# A process model for organizational data quality assessment

van Wierst, J.W.G.

*Award date:*
2019

Eindhoven, December 2018

# A Process Model for organizational Data Quality Assessment

By
J.W.G. (Joost) van Wierst

Student Identify number: 0819233

In partial fulfillment of the requirements for the degree of
**Master of Science**
**In Operations Management and Logistics**

Supervisors:
Dr. B. Ozkan (first supervisor), TU/e, IS
Dr. Ir. H Eshuis (second supervisor), TU/e, IS
M. Verschuren, ASML
W. Peperkamp, ASML

**TU/e, School of Industrial Engineering**

**Series Master Thesis Operations Management and Logistics**

Key words: Data quality assessment, data quality methodology, data quality measurement, data quality framework

# Management Summary

As the amount of available data is increasing exponentially, data-driven decision making is a rapidly growing phenomenon in today's organizations. The quality of this data is paramount to its success, and poor data quality can have disastrous consequences. Data quality is therefore becoming an important competence, and an increasingly interesting topic in research. Current research on data quality provides a variety of methodologies and frameworks. Often these methodologies consist of both data quality assessment and data quality improvement. This study focusses on data quality assessment: "the process of obtaining measurements of data quality to determine the current state of data quality" (Woodall et al., 2013). Based on such initial assessment, improvement plans can be made that balance data quality levels, costs, resources and capabilities across an organization or department.

## Problem and methodology

The majority of data quality frameworks and methodologies are either developed for a specific context, technique or problem, or they provide a generic assessment that often lacks practical guidance and is not operationalized for a specific context. This may cause organizations to adopt a data quality assessment methodology that does not suit their needs and current situation. Operationalizing a data quality assessment framework to a specific context requires the definition of data quality (i.e. customizing the selection of dimensions and subsequent measures) to be part of the assessment process, instead of using predefined fixed sets as is often suggested in generic methodologies. This study addresses this gap between data quality assessment research and practices. As existing generic methods (i.e. regardless of context) often lack practical guidance, the goal of this study is to enhance how-to knowledge of applying the critical activities of data quality assessment in a specific context, and to improve the ability of data quality practitioners to obtain a complete assessment of their data quality. This goal is achieved by designing a generic, but highly practical process model for data quality assessment. Since the goal of this research is to design and develop an artefact, it has a design science approach. Peffers et al., (2007) developed a methodology for design science research for information system research. This methodology served as a guide for this research. This methodology started with a problem identification and the motivation for the research. Then, objectives for the solution were defined. The solution objectives defined for the process model of this research can be found in Table 1.

| Objective | Reasoning | Relation to research problem |
|---|---|---|
| **Practical utility** | Any vagueness on how to conduct the activities in the designed process model must be eliminated | Existing generic data quality assessment methodologies often lack practical guidance. |
| **Comprehensiveness** | A generic process model should be comprehensive to be applicable independent of context. | A generic process is often not practical for specific contexts. |
| **Genericness** | The designed process model must be applicable independent of any context | A generic but practical model for data quality assessment is missing |
| **Understandability** | The process model must be presented in an understandable format | For a process model to be practical, it must be well understandable |
| **Completeness** | The final assessment must give a complete overview of the current state of data quality in a specific context | Existing methodologies often do not fit specific business needs, and may therefore give incomplete or irrelevant results |

*Table 1: Solution objectives for a process model*

The development of the process model required two research questions to be answered. For the process model to be practical, a clear definition of the roles that participate in the process and in what activities they are involved is required. Also, for the process model to be comprehensive, all critical activities of data quality assessment must be included. This lead to the following sub research question for this research:

- What are the critical activities in a generic data quality assessment process?
- What roles need to be assigned to these activities to effectively perform the data quality assessment process?

To answer these questions, a literature review is conducted, following by a synthesis of this literature. In this literature review, relevant existing data quality assessment methodologies (on its own or as a part of a bigger data management approach) are collected and analyzed on both the activities that they contain and, if any, the roles that they define. In the synthesis, the aim is to group both activities and roles across the different methodologies based on their similarity. This grouping is direct input for the identification of critical activities and roles. After synthesizing the literature, the actual process model is designed considering the critical activities and roles identified in the synthesis. During this design, the earlier defined solution objectives, that represent design goals, are taken into account. BPMN is chosen is used as the modeling language for the process model, as BPMN is activity based and allows for visually depicting both information flows and roles.

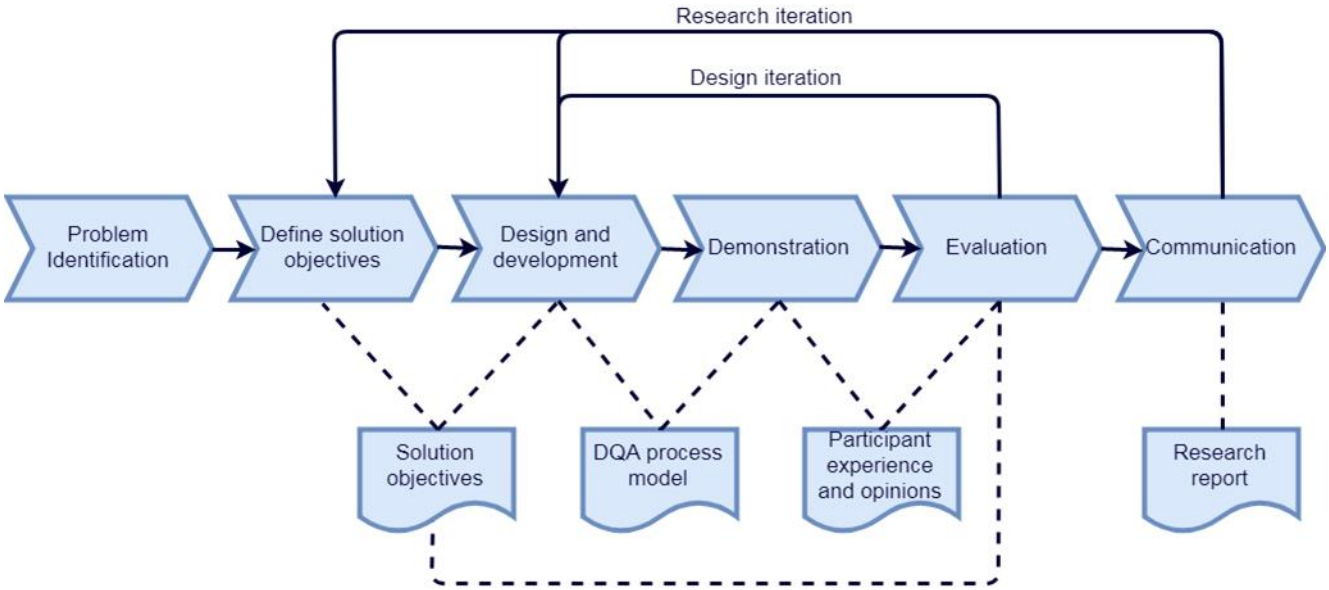



*Figure 1: Research Method*

## Demonstration

The process model is demonstrated in a case study at ASML. At ASML, they currently recognize the necessity to improve the quality on cycle time and labor hour data of the activities that are performed in the EUV factory. Although they have identified the root causes of data quality problems in a previous project, they miss an extensive data quality assessment method. The case of ASML provides the perfect opportunity to propose a data quality assessment process and to serve as the validation for this research. The case study was performed in a period of eight weeks. In total, there were 11 employees that participated in the process: 1 data quality expert, 1 data expert, and 9 data consumers. The process model

resulted in a measurement model consisting of 8 dimensions and 36 metrics. Three of these dimensions were measured subjectively, using questionnaire items, the other five dimensions were measured using objective measures.

## Evaluation

The last step of the research is to evaluate the process model and its results. The previously defined solution objectives are evaluated based on the observations and results during the demonstration. Since the defined solution objectives for this research mainly reflect qualitative characteristic (i.e. they are determined by the experiences and opinions of participants of the process model), a qualitative evaluation is deployed. This qualitative evaluation is achieved by performing semi-structured interviews with participants of the process in the case study. Semi-structured interviews are chosen for this evaluation as they allow for obtaining comprehensive experiences and opinions regarding the use of the process model for each of the solution objective. Three participants of the case-study were interviewed for evaluation. The evaluation of the proposed model showed that the model was considered practical, comprehensive, generic, understandable and complete by participants of the case study, indicating that the model is a solution to the research problem, and a valuable contribution for data quality practitioners in the field. However, also problems were identified for each solution objective, providing options to further improve the model. Further improvements were identified as possible subjects for further research including a configuration guide, the application of a data quality reference model, and the development of a normalization method for metric scores. More simple improvements included the addition of an extra validation loop after obtaining objective measurements and the inclusion of data collectors for data quality problem identification.

# Preface

This Master thesis is submitted for the Master program 'Operations Management & Logistics' at the Eindhoven University of Technology (TU/e) and has been performed for the Information Systems Group of the faculty of 'Industrial Engineering & Innovation Sciences'.

This master thesis is the final deliverable of six years of studying at the University of Technology Eindhoven. The last months in which this thesis was written have been both challenging and full of learning experience. I would like to take the opportunity to express my gratitude towards several persons who contributed to this thesis and provided support throughout the process. First, I would like to thank Baris, the first supervisor of the research and my mentor, for his valuable input in the thesis, the time that he made for me and for guiding me into the right directions. I want to thank Rik, my second supervisor, for his feedback and valuable input for improvement. Second, I want thank my company supervisors: Mark, for all his time, input and meetings that we had, Wout, for his supervision on the project, and Jelle, for giving me the opportunity to do this project.

Finally, I want to thank my friends, family and girlfriend for supporting me throughout the months of my thesis. It has been stressful at times, but your distractions and nice words kept me motivated to achieve a good result.

Joost van Wierst,

November 2018

# Table of Contents

# List of Figures



# List of Tables

# 1. Introduction

As the world is moving towards the big data era, data quality is becoming increasingly important for every organization (Abbasi et al., 2016). With the upswing of technologies such as cloud computing, the Internet of Things and social media, the amount of data being generated is increasing exponentially (Cai & Zhu, 2015). The enormous amount of data available in many forms forces organizations to come up with innovative ideas to find structure in this data and to deal with quality issues (Albala, 2011). Unstructured data from multiple sources make data quality management become a complex process. The causes of poor data quality are numerous: data entry by employees, external data sources (for example the web), poor data migration processes and system errors are some of them (Eckerson, 2002). As the amount of data being captured in organizations, stored in data warehouses, and mined for competitive use exploded over the last decades, maintaining the quality of it in order to support business processes is important, but difficult (Cappiello et al., 2004; Heinrich et al., 2009). The 'quality vs quantity' challenge is increasingly recognized by organizations (Kaisler et al., 2013): often, more data is considered more value, but this is not always true as more data can cause uncertainty and confusion if the quality of it is poor. Although maintaining high quality data is a challenging task for many businesses, it is a valuable asset. High quality data has become a prerequisite for world-wide business process harmonization, global spend analysis, integrated service management and compliance with regulatory and legal requirements (Hüner et al., 2009). Research shows that data quality has a critical impact on achieving strategic and operational business goals; high quality data positively impacts decision-making (Shankaranarayan et al., 2003), customer relationship management (Reid & Catterall, 2005) and supply chain management excellence (Kagermann et al., 2011). Decision making based on data is a rapidly growing phenomenon within organizations and enables managers and decision makers to make decisions more effectively. However, making decisions can be risky when it is based on data of poor quality (Chaudhuri et al., 2011). Poor quality data affects efficiency, risk mitigation, and agility by harming the decisions to be made in each of these areas (Friedman & Smith, 2011). In his paper, Redman (1998) aims to create awareness of the problem of poor data quality since the late 1990's. He classifies the impacts on poor data quality intro three levels: on the operational level poor data quality directly leads to customer dissatisfaction, increased costs and lowered employee satisfaction. On a tactical level poor data quality affects decision-making, the ability to reengineer, and internal organizational mistrust. Finally, on the strategic level, Redman argues that poor data quality makes it more difficult to set and execute a strategy.

These developments and research findings emphasize the increasing importance of data quality. Data quality therefore is becoming an increasing topic of interest in research. Data quality research areas involve among others data quality dimensions, models, techniques for measurement and improvement, tools and frameworks and methodologies (see literature review). Jaya et al., (2017) argue that data quality management models and data quality assessment methods are the essential deliverables in data quality research.

## 1.1. Scope of the study

The increasing need for high quality data has led to the definition and development of many data quality management models in the literature (see for example Total Data Quality Management (Wang, 1998), ISO 8000-61 (2016), or DAMA-DMBOK Guide (Dama International, 2009)). Although adopting different names, data quality management models often consist of comparable phases: the general approach to data quality management is a version of the iterative Deming cycle (Deming, 1986), better known as "plan-do-

check-act". See for example the data quality management process as defined in ISO 8000-61 (Figure 1.1). Typically, the plan-do-check-act translates to data quality management as follows:

- *Plan*: The plan phase of data quality management includes establishing data requirements and objectives for data quality, creating plans to meet these objectives and evaluating the performance of these plans. These plans aim to balance data quality levels, costs, resources and capabilities across an organization or department. The inputs for this phase are stakeholder needs and expectations and the feedback obtained from the act phase.
- *Do*: The do phase involves creating, using and updating data according to specified work instructions to deliver data that meet the requirements (defined in the plan phase). This phase also includes monitoring the quality by checking whether the data conform to pre-determined specifications (the required characteristics of data, based on the requirements).
- *Check*: The check phase measures the data quality levels and process performance related to data nonconformities or other issues that have arisen as a result of the plan or control phase. This measurement provides evidence by which to evaluate the impact of any identified poor levels of data quality on the effectiveness and efficiency of business processes. It consists of reviewing data quality issues, creating measurement criteria and an evaluation of results.
- *Act*: The act phase includes analyzing the root causes of data quality issues based on the results of the check phase. Based on this analysis, this phase corrects existing nonconformities and appropriately transforms processes to prevent future nonconformities.

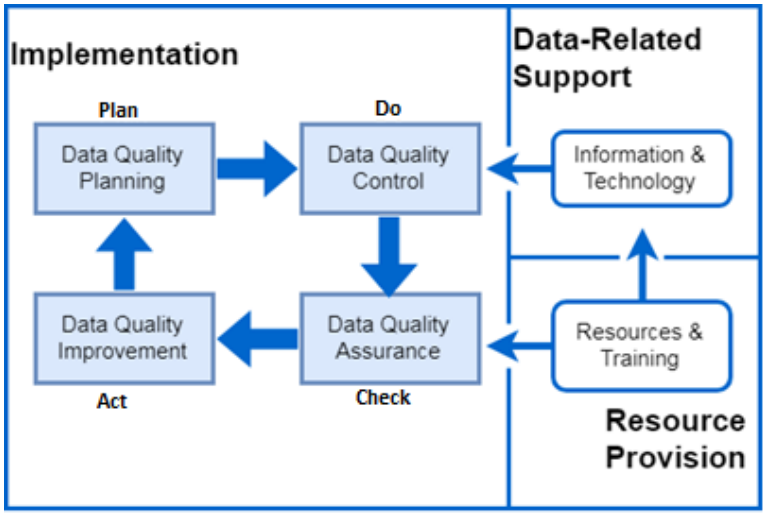



*Figure 1.1: Process reference model for data quality management (adapted from ISO 8000:61)*

However, as Stvilia et al., (2007) argue: "one cannot manage data quality without being first able to measure it meaningfully", which highlights the importance of data quality assessment before data quality control and improvement. Therefore, before an iterative data quality management process like ISO 8000:61 can be used, it is important to assess the current level of data (i.e. measure how well objectives and requirements are met) such that meaningful and effective improvements can be identified. In her book, Sebastian-coleman (2013) defines four data quality assessment scenario's, all having various assessment objectives (see Table 1.1).

| Assessment scenario | Goals | Deliverables |
|---|---|---|
| **Initial assessment** | • Obtain knowledge of data and the processes that produce it<br>• Identify data to be measured on an ongoing basis<br>• Measure baseline condition critical data | • Measurement results<br>• Improved data definitions<br>• Recommendations for ongoing measurements |
| **Improvement projects** | • Implement changes in data capture and processing<br>• Show measurable improvement over previous state | • Documented process changes<br>• Measurements showing data quality improvement |
| **Ongoing Measurement** | • Ensure that data continues to meet expectations<br>• Investigate changes in data quality patterns<br>• Identify opportunities for improvement | • Action plans for further improvement<br>• Reports on changes in data quality patterns |

*Table 1.1: Data quality assessment scenarios, adapted from Sebastian-coleman (2013)*

Considering these scenarios, the initial assessment is the topic of this study. Such initial assessment contributes to an effective execution of data quality management practices, as it provides a clear definition of data and related business processes, meaningful measures for data quality control, and a baseline condition of critical data. This study adopts the definition of a data quality assessment that is provided by Woodall et al., (2013): "a data quality assessment is the process of obtaining measurements of data quality to determine the current state of data quality". For this study, the following components are considered a part of a data quality assessment process:

- Obtain knowledge of data and the processes that produce it (corresponding to the goals of initial data quality assessment).
- Establishing data quality requirements and objectives (corresponding to a part of the plan phase in the ISO 8000:61 data quality management reference model).
- Measure the baseline condition of critical data (corresponding to the goals of initial data quality assessment) by measuring data quality levels using metrics that are defined based on data quality requirements, objectives and data quality issues (corresponding to parts of both the do and the check phase of the ISO 8000:61 data quality management reference model).

Based on such initial assessment, improvement plans can be made that balance data quality levels, costs, resources and capabilities across an organization or department. Making such plans, and the improvements that follow from them, are not part of the scope of this study.

