f=false. Assessed: 1.6.2022.

Appendices

Appendix 1. Questions sent to interviewees prior to the interview

What are the most common challenges for each of the data quality dimension listed below?

Share some good experiences and practices to improve each data quality dimension stated?

Dimension	Explanation
Relevancy	Data is essential to the task. Data is useful and suitable for work.
Completeness	All the necessary values of the data are included. Data covers the needs of tasks.
Free of Error	Data is flawless and accurate.
Understandability / Interpretability	Data is easy to interpret / Data is clear and easy to internalize. The units of measurement of the data are clear.
Accessibility	The data is available to consumers. Data is available and accessible.
Timeliness	The data is not old. The data contains up-to-date and valid information.
Appropriate amount	There is neither too much nor too little data. There is not too much data, and the amount of data meets the needs.
Ease of Operation	It is easy for the user to take advantage of the data. The data is easily editable and combined.
Consistent representation	The data is succinctly presented. Compact design.
Concise representation	The data is presented in the same format.
Reputation	Data is trusted. The data comes from good sources and has a good reputation.
Objectivity	The data is fair and open-minded. Data is based on facts.

Security	The data can be accessed by the right parties. The data is protected at an adequate level.
Believability	The data can be considered accurate. It is credible and reliable.

At the end of the interview, I ask you to select 3 dimensions that usually affect the quality of the data most negatively.