Furthermore, considering the three types of data quality that most authors distinguish (structured, unstructured and semi-structured, see (van Wierst, 2018), this study focusses on the assessment of structured data. Although data quality assessment for unstructured and semi structured is becoming a

more popular topic in recent research, most works focus on structured data as it is usually structured data that is to be assessed in today's organizations.

## 1.2. Motivation for the study

In the past decade, data quality has become a popular research topic. Data quality frameworks and methodologies, for both assessment and improvement, became increasingly available. However, as Woodall et al., (2013) argue, organizations have many different requirements related to data quality assessment, and the aspects of data quality that are of interest are highly dependent on the context. Organizations may be forced to adopt an assessment methodology that does not fully fit their needs and current situation.

An explanation for this is that the majority of data quality frameworks and methodologies are either developed for a specific context, technique or problem (see for example Aljumaili et al., (2016); Brown et al., (2013); del Pilar Angeles & García-Ugalde, (2009); Eppler & Muenzenmayer, (2002); Madhikermi et al., (2016); Neumaier et al., (2016); Shardt & Huang, (2013); Wan et al., (2015)), or they provide a generic assessment method (i.e. regardless of context or application) that often lacks practical guidance and is not operationalized for a specific context and business needs (for example Lee et al., (2002); Pipino et al., (2002); Wang, 1998)). Operationalizing a data quality assessment framework to a specific context requires the definition of data quality (i.e. the selection of dimensions and subsequent measures) to be part of the assessment process, instead of using predefined fixed sets (as is done in for example Cai & Zhu, (2015); Redman, (1996); Wand & Wang, (1996); Wang & Strong, (1996)). Various articles emphasize the importance of a free selection and definition of dimensions based on organizational context or business needs (De Amicis & Batini, 2004; Su & Jin, 2004; Woodall et al., 2013). This study addresses this gap between data quality assessment research and practices.

As existing generic methods (i.e. regardless of context) often lack practical guidance, the goal of this study is to enhance how-to knowledge of applying the critical activities of data quality assessment in a specific context, and to improve the ability of data quality practitioners to effectively (i.e. "doing the right things") and efficiently (i.e. "doing things right") obtain a complete assessment of their data quality. This goal is achieved by designing a generic, but highly practical process model for data quality assessment. For the process model to be generic (i.e. applicable independent of context), the inclusion of all critical activities of data quality assessment must be ensured. Additionally, for the process model to be practical, it requires a low-level definition of these activities along with a distribution of these activities among distinct roles. This requires the following questions to be answered before the design of a process model:

- What are the critical activities in a generic data quality assessment process?
- What roles need to be assigned to these activities to effectively perform the data quality assessment process?

Answering these questions provides the necessary knowledge for the development of a data quality assessment process model that is both generic but highly practical.

Besides aiming for a generic but practical process model, data quality assessment should both have a bottom up and a top-down approach: by reviewing existing methodologies, the majority can be divided over two categories (see Figure 1.2): methodologies are either problem-driven (bottom-up) or requirement- driven (top-down). A problem-driven approach aims to identify problems experienced by data consumers and creates adequate metrics that reflect these problems. Furthermore, problems can be

identified from the definition of data objects, attributes, their relations and subsequent rules (for example: the attribute "gender" can only have two values). Examples of problems driven methodologies can be found in Batini & Scannapieco (2006), Sebastian-coleman (2013), and Batini et al. (2005). On the other hand, methodologies can be requirement-driven: relevant dimensions and metrics are selected based on the functionality that data should have. This requires the identification of the goals of the tasks of data consumers related to the data and what they expect from it. Examples of requirement-driven methodologies can be found in Bicalho et al. (2017), Jeusfeld et al. (1998), Wang (1998) and Lee et al., (2002). This study aims to incorporate both approaches in a single process model.

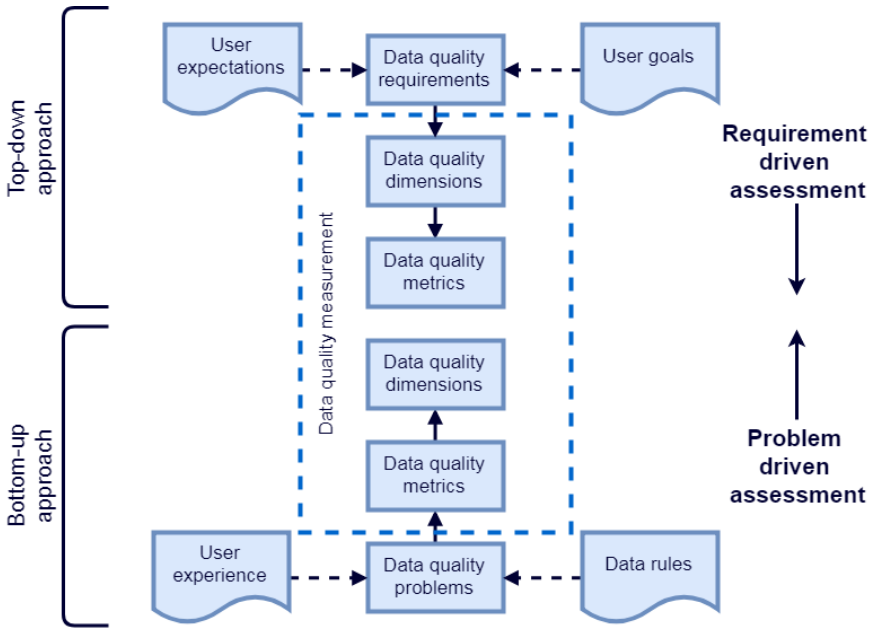



*Figure 1.2: Top-down and bottom-up approach to data quality*

## 1.3. Research design

The goal of this research is to develop a new artefact (a process model for data quality assessment) and follows a design science approach. Design Science is a research is an outcome-based research methodology, that focusses on the development of artefacts. As opposed to explanatory research, the research objectives in design science research are of a more pragmatic nature. Hevner & Chatterjee (2010) define design science as follows:

*"Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem."*

And that its first principle is:

*"The fundamental principle of design science research is that knowledge and understanding of a design problem and its solution are acquired in the building and application of an artifact."*

Peffers et al. (2007) provide a methodology for conducting design science research for the information systems discipline. Their methodology describes six steps which form the basis of the research method of this study. The first step of this methodology is to identify the problem and define the objectives for a solution. Then, an artefact is designed (the process model in this case). In order to design a data quality assessment process model, a literature review is conducted to identify the critical activities and roles in a generic data quality assessment process. A synthesis of this literature provides the input for the design of the actual process model. The application of the model is demonstrated in a case study and thereafter evaluated (using interviews with participants of the case study) based on the previously defined solution objectives. Figure 1.3 shows the process of the methodology of Peffers et al. (2007).

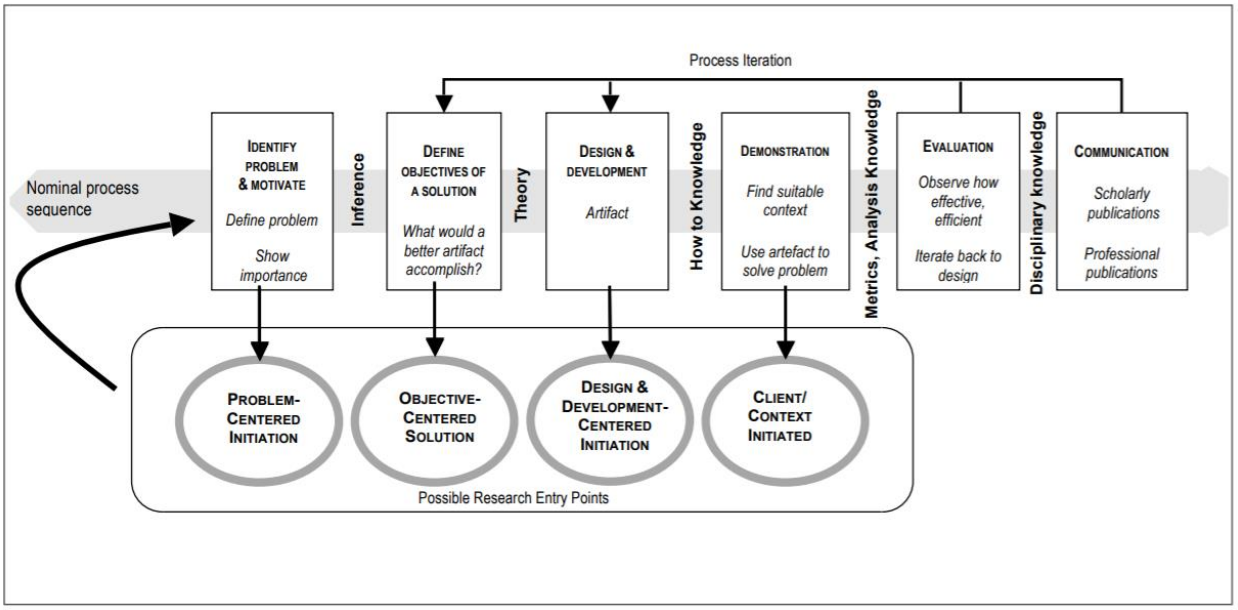



*Figure 1.3: Process model for Design Science Research (Peffers et al., 2007)*

Looking at the methodology of Peffers et al. (2007) in the figure above, there are four research entry points: a problem-centered initiation, an objective-centered initiation, a design and development-centered initiation, and a client/context-centered initiation. As this research starts with identifying and describing a problem, this research enters the methodology using with a problem-centered initiation.

# 2. Background and related work

## 2.1. Introduction to data quality

The most comprehensive definition of data quality is given by Juran & Godfrey (1998): "Data and information are of high quality if they are fit for their uses (by customers) in operations, decision-making, and planning. They are fit for use when they are free of defects and possess the features needed to complete the operation, make the decision, or complete the plan." Although throughout the data quality literature a wide range of definitions can be found, this subjective term 'fitness for use' is acknowledged by many researchers. Wang & Strong (1996) define data quality as "the distance between data views presented by an information system and the same data in the real world", indicating that data quality depends on the ability of an information system to represent real world objects. Karr et al. (2006) focus more on the functionality of data to make better decisions and define data quality as "the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions".

However, a more practical definition is needed to characterize the different aspects of data quality, and to be able to measure and assess it. Researches unanimously agree that data quality is a multi-dimensional concept, and a variety of data quality dimensions have been identified. In this section, the key data quality dimensions are presented. The dimensions presented constitute the focus of the majority of data quality researches (Scannapieco & Catarci, 2002).

### 2.1.1. Accuracy

*Accuracy* is the most widely used data quality dimensions (Huang et al., 1998), and is considered in majority of data quality methodologies. Although the definition of accuracy is often worded differently by researchers, its definition generally comes down to the following: accuracy is the closeness between a data value and the value of a real-world object that the data aims to represent. Batini & Scannapieco (2006) distinct between two kinds of accuracy:

- *Syntactic accuracy* is the closeness of a data value to the elements of the corresponding definition domain. In syntactic accuracy, a data value is not compared to the value of the real-world object it aims to represent. Rather, syntactic accuracy checks if a data value corresponds to any value in the domain that defines this data value (Batini & Scannapieco, 2006).
- *Semantic accuracy* is the closeness of a data value and the real-world object it aims to represent. In order to be able to measure semantic accuracy, the true value of the real-world object needs to be known (Batini & Scannapieco, 2006).

Batini & Scannapieco (2006) provide three measurements to calculate the weak accuracy error, *strong accuracy error* and the *syntactic accuracy*, given that correct values of the data are available.

### 2.1.2. Completeness

Wang & Strong (1996b) define completeness as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand." Pipino et al. (2002) identified three types of completeness:

- *Schema completeness* is the degree to which concepts and their properties are not missing from a data schema
- *Column completeness* is defined as a measure of the missing values for a specific property or column in a table
- *Population completeness* evaluates missing values with respect to a reference population

An important note needs to be mentioned when it comes to null values and completeness. When measuring the completeness of a table, it is important to know why a value is missing. Batini & Scannapieco (2006) argue that there are three reasons for a value to be null: either, the value is not existing (which does not contribute to incompleteness), or the value is existing but not known (which contributes to incompleteness), or it is not known whether the value exists (which may or may not contribute to completeness).

### 2.1.3. Time related dimensions

An important characteristic that defines data quality is their change over the time and to extent to which they are up to date. Most research recognizes three closely related time dimensions: currency, volatility and timeliness. Ballou et al. (1998) defined the three time-related dimensions and their relation. The *currency* of data concerns how often data is updated. It can be expressed by the time of the last update of a database or the time between receiving a data unit and the delivery of the data unit to a customer. *Volatility* is defined as the length of time data remains valid. Real-world objects that are subject to rapid change (for example wind speed) provide highly volatile data. *Timeliness* implies that data should not only be current, but the right data should be available before they are used. Ballou et al. (1998) defined a measure for timeliness, presenting the relation between the three time-related dimensions:

$$Timeliness = \max\{0, 1 - \frac{currency}{volatility}\} \tag{3.1}$$

### 2.1.4. Consistency

The consistency of data considers the violation of semantic rules (Batini & Scannapieco, 2006). These semantic rules are often expressed in so-called integrity constraints: properties that must be satisfied by all instances in a dataset. Batini et al. (2009) describes two fundamental categories of integrity constraints:

- Intra-relation constraints define a range of admissible values for an attribute. An example of a violation of such a constraint is a negative age in a database presenting persons (violating the integrity constraint that “age” must be a positive number).
- Inter-relation constraints involve attributes from other relational databases. An example of a violation of such a constraint is a different age of the same person (identified by a social security number) in two databases.

### 2.1.5. Other Dimensions

Even though the dimensions described above are recognized as key data quality dimensions and mentioned in most data quality research and methodologies, many other dimensions have been identified. Many papers aim to completely identify and describe all important characteristics and dimensions that define data quality. Generally, these proposals of sets and taxonomies of dimensions specify the data quality concept in a general setting (i.e. they apply to every context). Examples of well-known taxonomies and categorizations of data quality dimensions are given in (Cai & Zhu, 2015; Eppler, 2006; Kahn et al., 2002; Redman, 1996; Stvilia et al., 2007; Wand & Wang, 1996; Wang & Strong, 1996a) and described in (van Wierst, 2018). Data quality assessment methodologies often adopt one of these categorizations/taxonomies, creating a fixed set of dimensions. However, multiple papers suggest that the set of dimensions used in data quality assessment should be open, and that the selection of dimensions is part of the assessment process (De Amicis & Batini, 2004; Pipino et al., 2002; Su & Jin, 2004).

This way, an assessment method is developed that is customized to the data requirements in a specific context. However, an open set of dimensions always needs a reference set (from which dimensions are selected), for example the PSQ/IQ model described in Kahn et al. (2002). The most complete reference set is defined by Eppler (2006), who presents a list of seventy typical data quality dimensions (see Appendix I: 70 data quality dimensions provided by Eppler (2006). Eppler argues that during data quality assessment, this list should be shortened to twelve to eighteen criteria, as that amount provides an adequate scope of criteria (considering other assessment methodologies). However, he does not provide a method for selecting dimensions from his reference set.

### 2.1.6. Measurements for dimensions

Designing the right metrics is one of the most challenging tasks of data quality assessment, as they should identify all errors, without reflecting the same errors multiple times (del Pilar Angeles & García-Ugalde, 2009). An overview of data quality dimensions and their measures used throughout a variety of methodologies is presented by Batini et al. (2009) (see Appendix II: Collection of data quality dimensions and metrics from different methodologies (Batini et al., 2009). As can be seen in this overview, a user survey is included as a metric for each dimension, to assess the perceived quality of data users (i.e. subjective measures).

The most simple formula (referred to as the *simple ratio*) for obtaining the value of objective measures is by calculating a ratio like the following (Caballero et al., 2007; Y.W. Lee et al., 2006):

$$Ratio = 1 - [Number\ of\ undesirable\ outcomes / Total\ outcomes] \quad (3.2)$$

However, the calculation of such ratios is only possible when there are clear rules on when an outcome is desirable or undesirable. Besides the simple ratio, Pipino et al. (2002) describe two more functional forms for the definition of objective measures:

- *Min/Max Operation*: to handle dimensions that require the aggregation of two or more data quality indicators (e.g. the above described ratio's). The min operator is conservative as it assigns the lowest quality indicator to a dimension. An example of the max operator can be found in the formula for assessing timeliness by Ballou et al. (1998) (see formula 3.1)
- *Weighted average*: in which weights are assigned to metrics in order to calculate a score for a dimension. A typical formula looks as follows (Y.W. Lee et al., 2006):

$$DQ = \sum ni = 1(aiMi) \quad (3.3)$$

  In which $n$ is the amount of individual metrics, $ai$ is a weighting factor of measure $i$ with $0 \leq ai \leq 1$, $a1 + a2 + [\dots] + an = 1$ and $Mi$ is a normalized value of the assessment of the $i$-th metric.

The definition of data quality as 'fitness for use' implies that the quality of data is highly determined by the perceived quality of data by data consumers (i.e. those who use the data). However, most methodologies provide only objective measures for assessing data quality dimensions. Pipino et al. (2002) recognize the importance of the distinction between subjective and objective measures, and argues that a comparison between the two, is the input for the identification of data quality problems. The differences between objective and subjective can be found in Table 2.1. Pipino and his colleagues conclude that subjective measures are an important part of data quality assessment.

| Feature | Objective | Subjective |
|---|---|---|
| Measurement tool | Software | Survey |
| Measuring target | Datum | Representational information |
| Measuring Standard | Rules, Patterns | User Satisfaction |
| Process | Automated | User Involved |
| Result | Single | Multiple |
| Data Storage | Databases | Business Contexts |

*Table 2.1: Objective versus subjective measures, adapted from Pipino et al. (2002)*

## 2.2. Related work

This section describes research works that are closely related to the research goals and methods of this study.

### 2.2.1. Configuring a data quality assessment process

One objective of this study is to provide a data quality assessment process that conforms to the requirements that an organization may have for this assessment (i.e. its fits organizational needs and the current situation). This goal has been pursued by other researchers as well, for example Woodall et al. (2013). In their paper, they propose a configuration method that dynamically configures the data quality assessment process for specific business needs, while leveraging the best practices from existing methodologies. The input for this configuration method is a generic data quality assessment process containing recommended activities (critical activities that should always be included in data quality assessment) and optional activities (activities that can optionally be performed based on the requirements of the data quality assessment), and the dependencies between them (see Figure 2.1).

This generic assessment process was obtained by extracting and grouping activities and their definitions from a selected number of data quality assessment methodologies. Based on the inclusion of activities across different methodologies, the activities were categorized as either recommended or optional. The order and dependencies between activities were defined based on the activity definitions and their inputs and outputs. Considering this generic data quality assessment process, the configuration method that Woodall et al. describe consists of:

- Determining the aim of the assessment and the company requirements related to the assessment. The aim of the assessment is essential to inform data quality assessors of what the resulting assessment process should be used for. The company requirements related to the assessment follows from the determined aim of the assessment.
- Select the activities from the generic process model that contribute to the assessment aim and that meet company requirements.
- Configure the activities in the process: arrange the activities into a sensible order and include any activity dependencies.

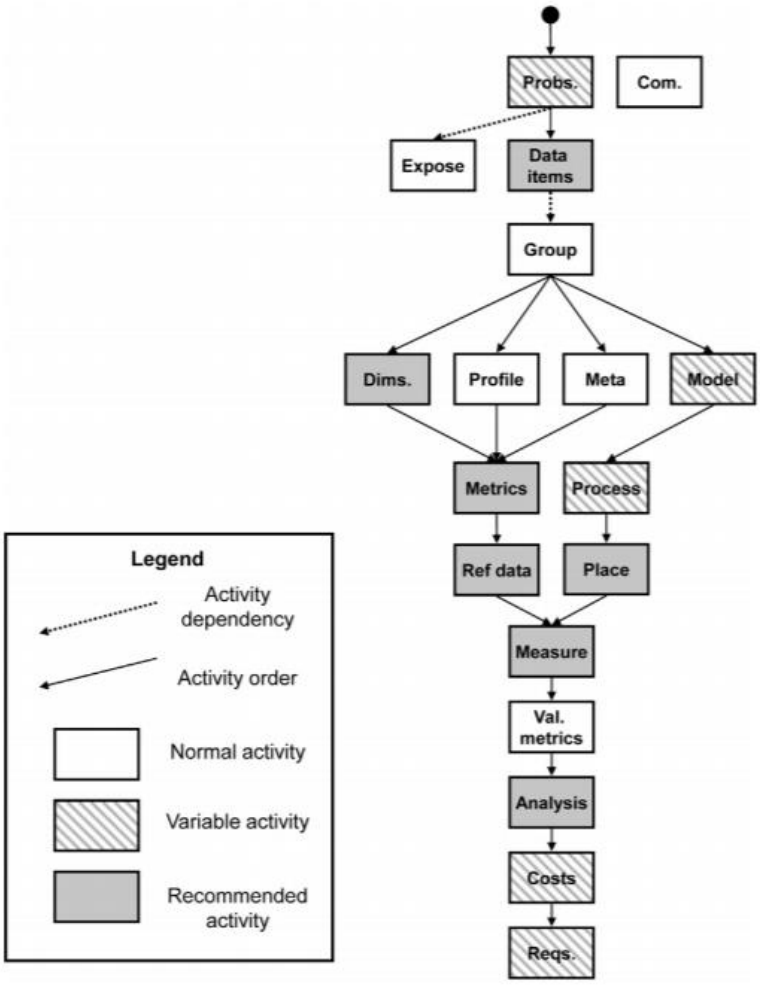



*Figure 2.1: A generic data quality assessment process (Woodall et al., 2013)*

### 2.2.2. Comparative analysis of data quality assessment methodologies

The process model designed in this study is based on a comparative analysis of existing methodologies in order to identify critical activities. A similar but more extensive comparative analysis has been done by Batini et al. (2009). In their paper they compare 13 data quality methodologies (both assessment and improvement methodologies) based on the following:

- The methodological phases and steps
- The strategies and techniques
- The data quality dimensions and metrics considered
- The types of data considered (structured/unstructured/semi structured)
- The types of information systems

The comparison of methodologies on their phases and steps is of most interest for this study. Batini et al. found that all methodologies organize the data quality assessment process in several steps, and the following common steps can be recognized (see also Figure 2.2 for the inclusion of these steps among methodologies):

- *Data analysis*: in which schemas are examined and interviews are performed to achieve a complete understanding of data and related architectural and management rules.
- *Data quality requirement analysis*: in which surveys are conducted to find the opinion of data users and administrators to identify data quality issues and set new quality targets.
- *Identification of critical areas*: in which the most relevant databases and data flows to be quantitatively assessed are selected.
- *Process modeling*: which provides a model of the processes producing or updating data.
- *Measurement of quality*: in which quality dimensions are selected that are affected by the data quality issues identified in the requirement analysis, and metrics for these dimensions are defined.

| Step/Meth Acronym | Data Analysis | DQ Requirement Analysis | Identification of Critical Areas | Process Modeling | Measurement of Quality | Extensible to Other Dimensions and Metrics |
|---|---|---|---|---|---|---|
| TDQM | + | | + | + | + | Fixed |
| DWQ | + | + | + | | + | Open |
| TIQM | + | + | + | + | + | Fixed |
| AIMQ | + | | + | | + | Fixed |
| CIHI | + | | + | | | Fixed |
| DQA | + | | + | | + | Open |
| IQM | + | | | | + | Open |
| ISTAT | + | | | | + | Fixed |
| AMEQ | + | | + | + | + | Open |
| COLDQ | + | + | + | + | + | Fixed |
| DaQuinCIS | + | | + | + | + | Open |
| QAFD | + | + | + | | + | Fixed |
| CDQ | + | + | + | + | + | Open |

*Figure 2.2: Comparing methodologies on their (assessment) steps included (Batini et al., 2009)*

Furthermore, Batini et al. describe an optional activity prior to the assessment called state reconstruction. If not yet available, the state reconstruction collects contextual information on organizational processes, quality issues and corresponding costs.

### 2.2.3. Value of additional research

Although the work of Woodall et al. (2013) is valuable for organizations to configure the process of data quality assessment based on organizational needs, it does not provide practical guidelines on how to perform each of the activities. Obtaining a practical interpretation of the activities described in the work of Woodall and his colleagues would be valuable in combination with a configuration guide. In order to find the critical activities of data quality assessment, this research uses a similar approach to the ones in the works of both Woodall et al. (2013) and Batini et al. (2009).

# 3. Research Method

As the goal of this study is to develop an artifact (i.e. a data quality assessment process), a design science approach is chosen for the development of a research method. Peffers et al. (2007) describe a methodology for conducting design science research in the field of information systems. They argue that design science is of importance in any discipline for the creation of successful artefacts, but recognized that little design science had been done in the discipline of information systems. The lack of a commonly accepted framework for design science research within the discipline may have contributed to this slow adoption (Peffers et al., 2007). In their paper, they provide such a framework. This framework incorporates principles, practices and procedures to carry out design science research for information systems research. The research method of this study will follow their methodology. It includes six steps, presented in Figure 3.1. This chapter provides the application of these steps for this research and a justification of the research techniques used in each step.

## 3.1. Problem identification and motivation

The problem identification and motivation for the study defines the specific research problem and justifies the value of a solution. This problem definition provides a motivation for the development of an artefact (a process model in this study) that can effectively provide a solution. Besides clearly defining the specific problem, it is important to provide a justification of the value of a solution. This justification ensures that the researcher and the audience are motivated to pursue the solution, and it helps to understand the reasoning of the researcher associated with the problem as well as the need for a solution. An extensive narrative literature review prior to this research has been conducted to describe and discuss the current state of research on organizational data quality assessment and improvement (van Wierst, 2018). Based on this literature review the following research problem can be identified:

The majority of data quality frameworks and methodologies are either developed for a specific context, technique or problem, or they provide a generic assessment method (i.e. regardless of context or application) that often lacks practical guidance and is not operationalized for a specific context and business needs. A generic but practical model for data quality assessment, that incorporates the context in which the assessment is conducted, is missing.

A solution to this problem in the form of a process model is valuable for data quality practitioners as it enables them to effectively and efficiently obtain a complete assessment of their data quality. Also, such a model ensures that this assessment is suitable for the context in which it is performed, by providing a method for selecting relevant dimensions for this context.

## 3.2. Definition of the objectives for a solution

The objectives for a solution are derived from the problem definition. Table 3.1 presents the identified objectives for this study, based on the problem definition. The table provides a reasoning for the inclusion of the objective and describes the relation to the research problem.

| Objective | Reasoning | Relation to research problem |
|---|---|---|
| **Practical utility** | Any vagueness on how to conduct the activities in the designed process model must be eliminated | Existing generic data quality assessment methodologies often lack practical guidance. |
| **Comprehensiveness** | A generic process model should be comprehensive to be applicable independent of context. | A generic process is often not practical for specific contexts. |
| **Genericness** | The designed process model must be applicable independent of any context | A generic but practical model for data quality assessment is missing |
| **Understandability** | The process model must be presented in an understandable format | For a process model to be practical, it must be well understandable |
| **Completeness** | The final assessment must give a complete overview of the current state of data quality in a specific context | Existing methodologies often do not fit specific business needs, and may therefore give incomplete or irrelevant results |

*Table 3.1: Solution objectives*

Practical utility refers to what degree the process model and the activities and roles that compose it are perceived as practical, and not abstract or high-level. This means that the activities in the model need to be defined on a low-level such that the activities and tasks are not interpretable in more than one way, and that any vagueness of the definitions, goals or description of activities is eliminated. Comprehensiveness of the process model ensures that all critical activities of data quality assessment are included. Dependent of the context in which data quality is assessed, some activities can be of more importance than others. Therefore, in a generic model, all activities that have the potential to be critical in a context, need to be included. Also, a comprehensive model includes both a top-down and bottom-up approach (as depicted in Figure 1.2). The genericness of the process model refers to what degree the model is applicable independent of context. This means that all activities and roles defined must make sense independent of context. The understandability objective refers to what degree the model is presented in an understandable format. This includes that graphical depictions of the model are clear and conform to general modeling rules, and activities are clearly described in an understandable way. Finally, the completeness of the model refers to how the final result of the process model is perceived as a complete assessment of the current state of data quality, thus that it represents all data quality goals and problems for a specific context.

## 3.3. Design and development

After clearly defining the problem and the objectives that a solution must satisfy, the next step is to create the artefact; for this study that is the development of a data quality process model. Peffers et al. (2007) describe that moving from objectives to design and development requires knowledge of theory to bear in a solution.

Before creating the actual process model, the following knowledge needs to be obtained: in order for the process model to be comprehensive, all critical activities of a generic data quality assessment process must be identified. Furthermore, for the process model to be practical in its use, a clear definition of the roles that participate in the process and in what activities they are involved is required.

Considering the solution objectives and the above described knowledge requirements, two (possibly overlapping) categories of objectives can be identified; on the one hand, there are objectives that reflect

design goals of the artefact, thus they are a result of an adequate design of the process (they should be constantly kept in mind during the actual creation of the artefact). On the other hand, there are objectives that require specific knowledge or theory to be satisfied, which needs to be obtained before the actual design of the process. Table 3.2 presents for each objective the corresponding category, and the required knowledge or goal to achieve each objective.

| Objective | Category | Required knowledge/ Design goal |
|---|---|---|
| **Practical utility** | Both | Identification of roles to be assigned in a data quality assessment process, activities in the process must be defined on a low level |
| **Comprehension** | Knowledge requirement | Identification of critical activities of a generic data quality assessment process. Inclusion of different data quality assessment approaches. |
| **Genericness** | Design goal | All activities in the process model need to interpretable independent of any context |
| **Understandability** | Design goal | The process must be presented in a clear presentation and conform to common modeling rules |
| **Completeness** | Both | The model must combine different perspectives of data quality in a final result |

*Table 3.2: Knowledge requirements and design goals for the solution objectives*

In order to obtain this required knowledge, a literature review is conducted, following by a synthesis of this literature. Based on the identified knowledge requirements, the following questions need to be answered by this literature review and synthesis:

- What are the critical activities in a generic data quality assessment process?
- What roles need to be assigned to these activities to effectively perform the data quality assessment process?

During this literature review, relevant existing data quality assessment methodologies (on its own or as a part of a bigger data management approach) are collected and analyzed on both the activities that they contain and, if any, the roles that they define. In the synthesis, the aim is to group both activities and roles across the different methodologies based on their similarity. This grouping is direct input for the identification of critical activities and roles.

After synthesizing the literature, the actual process model is designed considering the critical activities and roles identified in the synthesis. During this design, the earlier defined solution objectives, that represent design goals, are taken into account. BPMN is chosen is used as the modeling language for the process model, as BPMN is activity based and allows for visually depicting both information flows and roles.

## 3.4. Demonstration

Following Peffers et al. (2007) methodology, the next step is to demonstrate the use of the artifact. As this research aims to provide a solution for practicing data quality assessment in the field, its demonstration should be in the field as well. Therefore, a case study is the chosen method to demonstrate the use of the process model. Considering the different types of case-studies described by (Yin, 2003), for this research, an holistic single case study is applied. This means that the model will be applied for a single

case using one unit of analysis. The rationale behind is the following: a single case allows for revelation: the opportunity to observe and analyze the use of the process model in depth. As the study will be validated based on the opinion and experiences of individual participants of the case, a single unit of analysis is deployed, namely the individuals. This case study will be conducted at the EUV factory of ASML. More information on this case can be found in Chapter 6.

## 3.5. Evaluation

The goal of the evaluation is to measure how well the designed artefact supports a solution to the problem. To measure this, the previously defined solution objectives are to be evaluated based on the observations and results during the demonstration. Based on this evaluation, the research either iterates back to the design step to improve the effectiveness, or it leaves potential improvements to subsequent research or projects. Since the defined solution objectives for this research mainly reflect qualitative characteristic (i.e. they are determined by the experiences and opinions of participants of the process model), a qualitative evaluation is deployed.

This qualitative evaluation is achieved by performing semi-structured interviews with participants of the process in the case study. Semi-structured are chosen for this evaluation as they allow for obtaining comprehensive experiences and opinions regarding the use of the process model for each of the solution objective. For each solution objective, several standard questions (that will be asked to all participants) are defined (see Table 3.3). Based on the given answers, in-depth questions may be asked to obtain a good understanding of experiences and opinions.

| Objective | Interview Questions |
|---|---|
| **Practical Utility** | - Do you think that the proposed process model is practical? |
| | - Do you think activities and roles are defined on a low-level and are not abstract? |
| | - Have you experienced any vagueness in the definition or description of activities or roles? |
| **Comprehensiveness** | - Do you think that the process model includes all critical activities of data quality assessment? |
| | - Do you think there are critical activities missing in this model? |
| | - Do you think there are roles missing in this model? |
| | - Do you think that the model approaches data quality from a broad perspective? |
| **Genericness** | - Do you think this process model can be easily applied in other contexts? |
| | - Do you feel like every activity is defined independent of this context? |
| | - Do you feel like every role is defined independent of this context? |
| **Understandability** | - Do you think that the process model is clearly depicted? |
| | - Do you think the process model conforms to BPMN rules? |
| **Completeness** | - Do you feel like the final assessment gives a complete overview of the current state of data quality? |
| | - Do you feel like there are other data quality problems or goals that are not represented in this assessment? |

*Table 3.3: Evaluation interview questions*

## 3.6. Communication

The sixth activity described by Peffers et al. is communication. This involves presenting the problem and its importance, and the artefact with its novelty and effectiveness to the relevant audiences and practicing

professionals. There are two main groups of relevant audience for this study. On the hand, the results of this study are of value for data quality practitioners in the field, as it supports them in obtaining a complete and effective data quality assessment. On the other hand, the results of this study provide input for data quality researchers, as it provides future research directions for further evaluation and improvement of the model. This report is the main means of communication of this research and will be included in the research repository of the University of Technology Eindhoven, where it is available for the public.

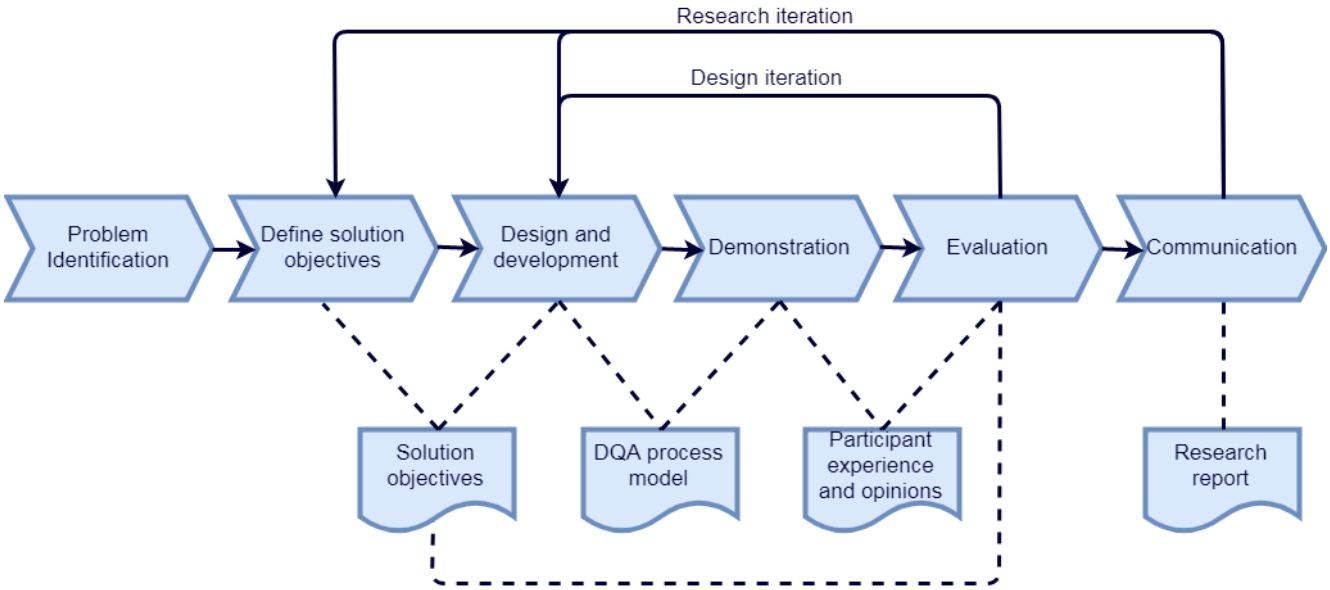



*Figure 3.1: Research roadmap*

# 4. Analysis of existing methodologies

To obtain the required knowledge for the design of the process model, a literature review is conducted. As is argued in chapter 4, for the process model to meet the solution objectives, certain specific knowledge is required; in order for the process model to be complete, all critical activities, regardless of any context, of a data quality assessment process must be included. In addition, for the process model to be practical (i.e. with specific guidelines on how to perform activities) it requires the identification of the roles that participate in the process and their involvement in each of the activities. This leads to the following research questions that are to be answered by this literature review:

- What are the critical activities in a generic data quality assessment process?
- What roles need to be assigned to these activities to effectively perform the data quality assessment process?

Figure 4.1 shows the process of this literature review. This process is adapted from the literature review process provided by Budgen & Brereton (2006). Although this paper describes a systematic literature review, the literature review in this study is designed more flexible to allow the inclusion of papers based on a subjective assessment by the researcher. The process is as follows: based on the research questions, a search strategy is determined. This search strategy consists of the definition of search terms and considered databases. The research questions also provide input for the definition of the inclusion criteria (described in 4.2). Subsequently, the chosen databases are searched, and relevant articles are collected based on these inclusion criteria (and keeping in mind the research questions that need to be answered). Each article is then analyzed based on the activities and roles that they define. Finally, the results are synthesized: similar activities across the different methodologies presented in the papers are grouped together. This grouping on similarity is based on a subjective judgement of the researcher (e.g. considering their inputs, outputs, goals and techniques). Part of the synthesis is to assign (and justify this assignment) the identified roles to the identified critical activities. Finally, based on this synthesis conclusions are drawn in which the research questions are answered.

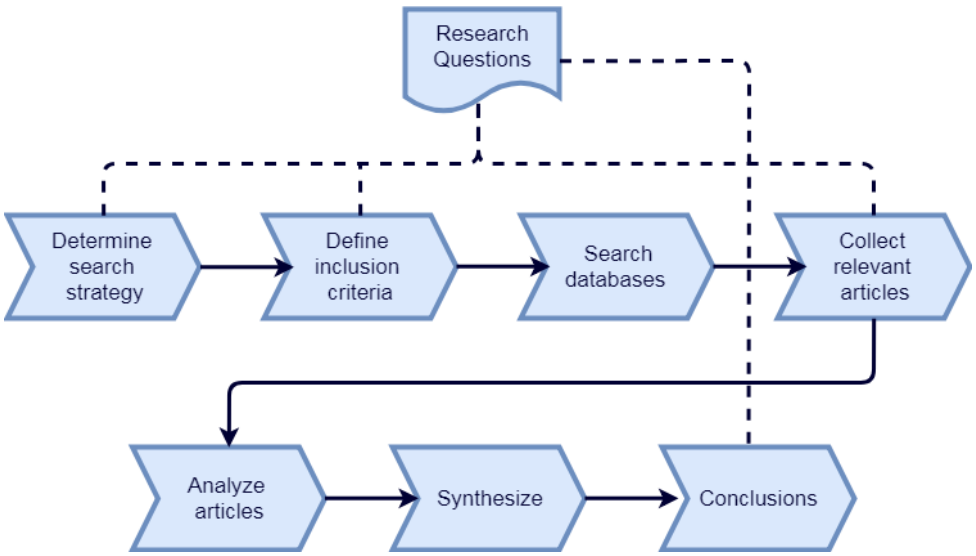



*Figure 4.1: Answering the research questions*

## 4.1. Search strategy

The LibrarySearch tool provided by the University of Technology Eindhoven is used to execute search queries. This tool executes the queries over 42 online databases (see Appendix III: Databases searched for literature review). A set of search words is defined based on the context of this research and the research questions. Based on the amount of results per search query and a quick judgement of the relevancy of these results, search terms are added, refined and combined (using Boolean operators) to filter out irrelevant results. The relevancy of the results is assessed based on their title and abstract or description. If a result is found relevant, a decision is made for inclusion in this review by reviewing the work and applying the inclusion criteria described in section 5.2. Other than finding research directly from the databases, contributions are found by checking relevant references to other work as well (for example provide Batini et al. (2009) many relevant references). Appendix IV: Search words used for literature review shows the final set of search words used.

## 4.2. Inclusion criteria

To decide whether the research contributions provide valuable input for this literature review, and for the questions that need to be answered, the following inclusion criteria are applied to assess the article. These criteria are subjectively assessed by the researcher.

- The work must present a methodology or a process for data quality assessment. This can either be focused on data quality assessment specifically, or as a part of a larger data quality approach.
- The work goes into detail on the assessment phase (i.e. it does not primarily focus on data quality improvement or other data quality management activities).
- The methodology or process presented must be applicable to other contexts. This does not mean that only generic methodologies are considered, but they cannot be too focused on specific situations, problems, or data (for example in Ahmed, 2018; Madhikermi et al., 2016; Shardt & Huang, 2013). The steps, activities and goals should make sense in other contexts as well.
- The methodology or process presented must be validated, either through experimentation or through appliance in case study. This ensures that it has some proven value for data quality assessment practices.

## 4.3. Included research for analysis

This section summarizes the researches that are found in the literature search and that meet the inclusion criteria. In total, eight methodologies are included (see Table 4.1). Each methodology is shortly described on their approach, goals and unique elements. Also, a graphical representation (see Figure 4.2 for a legend) of the activities, inputs and outputs and roles (if mentioned) is given for each methodology. Considering the focus of this research (see section 1.2), a clear distinction is made between data quality assessment activities and activities that are part of other data quality management competences. The latter are not included in the analysis and overviews.

| Methodology | Acronym | Reference |
|---|---|---|
| Total Data Quality Management | TDQM | Wang, 1998 |
| Data Quality Assessment | DQA | Pipino et al., 2002 |
| A Data Quality Assessment Framework | DQAF | Sebastian-Coleman, 2013 |
| Data Quality Assessment: The Hybrid Approach | Hybrid | Woodall et al., 2013 |
| A Methodology for Information Quality Assessment | AIMQ | Lee et al., 2002 |
| Framework and Methodology for Data Quality Assessment | ORME-DQ | Batini et al., 2007 |
| Data Warehouse Quality Methodology | DWQ | Jeusfeld et al., 1998 |
| Data Quality Assessment for Life Cycle Assessment | DQALCA | Bicalho et al., 2017 |

*Table 4.1: Included methodologies for analysis*

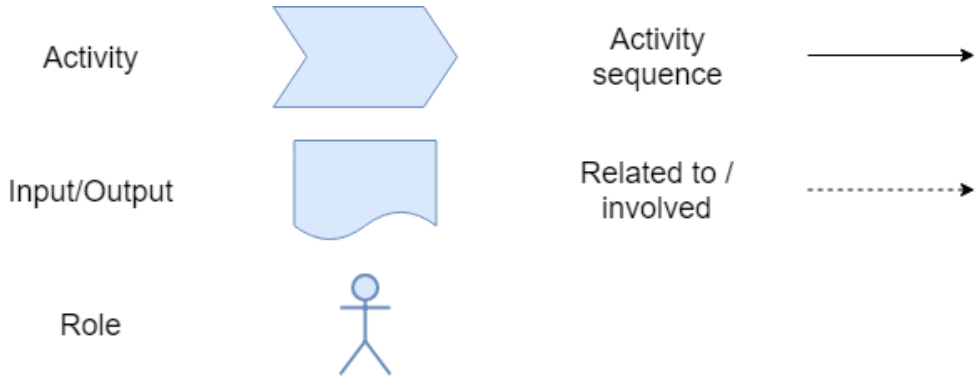



*Figure 4.2: legend for graphical presentation of methodologies*

### 4.3.1. Total Data Quality Management (TDQM)

The Total Data Quality Management (TDQM) methodology (Wang, 1998) was the first general methodology proposed in data quality literature. It was based on academic research, and its fundamental objective is to extend the principles of Total Quality Management (TQM) (Oakland, 1989) to data quality: like raw materials are needed for the manufacturing of product, raw data is needed in the manufacturing of information. Likewise, like the process in product manufacturing consists of an assembly line, the process in information manufacturing flows through information systems. Finally, as the output of product manufacturing is a physical product, the output of information manufacturing is an information product (IP). A schema of the TDQM methodology is shown in Figure 4.3. Considering that the focus of this research is on data quality assessment, only the definition and measurement phase of TDQM are considered in this paper. The first step is to define the characteristics of the information product. This is done on two levels: at the higher-level, the functionalities for the information consumers are defined (what functionalities are needed to perform the task at hand). On a lower level, the basic units of the IP and their relationships are defined and presented in, for example, an entity-relationship model. Then, based on the perspectives of different roles (TDQM differentiates between IP suppliers, manufacturers,

consumers and managers), the IP requirements are defined using surveys and dimensions for the assessment are chosen. Finally, the information manufacturing system is defined, that describes how the IP is produced. After defining the IP characteristics, requirements and manufacturing system, metrics (subjective and objective) are defined for the chosen dimensions. TDQM differentiates between basic data quality measures defined in the literature, and specific measures based on business rules. Using the data quality metrics, data quality measures can be obtained along various data quality dimensions for analysis. Low scoring metrics and dimensions are direct input for identifying data quality problems. This process is presented in Figure 4.4.

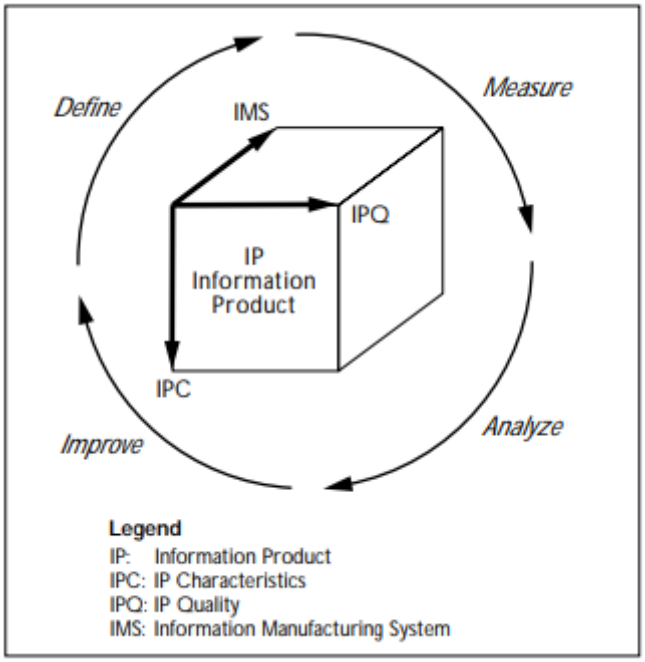



*Figure 4.3: Total Data Quality Management (Wang, 1998)*

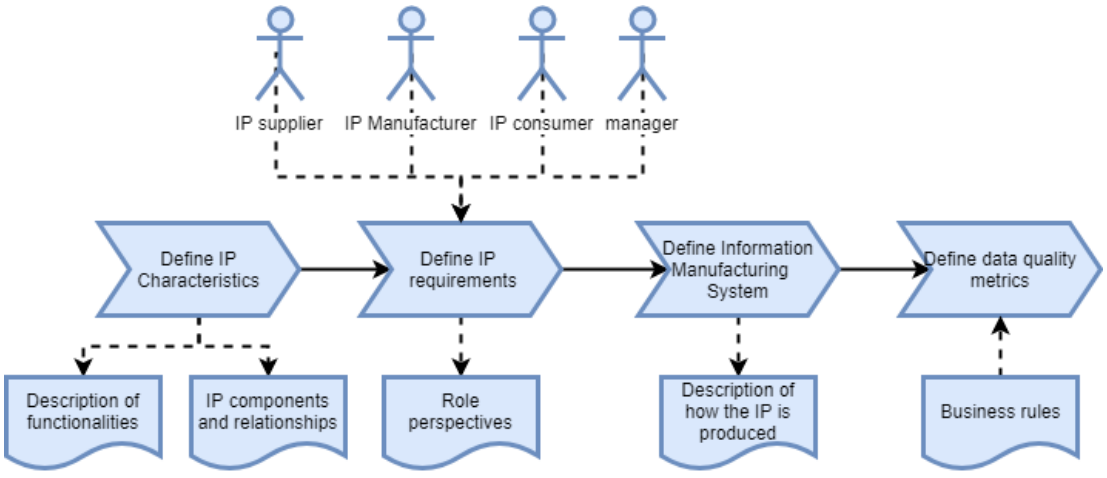



*Figure 4.4: TDQM process*

### 4.3.2. Data Quality Assessment (DQA)

Pipino et al. (2002) argue that data quality assessment requires awareness of the "fundamental principles underlying the development of subjective and objective data quality metrics". In their paper, they present a methodology in which the comparison between subjective and objective measures is the foundation for identifying improvement directions. Data quality is subjectively assessed using a questionnaire among different roles (data consumers, data custodians, data providers and managers). This assessment obtains a quality score (1 to 10) for each of the dimension assessed (a fixed set of dimensions is proposed in the paper, but the method is extendable to other dimensions as well). Also, the data is objectively assessed using objective quality metrics, for which the paper presents three functional forms (see section 2.1.6) to create them. A comparative analysis between the subjective assessment and the objective assessment (using the matrix in Figure 4.5: the quadrants I, II, and III indicate a data quality problem that needs improvement) finds discrepancies and is the input for the identification of improvements. This process is presented in Figure 4.6.

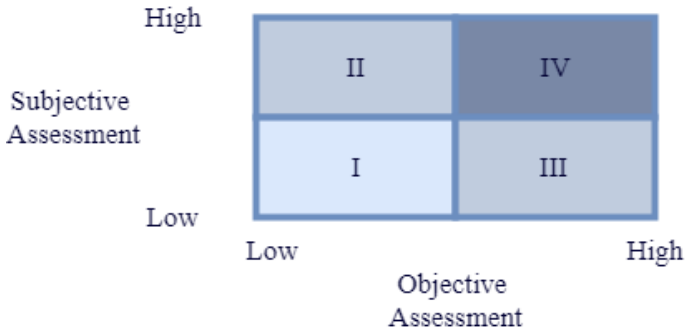



*Figure 4.5: Comparing subjective and objective measurement (Pipino et al., 2002)*

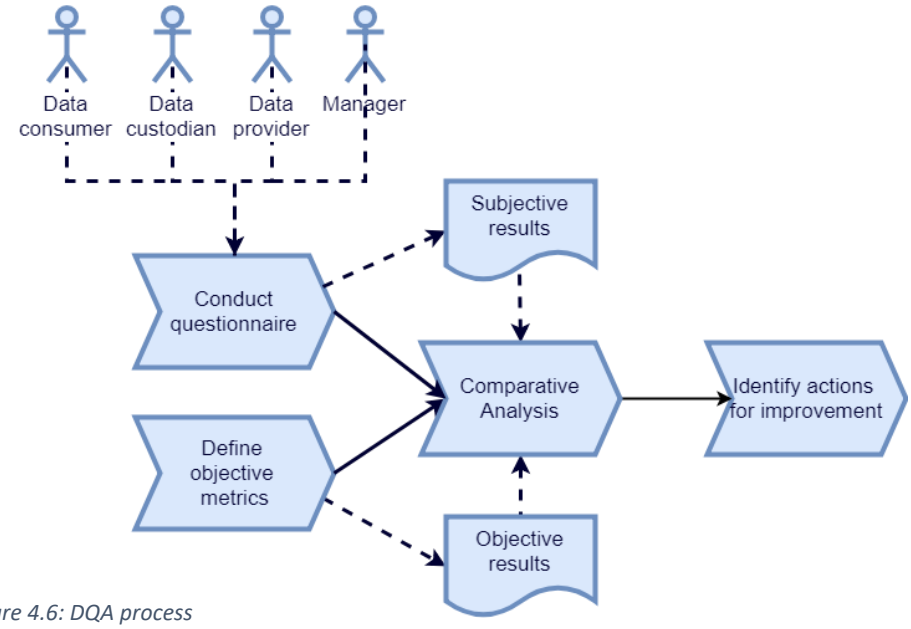



*Figure 4.6: DQA process*

### 4.3.3. A Data Quality Assessment Framework (DQAF)

In her book, Sebastian-coleman (2013) teaches how to measure and monitor data quality over time. The author defines four different assessment scenarios, all having different goals and deliverables: an initial assessment identifies a measure baseline and identifies the data to be measured on an ongoing basis. Data quality assessment in improvement projects aim to show the improvement in data quality as process changes are implemented. Lastly, in-line measurements and periodic measurements ensure that data continues to meet expectations. Since the latter three are not considered in the scope of this research (they focus on other disciplines of the data quality management model described in section 1.1), the initial assessment scenario is analyzed here. This assessment starts with data profiling: identifying and reporting the data structure, content, rules and relationships by applying statistical methodologies to return a set of standard characteristics about data (data types, field lengths, cardinality of columns, granularity, value sets, format patterns, implied rules, and cross-column and cross-file data relationships, as well as the cardinality of these relationships). Data profiling consists of both column profiling (identifying characteristics of individual columns) and structure profiling (identifying the relationships between columns or between tables and the rules that govern those relationships). Based on this data profiling, expectations from both data users and data producers are defined (for example: if a record representing a person has marital status "married", then it is expected that the column "spouse" contains a name). The expectations are compared to the actual measures (from example: only 80% percent of records with marital status "married" have a name in the column "spouse"), and from this comparison, improvement directions are identified. This process is presented in Figure 4.7.

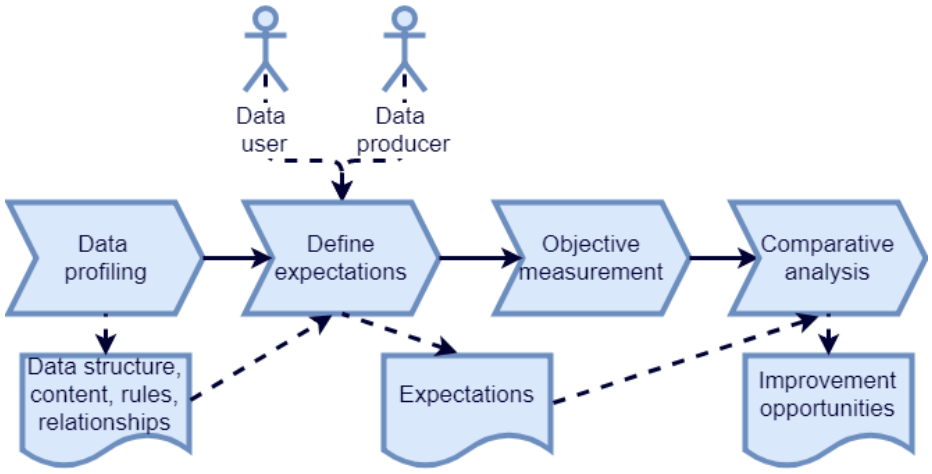



*Figure 4.7: DQAF process*

### 4.3.4. Data Quality assessment: The Hybrid approach (Hybrid)

Woodall et al. (2013) argue that organizations have different requirements for data quality assessment but that there are no methods to configure existing data quality assessment methods to organizational needs. In their paper, they propose an approach to dynamically configure an assessment technique while leveraging the best practices from existing assessment techniques. Based on a literature review, they classify data quality assessment activities as recommended or optional and create a generic assessment

process containing both these recommended and optional activities. The first step of their approach is to determine the aim of the assessment for example, to determine and prioritize an organization's data quality problems and obtain measurements for each problem). Then, the company requirements related to the assessment are identified (for example: determine the costs caused by low data quality and model the way data is created and how it flows). Finally, activities are selected, and their order and dependencies are defined. Although the paper does not provide a practical application of the activities to be performed, the results of their literature review and the recommended activities that they have identified are valuable input for this research. These recommended activities are shown in Figure 4.8

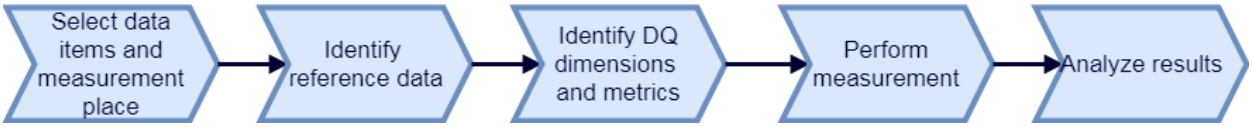



*Figure 4.8: Hybrid process*

### 4.3.5. A Methodology for Information Quality Assessment (AIMQ)

The AIMQ (A Methodology for Information Quality Assessment) Methodology was developed by Yang W Lee et al. (2002), and consists of three main components: The PSP/IQ model (Kahn et al. (2002) see Table 4.2) organizes the key data quality dimensions in four dimensions so that meaningful decisions can be made about improving data quality (a first pilot questionnaire is used to identify relevant quality dimensions and attributes). The IQA instrument measures data quality for each of the data quality dimensions (dimensions from the same quadrant are averaged to obtain a measurement for each quadrant). The IQA instrument is a questionnaire that is conducted among information consumers and IS professionals in different organizational roles. Finally, based on the questionnaire results, gap analysis techniques are applied. Benchmarking is used to compare the results of the questionnaire to the results of competitors, industry leaders and other sources of best practices. A role gap analysis compares the questionnaire results from respondents in different organizational roles, IS professionals and information consumers. The role gap analysis aims to explain whether differences between roles can cause different assessments of data quality. This comparison across roles serves to identify data quality problems and lays the foundation for data quality improvement. The AIMQ process is presented in Figure 4.9.

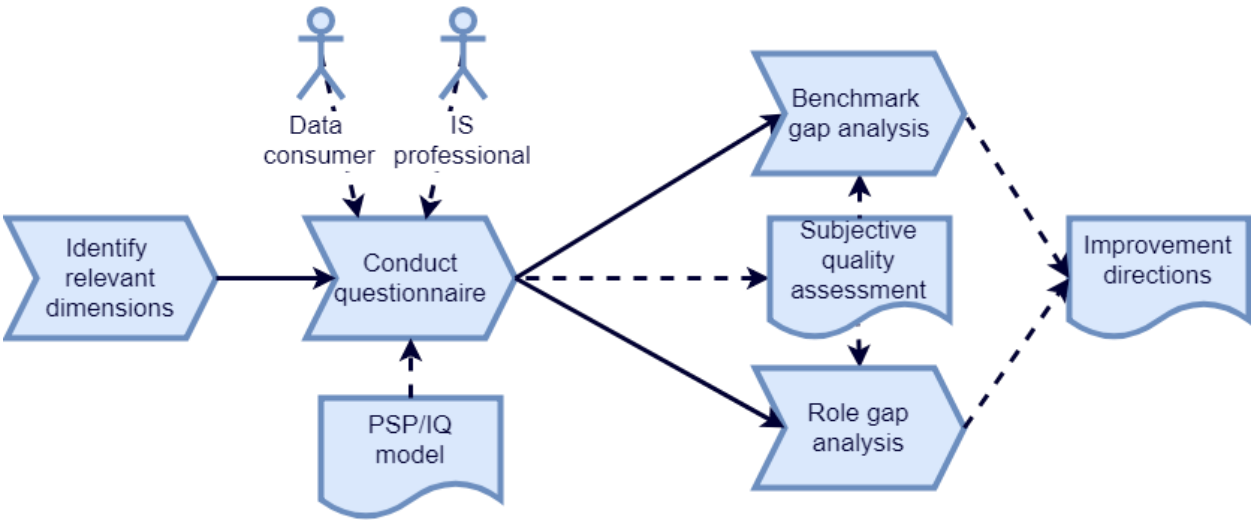



*Figure 4.9: AIMQ process*

| | Conforms to specifications | Meets or exceeds consumer expectations |
|---|---|---|
| **Product Quality** | Sound information | Useful information |
| **Service Quality** | Dependable information | Usable information |

*Table 4.2: PSP/IQ model (Kahn et al., 2002)*

### 4.3.6. Framework and Methodology for Data Quality Assessment (ORME-DQ)

Batini et al. (2007) propose a data quality assessment methodology (ORME-DQ) that is based on applying the relevant principles of a well-known approach for operational risk evaluation to information and data quality and its effects on operational risk. The first step of this methodology is to develop a state reconstruction to identify all relationships between organizational units, process, services and data. This step aims to provide a clear picture of the main uses of data, of providers, and of consumers of data flows. After the state reconstruction, a loss analysis is performed. This loss analysis identifies loss events caused by low data quality and provides an economic value of the expected loss (using a predefined hierarchy of costs caused by low data quality, and appropriate metrics). Given the loss events with the largest economic impact, the critical business processes related to these loss events are selected and the datasets provided or consumed by these processes are identified. Lastly, the relevant datasets are assessed by selecting quality metrics from existing literature (by a data quality expert). Using these measurements, further analysis is done on the conditional probability of loss events and their relation to historical series of data quality dimensions quantitative measures. This process is presented in Figure 4.10.

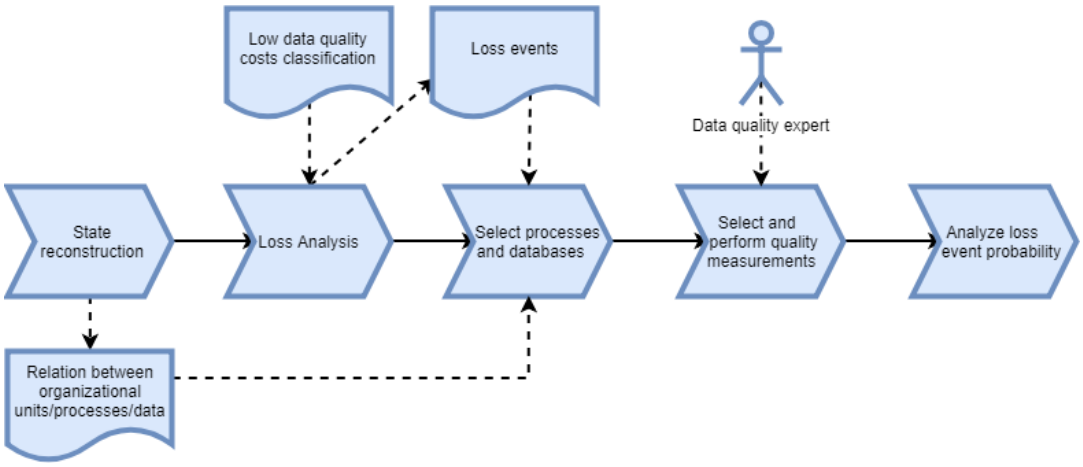



*Figure 4.10: ORME-DQ process*

### 4.3.7. Data Warehouse Quality Methodology (DWQ)

The Data Warehouse Quality (DWQ) methodology (Jeusfeld et al., 1998) studies the relationship between quality objectives and design options in data warehousing. Jeusfeld et al. (1998) propose a model in which the components of a data warehouse are linked to a quality model as presented in Figure 4.11, and show how this model can be used for quality goal formulation and quality assessment. Their proposed model allows distinctive stakeholder groups to design abstract quality goals (for example: “increase the efficiency of the data loading process”) that are translated into executable analysis queries on quality measurements in the data warehouse’s meta database. Based on these quality goals, the methodology allows for a free selection (and definition) of quality dimensions by different stakeholders. First, abstract quality goals are obtained from different stakeholders. Based on these quality goals and the data warehouse context (which is not considered for this analysis), relevant dimensions of data quality are identified. Stakeholders identify weights to these dimensions based on their importance. The obtained data quality goals are translated into executable queries that can run over a database’s metadata (to retrieve timestamps for example). Finally, the obtained results are compared to the previously defined quality goals to identify directions for improvement. This process is shown in Figure 4.12.

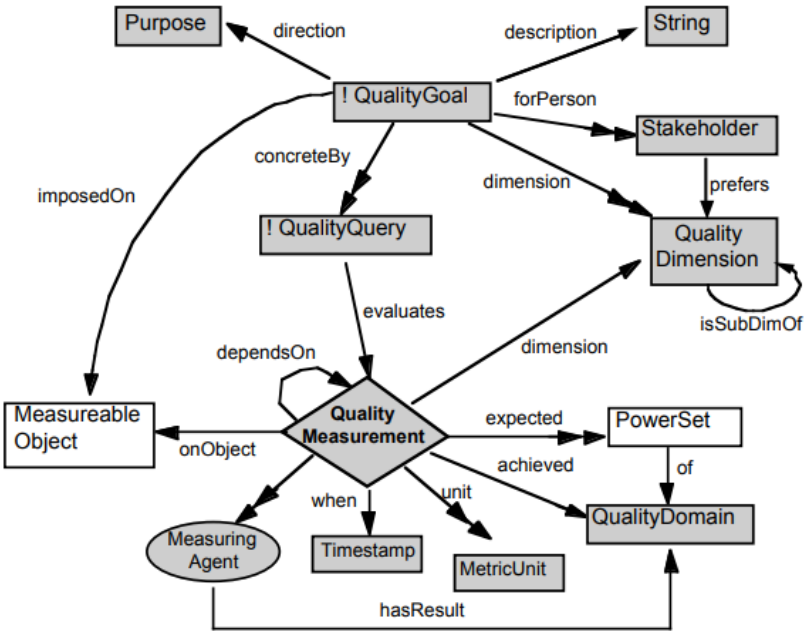



*Figure 4.11: Data quality concept model (Jeusfeld et al., 1998)*

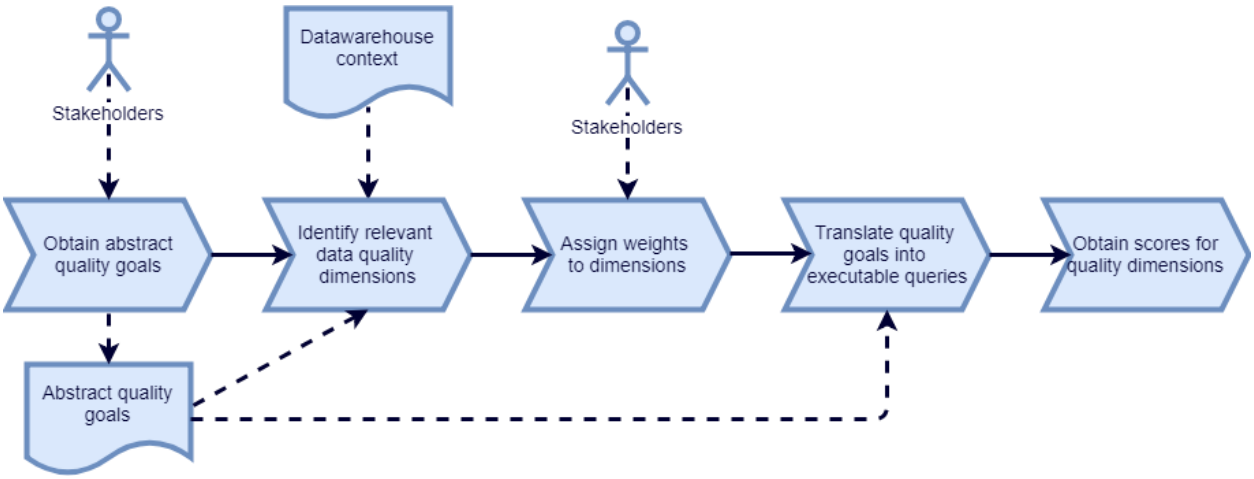



*Figure 4.12: DWQ process*

### 4.3.8. Data quality assessment for Life Cycle Assessment (DQALCA)

The aim of the paper of Bicalho et al. (2017) is to investigate the adequacy of the current approach for data quality assessment for Life Cycle Assessment (LCA). Although this paper focusses on a specific problem (LCA data) and it aims to identify problems in the current way of assessment, the methodology that it presents is valuable for this research. The process of assessing data quality starts by identifying the data quality goals. These quality goals are specific for a LCA. They are defined by the users of this data, as data quality depends on what users expect form it (an example from the paper: use representative data of an oil palm production located in Para, Brazil that applies modern farm techniques). Based on the goals of the LCA, the required data is selected (determine what data is needed) and collected (find the sources where to find this needed data). Thereafter, the data is assessed using the pedigree matrix (proposed by

Weidema & Wesnæs (1996)) and known as the main reference for data quality assessment in LCA). This matrix assesses data quality using five predefined dimensions by giving scores (1 to 5) to each dimension, based on descriptive quality indicators. The assignment of scores to these dimensions is based on physical measurements and expert judgements. Some dimensions (temporal, geographical and further technical correlations) are dependent on the defined quality goals and are therefore subjectively assessed considering these quality goals. An overview of this assessment process can be found in Figure 4.13.

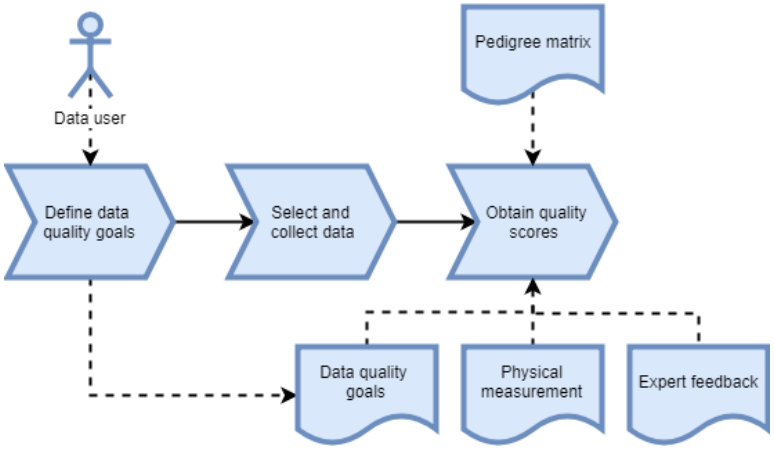



*Figure 4.13: DQALCA process*

## 4.4. Synthesis

This section synthesizes the analyzed methodologies in order to answer the research questions.

### 4.4.1. Identifying critical activities of data quality assessment

First, for all activities that are identified in the analysis of the selected methodologies, the inputs and outputs are described (see Table 4.3). Based on these inputs and outputs, and a subjective assessment of similarities between activities based on the analysis of the methodologies, the activities across the methodologies are grouped together see Figure 4.14. For this synthesis, a group is created only if three or more activities can be assigned to this group. This grouping results in the identification of four main activities (define context, define measurement method, perform measurement and analysis) and a total of eight critical activities (define business processes, define data and relations, define goals and requirements, identify dimensions for assessment, select objects for assessment, subjective measurement, objective measurement and analysis) of a data quality assessment process. In total, four activities could not be grouped as they did not have similarities with activities from other methodologies, either because they are too specific for a given methodology or because they just did not appear in other methodologies. A big challenge in this grouping process is that throughout methodologies, activities are often defined on different levels of abstraction and detail.

| Methodology | Activity | Input | Output |
|---|---|---|---|
| TDQM | Define IP characteristics | - | Data functionalities, components and relationships |
| TDQM | Define IQ requirements | Perspectives from different roles | Relevant IQ dimensions |
| TDQM | Define Information Manufacturing System | - | Data production process |
| TDQM | Define data quality metrics | Relevant dimensions, business rules | Data quality metrics |
| DQA | Conduct questionnaire | Data quality dimensions | Subjective DQ dimensions scores |
| DQA | Define objective measures | Functional forms for objective measures | Objective DQ dimension scores |
| DQA | Comparative Analysis | DQ dimension scores | Discrepancies |
| DQA | Identify improvement directions | Discrepancies | Improvement directions |
| DQAF | Data profiling | - | Data structure, content, rules and relationships |
| DQAF | Define expectations | Data structure, content, rules and relationships | Expected data quality values |
| DQAF | Objective measurement | Data rules | Objective quality scores |
| DQAF | Comparative Analysis | Data rules, quality scores | Improvement directions |
| Hybrid | Select data items and measurement place | - | Data items for quality measures |
| Hybrid | Identify reference data | - | Reference data for comparative metrics |
| Hybrid | Identify DQ dimensions and metrics | Data items | DQ dimensions and metrics |
| Hybrid | Perform measurement | DQ metrics | Measurement results |
| Hybrid | Analyze results | Measurement results | - |
| AIMQ | Identify relevant dimensions | PSP/IQ model, stakeholder perspectives | Relevant dimensions |
| AIMQ | Conduct questionnaire | Relevant dimensions, questionnaire items | Subjective DQ dimensions scores |
| AIMQ | Benchmark gap analysis | Dimension scores, benchmarks | Improvement directions |
| AIMQ | Role gap analysis | Dimension scores across roles | Improvement directions |
| ORME-DQ | State reconstruction | - | Organizational units, processes and data |
| ORME-DQ | Loss event analysis | Cost classification | Loss events |
| ORME-DQ | Select processes and databases | Loss events | Critical processes and databases to be measured |
| ORME-DQ | Select and perform quality measurements | Data quality metrics | Qualitative and quantitate measurement results |
| ORME-DQ | Analyze loss event probability | Measurement results | Loss events probabilities and criticality |
| DWQ | Obtain abstract quality goals | Stakeholder goals | Abstract quality goals |
| DWQ | Identify relevant data quality dimensions | Abstract quality goals, data warehouse context | Relevant DQ dimensions |
| DWQ | Assign weights to dimensions | Stakeholder opinions | Dimension importance weights |
| DWQ | Translate quality goals into executable queries | Abstract quality goals, data warehouse context | Data quality measurement queries |
| DWQ | Obtain scores for quality dimensions | Data quality measurement queries | DQ dimensions scores |
| DQALCA | Define data quality goals | Data user goals/expectations | Data quality goals |
| DQALCA | Select and collect data | Data quality goals | Databases and objects for measurement |
| DQALCA | Obtain quality scores | Pedrigree matrix, physical measurements, expert feedback | Data quality scores |

*Table 4.3: Activity inputs and outputs*

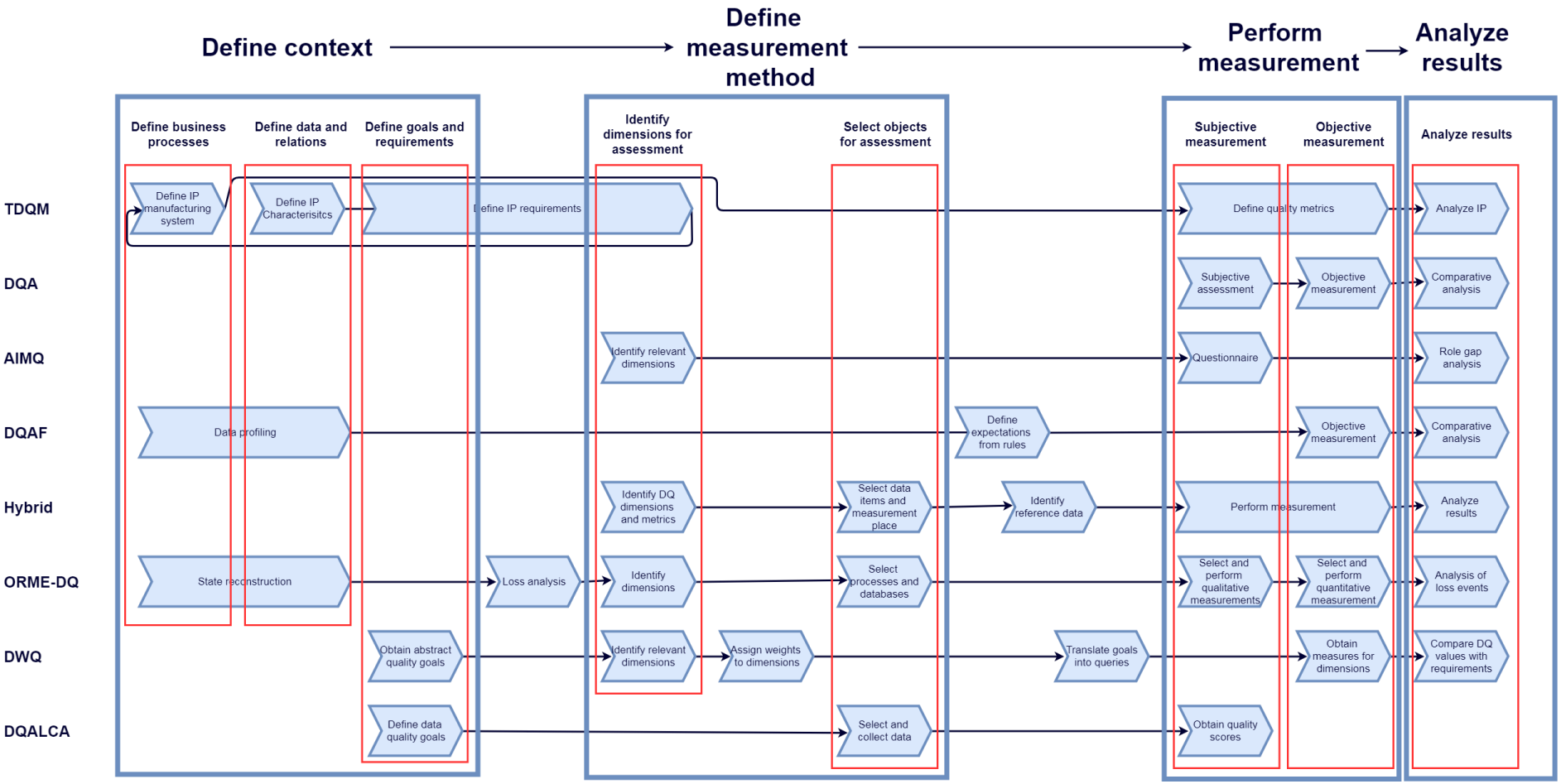



*Figure 4.14: Activity grouping and identification of critical activities*

### 4.4.2. Identifying roles in data quality assessment

A similar synthesis is performed on the roles that are mentioned throughout the methodologies: for each methodology, the roles mentioned are identified and grouped on their similarity. First, all roles throughout the methodologies have been identified along with the activities that they are involved in. This is presented in Table 4.4.

| Methodology | Role | Responsibility |
|---|---|---|
| TDQM | Information suppliers | Define IP requirements |
| TDQM | Information manufacturers | Define IP requirements |
| TDQM | Information consumers | Define IP requirements |
| TDQM | IP managers | Define IP requirements |
| DQA | Data consumer | Subjective assessment |
| DQA | Data custodian | Subjective assessment |
| DQA | Data provider | Subjective assessment |
| DQA | Manager | Subjective assessment |
| DQAF | Data user | Define expectations from data rules |
| DQAF | Data producer | Define expectations from data rules |
| AIMQ | Data consumer | Subjectively assess data quality (by a questionnaire) |
| AIMQ | IS professional | Subjectively assess data quality (by a questionnaire) |
| ORME-DQ | Data quality expert | Select and perform quality metrics |
| DWQ | Stakeholders | Define quality goals, assign dimension weights |
| DQALCA | Data user | Define data quality goals |

*Table 4.4: Roles throughout methodologies*

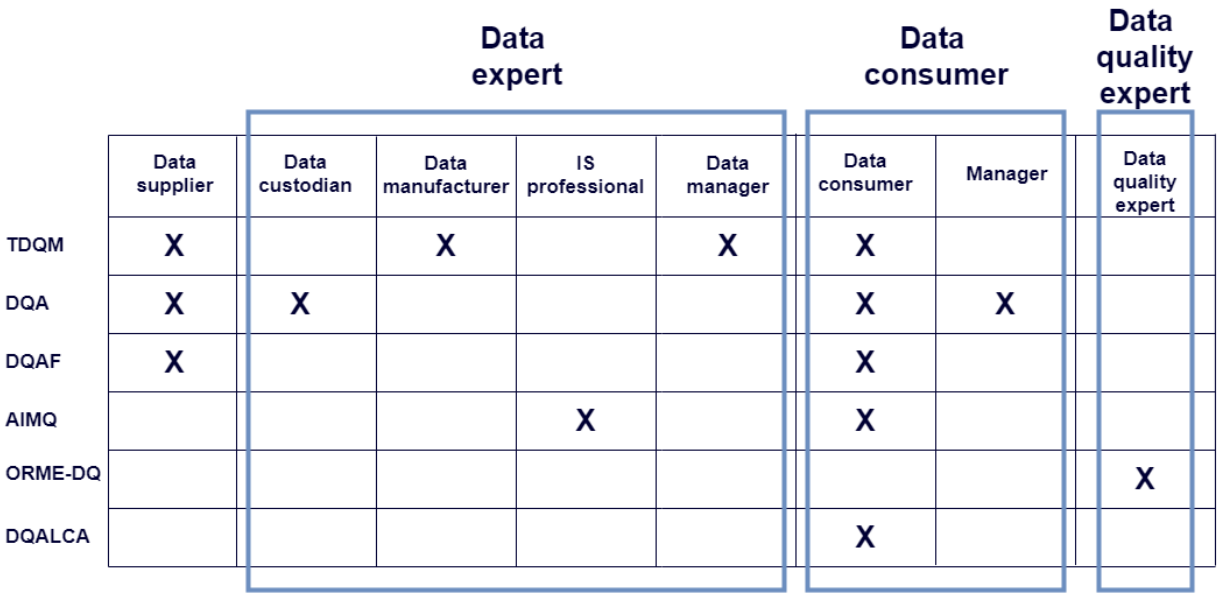



*Figure 4.15: Role grouping and synthesis*

Although adopting different names, there are eight different roles that are identified throughout the methodologies (for example the roles of information supplier in TDQM and data provider in DQA are considered under the same name in the synthesis: Data supplier). The appearance of these eight roles throughout methodologies can be found in Figure 4.15. Three roles that are of importance can be identified for data quality assessment: data experts, data consumers and data quality experts. More information and definition of these groups can be found in section 4.5.2.

## 4.5. Conclusions

Based on the synthesis of the analyzed methodologies the research questions are answered. The critical activities and the roles that play a part in the process are identified and further defined. These results are direct input for the development of a process model.

### 4.5.1. Activities

This section describes the critical activities identified in the synthesis. As can be seen in Figure 4.14, four main activities can be identified from the synthesis. Each activity is numbered to ensure that they are included in the process model designed in Chapter 5.

Main activity 1: Define context

As the context provides important input for the selection of relevant dimensions and metrics for that context, most methodologies include a definition of this context in some way. In the methodologies analyzed, the context is typically defined by the following:

| | |
|---|---|
| 1.1 Define business processes: | Provide a clear description and contingent graphical representations of the business processes that create, modify or consume the data to be assessed. |
| 1.2 Define data and relations: | Provide a definition of the data that is to be assessed: identification of data objects, types and relations. |
| 1.3 Define goals and requirements: | Identify the goals of the data consumers related to this data and the subsequent requirements to successfully perform their tasks. |

Main activity 2: Define measurement method

With the context defined, the next step is to find a way to measure relevant quality characteristics for that specific context.

| | |
|---|---|
| 2.1 Identify dimensions for assessment: | Translate the goals and requirements into data quality dimensions and select these relevant dimensions. |
| 2.2 Select objects for assessment: | Select the exact information systems, datasets, tables and data objects, on which the measurements are performed based on goals, requirements, dimensions and data availability. |

3. Main activity: Perform measurement

After relevant quality dimensions are specified and the objects for the measurements are selected, the actual measurement can be performed. Measurements can be subjective (i.e. based on opinions and typically measured with a questionnaire) or objective (based on calculations of rule compliance).

| | |
|---|---|
| 3.1 Subjective measurement: | Create and obtain subjective measures (typically questionnaire items) |
| 3.2 Objective measurement: | Create objective measures: define the calculations to be made for the metrics and conduct these calculations on the selected data objects. |

4. Main activity: Analysis

Typically, the last step of data quality assessment is to analyze the measurement results. The purpose of this analysis varies widely throughout the methodologies. It can either compare the obtained results to industry benchmarks (AIMQ), compare objective scores with subjective scores (DQA) or compare the measured values to previously set targets. In all methodologies, the analysis aims to identify problem areas and improvement directions.

### 4.5.2. Roles

Six of the eight analyzed methodologies mention one or more roles that somehow participate in the assessment process (see Figure 4.15). Wang (1998) is the only paper of the analyzed methodologies that provides a definition of the roles (see Table 4.5), in the other papers they are only mentioned.

| Role | Definition by Wang (1998) |
|---|---|
| **Data supplier** | those who create or collect data for the Information Product |
| **Data manufacturer** | those who design, develop, or maintain the data and systems infrastructure for the Information Product |
| **Data manager** | those who are responsible for managing the entire Information Product production process throughout the Information Product life cycle |
| **Data consumer** | those who use the Information Product in their work |

*Table 4.5: Roles defined by Wang (1998)*

For this study, the following roles are identified that participate in a data quality assessment process:

Data consumer: most methodologies that mention roles, mention data consumers: those that use the data in their work. Considering the definition of data quality as “fitness for use”, the opinion and experience of the usage of data by data consumers is important input for identifying the goals and requirements of data.

Data expert: Even though the papers distinct between data custodians, data manufacturers, IS professionals and data managers, for this study we combine these roles as they have the same

functionality in an assessment process: they provide contextual knowledge. Data experts know how the data is collected, modified and consumed, and how data objects are defined and related to each other.

Data quality expert: Although the role of a data quality expert is mentioned by only one paper, in this study it is considered as an important role in data quality assessment. A data quality expert takes the lead in the data quality assessment process: by combining the knowledge of data experts, the experience of data consumers and his own knowledge of data quality, he ensures a complete and correct data quality assessment.

Although data suppliers are mentioned in three papers, for this study they are omitted for the following reason: data suppliers have a role in the way that data is obtained, not necessarily in the way that this data is used (which highly determine the quality of data considering the definition “fitness for use”). Data suppliers therefore do not possess the right experience or opinions to identify data goals, requirements or experienced problems. In data quality improvement however, data suppliers can have an important role, as they know how data is obtained.

# 5. Process Model Development

This chapter describes the development of a process model for data quality assessment, based on the results of the literature review. The solution objectives defined in Chapter 3.2 are satisfied by performing and including the following:

- Practical utility: for the process model to be practical, activities are defined on a low level, with descriptions on how to perform them. This will eliminate any vagueness and makes the model easy to interpret.
- Comprehensiveness: for the process model to be comprehensive, all the critical activities identified in the literature review are included.
- Genericness: for the process model to be generic, it should be kept in mind during the design that it has to be applicable independent of context. Each activity and technique used for in the process must be achievable and relevant in any context.
- Understandability: to ensure understandability, the process model must comply to modelling rules, and it must have a clear and easy to understand presentation.
- Completeness: to ensure completeness of the assessment results of the process model, different approaches to data quality (problem-driven and requirement driven, subjective and objective measurement) are included.

The designed process model can be found in Figure 5.1,Figure 5.2,Figure 5.3 and Figure 5.4. To ensure that all critical activities (described and numbered in section 4.5.1) are included, their related activities in the process model include the numbers as they are defined in section 4.5.1. The practical descriptions of the activities and definitions of the data objects in the process model can be found in Appendix V: Detailed descriptions of activities and data objects.

## 5.1. Explanation of the model and design choices

This section explains the model and the choices made while designing the process model.

### 5.1.1. Scope definition

As is identified in the literature synthesis, the first step of a data quality assessment process is to define the context, consisting of the business processes, the data and relations between data and the goals and requirements of data consumers. These business processes and the data to be assessed together form the scope of the assessment, and therefore these activities are grouped together in the subprocesses ‘Define Scope’ (Figure 5.2). The activities consist of the following:

- Defining the business processes implies providing a BPMN model of the business process related to the data to be assessed. Although there are many ways to describe a business process, BPMN is used as it provided the opportunity to model both activities and data objects involved with these activities.
- Defining the data and its relations implies creating a UML class diagram of the data objects. A UML Class diagram is chosen as it clearly maps out the structure, relation and attributes of and between objects.

Furthermore, to obtain a clear definition and better understanding of the context, a mapping of the business processes to the data objects is added, describing how the processes consume, create or modify data. Also, as defining the scope of the assessment includes deciding on which people to involve, a

stakeholder analysis is included, and based on this analysis, relevant stakeholder groups can be selected for participation, and subsequently, the roles identified in this study can be assigned to individuals. In order to evaluate this scope definition (i.e. definition of business processes, data objects and relations and stakeholders involved), the model includes a review iteration with a data expert.

### 5.1.2. Define dimensions and metrics

As is described in section 1.2, the process model should include both a bottom-up (problem-driven) and top-down (requirement-driven approach). A bottom-up data quality assessment assesses data quality based on experienced problems by data consumers and the compliance to data rules that follow from referential integrity, functional dependencies and attribute analysis. Therefore, after defining the scope, the activity 'define rules' is included in which these rules are identified. The experienced problems are identified through semi-structured interviews with data consumers. These interviews also used to identify the goals of the data consumers (to include a top-down assessment approach). Semi-structured interviews are chosen as they allow for asking standardized questions to all consumers, and for going into more depth on specific goals or experienced problems. After conducting these activities, the subprocess 'Defining dimensions and metrics' can start. To model both a top-down and a bottom-up approach, this subprocess contains two parallel paths:

- The top-down approach: in which a set of dimensions is defined based on the identified goals from the interviews. After this set of dimensions is defined, metrics (both subjective and objective) can be designed for each dimension.
- The bottom-up approach: in which metrics (both subjective and objective) are created for each rule and for each identified problem experienced by data consumers. Thereafter, these metrics are grouped into dimensions.

By combining the results of the top-down and the bottom-up approach, the complete set of dimensions and metrics can be created. Based on this set, the data-objects objects (information systems, tables, attributes, history etc.) on which the measurements are performed can be selected. After obtaining these results, the metrics are reviewed with both data consumers (to evaluate whether they reflect the experienced problems and goals of data consumers) and data experts (to evaluate whether the metrics are valid and measure what they intend to measure). . Criteria for metrics are defined by RUMBA; metrics should be Reasonable, Understandable, Measurable, Believable and Achievable (see Kovac et al., 1997) for developing RUMBA data quality metrics). Also, weights are assigned to metrics by data consumers and data experts based on their opinion of the extent to which the metric represents the intended dimensions.

### 5.1.3. Perform Measurement

As measurements are performed subjectively and objectively, this subprocess contains two parallel paths. The subjective measurement implies the conduction of a questionnaire, of which the items are created during the development of metrics (in the previous subprocess). This questionnaire also serves to obtain the dimensions weights by asking the participants to their perceived importance of each dimensions to measure data quality. Parallel to the conduction of a questionnaire, the objective metrics can be performed (i.e. calculated over the selected objects, tables, attributes and data history). This subprocess yields a subjective measurement in the form of answers to questionnaire items, and objective measurement in the form of calculated formulas.

### 5.1.4. Analysis and reporting

Finally, the results of the questionnaire and the objective measurements are combined. Using the metric weights, a final score can be obtained for each dimension, and using the dimensions weight, a final overall data quality score can be obtained. Reporting includes the creation of a data quality report (describing the results and a description of the process) and distributing it to stakeholders.

## 5.2. Process Models

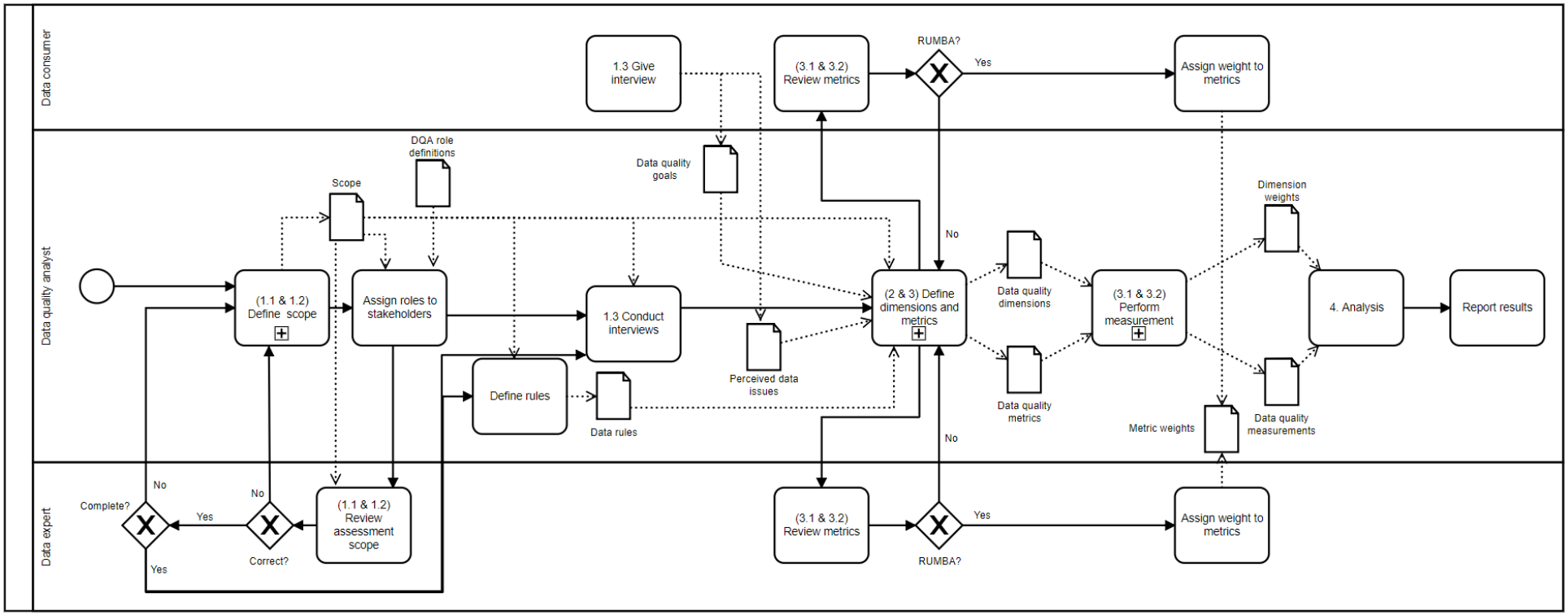



*Figure 5.1: Process model for data quality assessment*

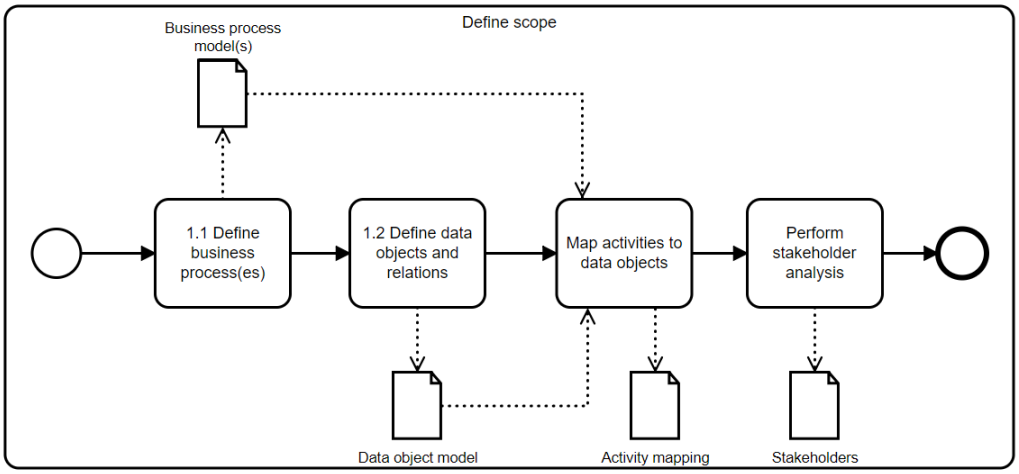



*Figure 5.2: Subprocess Define Scope*

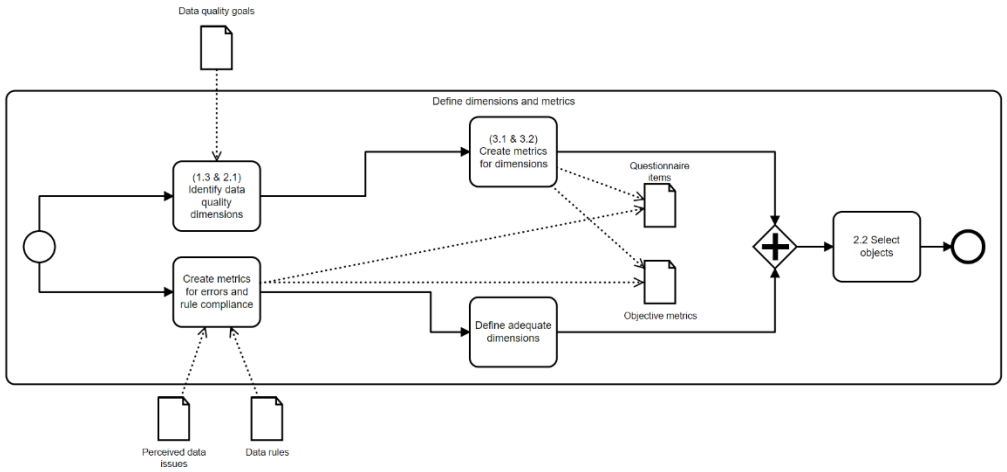



*Figure 5.3: Subprocess Define dimensions and metrics*

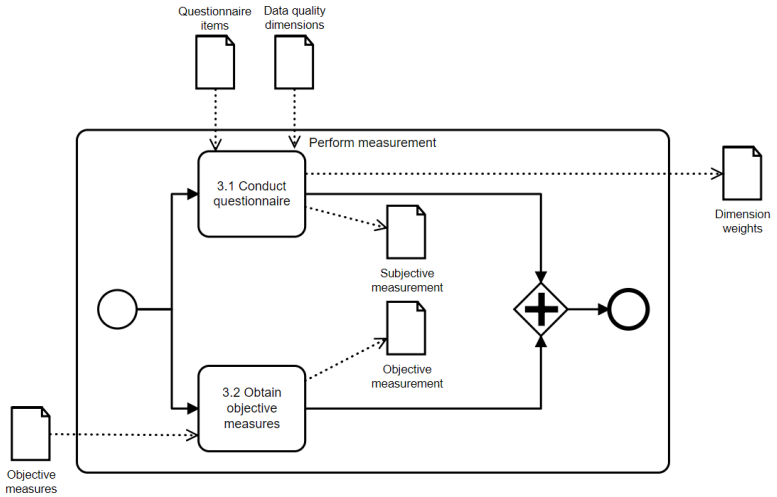



*Figure 5.4: Subprocess Perform measurement*

# 6. Demonstration and Evaluation

A case study will be conducted to demonstrate the use of the proposed assessment process, and to validate the process according to the defined solution objectives. Case Study research methodology by Yin (2003) was used as a guideline to set up and execute the case study. Based on the book of Yin (2003), the following steps are applied for this case study:

- Describe the case study context: the context in which the case study is conducted is described regarding the current data quality practices at ASML and for the specific case of the EUV factory.
- Define Case Study design: the case study design describes the sort of case study (i.e. multiple or single case, holistic or embedded) and the case study protocol (the procedure of conducting the case study, including the time frame and planning and the selection of participants).
- Preparing for data collection: including general training, training for the specific case, and providing the participants of the required knowledge.
- Collecting evidence: describing the process of conducting the case study (i.e. applying the process model in the case of ASML and the results thereof and collecting data for validating the process according to the defined solution objectives).
- Analyze evidence: determining the appropriate analysis strategy and techniques.

## 6.1. Case Description

This case study will be conducted at the EUV factory of ASML. At the EUV factory, they currently recognize the necessity to improve the quality on cycle time and labor hour data of the activities that are performed in the EUV factory. The production process in the EUV factory is extremely complex (i.e. consists of many steps, with many dependencies between these steps and many exceptions and deviations from standard procedures, frequent job reworks, machine failures/repairs/tests and other uncertainties induced by the complexity of the end product), causing many errors in their cycle time and labor hour records. The case of ASML provides the perfect opportunity to propose a data quality assessment process and to serve as the validation for this research.

The production of an EUV machine is divided into orders and orders are divided into milestones: pre-defined packages of steps to build a part of a machine, assemble parts, conduct repairs or maintenance, or other activities that support production. Because ASML wants to obtain transparency in the cycle times and required labor of each of these milestones and steps, the execution times of each step is logged, using a software tool called SAP Manufacturing Execution (SAP ME). This implies that an operator in the factory has several buttons he must click (for example start/stop/absent etc.) to log his activity and step durations. The data obtained from these loggings is used for cycle time and labor hour reporting, and ultimately for setting up cycle time improvement projects and optimal labor allocation. However, the quality of these loggings is currently of low quality. In practice, there are a lot of data cleaning activities on cycle time data, before it is used for reporting and analysis. The labor hour data that is collected in SAP ME is not used at all, currently a theoretical calculation of labor hours used is deployed for labor hour analysis.

## 6.2. Case Study context

### 6.2.1. Data Quality Management at ASML

ASML is a high-tech company that produce very complex machines (see Appendix VI: Organizational and departmental background of ASML EUV for organizational background) for chip manufacturers. Because of the complexity of ASML's end products, the company is highly data-driven, in the sense a lot of decisions are made based on data and data-analytics rather than intuition and personal experience. Because the company is so dependent on their data they have extensive data management practices in place. Considering the data management maturity model (DMM) described in (IT Governance Institute, 2007), overall ASML has a level 5 of data management maturity (optimized), meaning that they continuously aim to improve and optimize their data management practices based on changing organizational goals. However, ASML has a complex landscape when it comes to information systems and supporting business applications; some information systems are used on a large scale (across various departments) while others are used only within departments or even within specific teams. Also, there is a constant need for new business application solutions, causing a continuously changing set of information systems and applications in use. Implementing new information systems requires time before this system is fully customized to business needs and able to generate high quality data. This complex landscape of information systems and business application requires data quality management practices to be executed on a smaller scale; they are done for specific departments, information systems business application. For the case of the EUV factory, after the implementation of the software (SAP Manufacturing Execution) data quality management practices have been slowly picked up (see next section).

### 6.2.2. Data Quality Management practices for the case

As ASML recognized the necessity to improve the quality of cycle time and labor hour loggings, such that the quality of both labor hours and cycle times can be improved. A study is conducted to find the root causes of low data quality. By performing open interviews, this study identified the problems that cause this low quality, and found that problems can be categorized in software related issues (the software is not always suitable for the loggings to be done), the sequence of activities (which is not always logical and makes operators deviate from this sequence and therefore unable to log), the behavior of operators (operators refuse to accurately log, because they don't see the value of it) and the lack of knowledge and training on how to log. The study found a total of hundred root causes for the identified problems, indicating that there is a lot to be improved. However, as is mentioned before in this study, "one cannot manage data quality without being first able to measure it meaningfully" (Stvilia et al., 2007). Although there have been ad hoc analyses before on the quality of the cycle time and labor hour data, there are no standard and consistent methods for data quality measurement. Performing an initial assessment (as described in Table 1.1) is valuable for ASML, as it defines a measurement method that is consistent over time, and thus enables to monitor data quality improvement. By applying the process model described in the previous chapter, a complete assessment of the current quality of data is provided, and a measurement method for monitoring data quality improvement is created.

## 6.3. Case Study Design

### 6.3.1. Type of case study

Considering the types of case studies defined by Yin (2003), this study applies a single holistic case study. This means that the model will be applied for a single case using one unit of analysis. The rationale behind is the following: a single case allows for revelation: the opportunity to observe and analyze the use of the

process model in depth. As the study will be validated based on the opinion and experiences of individual participants of the case, a single unit of analysis is deployed, namely the individuals.

### 6.3.2. Case Study Protocol

The process model developed in Chapter 5 is applied for the case at ASML EUV to assess the data quality of cycle time and labor hour loggings; each activity in the process model was executed, and an assessment of data quality is obtained. The process model was executed over a period of eight weeks. In total, there were 11 employees that participated in the process: one was given the role of a data quality expert, one was given the role of a data expert (the researcher himself), and 9 were given the role of data consumers.

#### *6.3.2.1. Preparing for data collection*

The data to be collected is the experience and opinions of the participants of the process. By analyzing these experiences and opinions, the process model can be evaluated on the defined solution objectives. Before the process model can be implemented, it is important that all participants know what role they have (data expert/ data consumer/ data quality expert) and what it means, what the activities in the process model are and what their responsibilities and tasks are. To ensure this, the chosen participants were informed through a few meetings, in which the process model was presented, and the roles and responsibilities are explained. After all participants were informed, the execution of the process model started.

#### *6.3.2.2. Collecting evidence*

The next step is to collect evidence that process model conforms to the defined solution objectives (the process model is practical, comprehensive, generic, understandable and complete). Since the evaluation of the process model on the defined solution objectives is based on the experience and opinions of participants of the case study (i.e. those involved in the execution of the process model), evidence is collected using interviews. After all steps of the process model are executed, semi-structured interviews are performed with three (closely involved) participants of the process: the data expert and two data consumers. Semi-structured interviews are chosen for this evaluation as they allow for obtaining comprehensive experiences and opinions regarding the use of the process model for each of the solution objectives. For each solution objective, several standard questions (that will be asked to all participants) are defined (see Table 6.1). Based on the given answers, in-depth questions may be asked to obtain a deeper understanding of experiences and opinions. The interviews are held with each participant individually. The interviews are recorded so that the answers could be analyzed later. Each interview had a duration of approximately twenty minutes.

| Objective | Interview Questions |
|---|---|
| **Practical Utility** | - Do you think that the proposed process model is practical? |
| | - Do you think activities and roles are defined on a low-level and are not abstract? |
| | - Have you experienced any vagueness in the definition or description of activities or roles? |
| **Comprehensiveness** | - Do you think that the process model includes all critical activities of data quality assessment? |
| | - Do you think there are critical activities missing in this model? |
| | - Do you think there are roles missing in this model? |
| | - Do you think that the model approaches data quality from a broad perspective? |
| **Genericness** | - Do you think this process model can be easily applied in other contexts? |
| | - Do you feel like every activity is defined independent of this context? |
| | - Do you feel like every role is defined independent of this context? |
| **Understandability** | - Do you think that the process model is clearly depicted? |
| | - Do you think the process model conforms to BPMN rules? |
| **Completeness** | - Do you feel like the final assessment gives a complete overview of the current state of data quality? |
| | - Do you feel like there are other data quality problems or goals that are not represented in this assessment? |

*Table 6.1: Interview questions for semi-structured interviews for evaluation*

#### *6.3.2.3. Analyze evidence*

For the analysis of the evidence, the five-step approach described by LeCompte (2000) is used as a guideline. First, to get to know the data (i.e. the interview answers) each recording has been replayed several times. Step two is to focus the analysis. As we want to obtain an evaluation of the model on each solution objective, the analysis is focused by solution objective (which suits the setup of the interview questions). The third step is to categorize the information (also known as coding or indexing). By listening to each recording several times, relevant words, phrases, sentences and sections are noted for each solution objective. Based on these notes, the answers are summarized per solution objective per participant. The next step of LeCompte (2000) is to identify categories: categories can be identified through the identification of themes and patterns. For this study, the analysis of the interviews is kept simple and the solution objectives are used as categories. This resulted in an overview of the main comments of participants on each solution objective. These results are discussed and are the basis for conclusions (the last steps of the approach of LeCompte (2000)).

## 6.4. Case Study Results

### 6.4.1. Execution of the process model

The first step of the process model is to define the scope: consisting of the business process(es), the data objects, a mapping the business process to the data objects and a stakeholder analysis. This yielded the following:

- The presentation of a process model (in BPMN) of an operator performing all possible loggings in the factory (see Appendix VII: Case Study Results of the process model.
- The presentation of the data objects, relation and attributes using an UML class diagram. (see Appendix VII.2).

- A mapping of activities to data objects: for each activity in the process model, the relation (if any) to each object in the UML class diagram is described (see Appendix VII.3)
- A stakeholder analysis: identifying the different stakeholders that have an interest in high quality loggings.

After the scope was defined, the roles of data quality expert, data expert and data consumers were assigned to individuals based on the stakeholder analysis. Considering the scope and time frame of the case study, only members of the business engineer team (the team in which the case study is conducted) that work with the defined data are involved: in total 11 participants were selected; 1 data quality expert (the researcher himself), 1 data expert and 9 data consumers. The outputs of the scope definition were reviewed with the data expert, after which minor changes were made. Then, data rules were defined based on functional dependencies, attribute analysis and referential integrity. This resulted in 11 integrity rules, 8 functional dependency rules and 5 rules for individual attributes. After defining the rules, the interviews were conducted: all the data consumers were interviewed for identifying the data quality goals (what is the data used for and thus, what characteristics should it have), and the experienced problems. This yielded a set of five main goals, and 12 problems (see Appendix VII.4 and Appendix VII.5). Problems were only considered in the rest of the assessment when they were mentioned (or a similar problem was mentioned) by 2 or more data consumers. For each of these goals, problems and rules, metrics were created, resulting in a measurement model consisting of 8 dimensions and 36 metrics (see Appendix VII.6). The definition of these dimensions within the defined context are the following:

- Integrity: the referential integrity of attributes with unique values and ID's (i.e. the same ID's or unique attributes are not combined with other ID's or unique attributes if they have a 1-to-1 or 1-to-many relation.
- Consistency: The compliance of attributes to specific data rules based on functional dependencies.
- Validity: The compliance of data values to defined domains and data types.
- Accuracy: The ability of the data to reflect the actual cycle times and labor hours in the factory.
- Completeness: The completeness of records in the data set.
- Rationality: The degree to which the definition of data objects is rational for the tasks to be done.
- Comprehensiveness: the degree to which the data can provide insightful and the required information.
- Obtainability: the degree to which the data is obtainable at an acceptable quality level.

Each metric has been defined, such that it scores between 0 and 1. Three of these dimensions were measured subjectively, using questionnaire items, each item consisting of multiple questions referring to specific problems/ goals or rules (see Appendix VII.7). The weights of the objective metrics were determined in a meeting with two data consumers and the data quality expert. For the subjective measures, the questionnaire items were averaged. The objective measures were obtained over all the data records that had been collected in the current quarter (the fourth quarter of 2018). Combining the results of the objective metrics scores, the questionnaire scores (subjective measurement), and the weights, resulted in a final data quality score (see Appendix VII.8).

### 6.4.2. Interview results

As described in section 6.3.2.3, the interview results are obtained by getting to know the data (i.e. listening to the interview recordings multiple times), by noting relevant words, phrases, sentences and sections,

and subsequently, based on these notes, summarizing the essence of the answers of each participant on each solution objective. These results can be found in Table 6.2.

| | Participant 1 | Participant 2 | Participant 3 |
|---|---|---|---|
| **Practical Utility** | - Basically cannot be simpler<br>- In practice, probably not every step will be followed<br>- I appreciate the simplicity of roles | - In practice, probably not every step will be followed<br>- In practice, hard to make clear distinctions between these roles | - The translation of quality goals into dimensions is done without any argumentation and naming of dimensions is discussable |
| **Comprehensiveness** | - Add an extra validation loop after obtaining the measures | - You should validate your metric calculations with a data quality expert<br>- If data is defined "fitness for use" it's assessment should differ among different consumers | - There have been decisions made that are not included in the process model |
| **Genericness** | - It is generic as the definition of the data quality model is fully customized for the case | - In other context data collectors can play an important role | - It is generic, no further comments |
| **Understandability** | - BPMN is a good way to present the model and it is presented in an understandable format | - It does need a clearer set of instructions for execution, on itself, the model is hard to interpret | - Consider adding a simplified model containing only main phases |
| **Completeness** | - It is complete, no further comments | - This data is used by a bigger audience then considered, therefore goals and problems might not be complete | - Metrics should be normalized for better interpretation of results |

*Table 6.2: Interview results*

## 6.5. Findings

First of all, each of the participants were rather positive about the model considering each solution objective, indicating that the model is a good solution to the identified research gap. The analysis of the interviews found the main comments (both positive and negative) on each of the solution objectives by each participant. This can be summarized in the following findings:

- In practice, companies are unlikely to follow every step of the model, but rather select the steps to their interest.
- In practice, it can be hard to make a clear distinction between the roles of data expert and data consumer, as employees often have responsibilities of both of these roles.
- The step "translating goals into quality dimensions" is done without argumentation and the naming of dimensions is discussable.
- Adding a validation loop with a data expert after obtaining measurement results is valuable.
- The model can be further extended by creating different data quality models (different sets of dimensions and metrics) for different data consumers (since data quality is defined as "fitness for use", it should be assessed specifically for specific data consumers.
- The model does not contain some decisions that need to be taken in practice
- The model is highly generic, as the applied data quality model (set of dimensions and metrics) is fully customized.
- Data collectors do not play a role in the process model, although in some contexts they can provide valuable input for goal- and problem identification.
- The presentation of the model is understandable (by using BPMN), however a clear set of instructions for execution is required. Also, a simple, high-level overview of the process can be valuable.
- The model does not normalize metrics, even that would make the interpretation of results easier.

## 6.6. Discussion of results

A discussion of the results of the semi-structured interviews are described for each solution objective in the following.

### 6.6.1. Practical Utility

All three of the interviewees answered that they think that the proposed process model is practical. One noted that the process model cannot be simpler that it is, that it is ready for execution and easy to interpret. However, two of the three interviewees also mentioned that, according to their experiences, chances are that within an organization the model will most likely not be followed from beginning to end. Rather, users of the model will use it as a guideline and execute the activities that they consider relevant for obtaining a quality assessment. Although this means that the model is still useful, a complete assessment of data quality cannot be guaranteed, as all activities are closely related in terms of their inputs and outputs. A configuration guide could solve such a problem. Such a guide enables practitioners to pick the activities they consider important for their case, while ensuring that the required inputs for each activity is obtained (like the configuration method presented by Woodall et al. (2013)). Regarding the roles, all three interviewees agreed that the role definitions are clear. One interviewee noted that he appreciates the simplicity of the roles, and that there is no further distinction between different consumers and experts, making execution of the model easier in practice. However, one also noted that,

it will be hard to make clear distinctions between the defined roles in practice, as you will often see that data consumers often know a lot about the data production process (i.e. the knowledge of a data expert) and vice versa: a data expert often has tasks that would categorize him in the role of a data consumer. Although in practice the line between data expert and consumer is rather subjective, their influence in the process model will differ greatly (i.e. a data consumer does not play a role in context definition, and a data expert does not provide data goals and experienced problems). Furthermore, one interviewee mentioned his doubts for an activity of the process model: the translation of the quality goals into dimensions. The activity is done without any argumentation, and the naming and definition of dimensions is discussable.

### 6.6.2. Comprehensiveness

All three interviewees agreed that all critical activities of data quality assessment are included. However, two of the three interviewees mentioned the same element that can be added; an extra validation loop after obtaining the objective measures. In the current process model, there are reviews after the context definition and after the creation of metrics. However, as the case study pointed out, the process of obtaining the measures is not as easy as it seems: it requires data to be collected, merged, transformed and compared, and during this process, mistakes or false assumptions can easily be made. For this reason, another review session after obtaining the results can be valuable, to filter out such mistakes or false assumptions and to make sure that the used measures are valid and correct. Furthermore, one interviewee pointed out that, when data is defined as ''fitness for use'' for data consumers, the assessment of data quality will also differ for different consumers (i.e. data might be of high quality for one consumer, but of low quality for another depending on their tasks). The current model assumes all data consumers to be the same, and thus, the assessment to be valid for all consumers. Another interviewee however, pointed out that the simplicity of roles (and not making further distinctions) contributes to the practical utility of the process model. Finally, one interviewee pointed out that there have been one important decisions made that are not included in the process model. That is, the amount of history of data is considered for the objective measures; these measures can be calculated over data collected in the past two years, but also over data collected in the past two weeks. This decision will have a great impact on the final assessment, but the process model provides no guidelines to make this decision. All interviewees agreed that by defining metrics based on goals, problems and rules, the process model approaches data quality from a broad perspective.

### 6.6.3. Genericness

Two of the three interviewees informed that the process is generic and can be just as easily applied to another case, especially because the definition of the data quality model (the set of dimensions and subsequent measures and weights) are fully customized for the given context. One interviewee noted that the process model might be somewhat designed for the case specifically. This is mainly expressed by the fact that data collectors do not participate in the process: for this specific case that assumption can be made, as the data consumers have most of the knowledge that data collectors (operators in the factory) have, and they know about their experiences and opinions. However, in other contexts, data collectors and consumers might be far away from each other with no shared knowledge and experiences. In such contexts, this process model might miss valuable input (i.e. problems/goals) from data collectors that are of importance of data quality assessment. For this reason, including data collectors as an optional role in the process, might improve the ability of the model to identify all data problems and goals.

### 6.6.4. Understandability

All three interviewees answered that the process model is clearly presented, and that BPMN is a good way to present the process. Two comments were made for this matter: the process model does need a set of instructions (provided in the activity descriptions: Appendix V: Detailed descriptions of activities and data objects). On itself, the process model is hard to interpret. The other comment concerned the ease of presentation: a simplified overview of the process might be valuable, so that the main phases of the process can be seen in the blink of an eye.

### 6.6.5. Completeness

All three interviewees answered that, for the scope of the case study considered, the final assessment gives a complete overview of the current state of data quality. However, one mentioned that, the considered scope (the BE team) is rather small compared to the application of the data that has been assessed; throughout the department there are hundreds of data consumers and data experts that are involved with this data, and the process model could be applied on a much greater scale. However, for the scope considered (the people and data involved) it gives a complete overview of the current state of data quality. None of the interviewees could mention other goals, problems or rules that were not captured by the existing measures. One interviewee emphasized the importance of another review loop after obtaining the objective measures, as there were two measures for which the outcomes were hard to believe, and the suspicion existed that this was due to a false assumption made during the calculation of objective scores. Furthermore, one interviewee mentioned a problem concerning the scaling of metrics. In the current assessment, each metric is defined such that it obtains a result on a 0 to 1 scale. Subsequently, each measure is a given weight with which the final score of a dimension can be calculated, assuming that the weights accurately represent the importance of each measure for that dimensions. However, the score of a metric might also say a lot about its importance. For example; a metric that scores 0.98 would be considered good. But when this metric is defined as the uptime of a system, 0.98 is not that good, and the difference between 0.98 and 0.99 is a big difference. Although this is a big difference, it does not have a big impact on the final score of the dimension. A way to normalize metrics based on their impact and fluctuation would be a valuable improvement of the model.

# 7. Research Validity Threats

This research followed the design science research methodology for information systems research developed by Peffers et al. (2007). Following this methodology provided structure in the research process by providing clear guidelines on how to approach the research problem, how to come to a solution to this problem, and how to evaluate the solution. However, the following threats of the validity of study are identified that are worth mentioning.

An element that needed to be added to the research steps provided by the methodology, was the answering of two sub research questions that was required for the development of a new artefact: to develop a process model, the critical activities and roles of data quality assessment needed to be identified. These sub research questions were answered by conducting a literature review, analyzing the methodologies and papers found, and subsequently synthesizing the methodologies based on this analysis. This synthesis lead to the identification of critical activities and roles for data quality assessment. However, this synthesis is rather subjective, and in terms of reproducibility, could yield somewhat different results were the same synthesis to be done in a different research. Also, the results of this synthesis depend on the papers and methodologies selected (and thus on the search strategy), as well as on the (subjective) analysis of each of these methodologies (i.e. what activities and roles compose a methodology, and what inputs and outputs are there). To further validate this research method and its results, it would be useful the compare the results of the synthesis (and thus the answers of the research questions) to similar work (for example the common activities in data quality assessment described by Batini et al., 2009).

The development phase of the research included the development of the process model and providing descriptions of activities in the model. While designing this model, it was ensured that each of the identified critical activities in the synthesis were represented. However, these identified critical activities are defined on a somewhat abstract level. For the process model to be practical, a more practical interpretation of these activities was required. This translation from abstraction to practice also depend on the subjective interpretation of the researcher (i.e. there are more ways to translate the abstraction to a practice). For example, the critical activity “define data and relations” was given a practical interpretation by creating a data object model using an UML class diagram. However, there are more ways to define the data and its relations. Variations of the process model containing different practical interpretations are an interesting topic for further improvement of the model in further research.

The process model was evaluated using a set of solution objectives defined at the beginning of the study. These solution objectives are based on the problem that the artefact aims to solve and were obtained by straightforward reasoning of the researcher. However, there are possibly more solution objectives that are interesting for evaluation for this research, perhaps there are pre-defined sets of solution objectives for designing information system artefacts in the literature. A more complete set of solution objectives can provide for a better evaluation of the model, and thus can provide a more complete overview of possible improvements of the model.

Finally, to further validate the usefulness of the model, it will need more validation. Currently, the model is demonstrated in a single case study, and evaluated based on the opinion and experiences of participants in this case study. This can cause the evaluation to be biased to the results of this single case. Think for example of the solution objective genericness (the process model should be applicable independent of

context): it is hard for the participants of the case study to evaluate its ability to be applicable in another context. Therefore, applying this model in more cases, will result in a more complete evaluation.

# 8. Conclusion

Current research provides many data quality assessment methodologies and frameworks, containing extremely divergent activities, techniques and tools, not to mention the many different data quality dimensions and their taxonomies for which there is little consensus in current research, making data quality assessment a complex task in practice. Although obtaining a meaningful assessment can be challenging, the results are valuable, as they provide direct input for data quality improvement, and thus for better data.

This research proposed a process model for data quality assessment which is based on a synthesis. The identified research gap stated that existing data quality assessment methodologies are either generic or not operationalized for specific contexts, or they are developed for a specific context, technique or problem and therefore not practical for other contexts. This causes today's organization to adopt data quality assessment methodologies that do not fully fit their requirements or business needs. To tackle this problem, this study proposes a process model for data quality assessment that is both practical (i.e. operationalized) while still being generic (by including a customized measurement method, and multiple perspectives to approach data quality), and thus applicable independent of context. To ensure that the proposed model is indeed a solution to the identified research gap, five solution objectives were defined, and the proposed model was qualitatively evaluated on each of those, using interviews. The evaluation of the proposed model showed that the model was considered practical, generic, understandable and complete by participants of the case study. Several improvements can be made on the comprehensiveness of the model. Overall, the evaluation of the process model showed that the model is a solution to the research problem, and a valuable contribution for data quality practitioners in the field. However, also problems were identified for each solution objective, providing options to further improve the model. Potential improvements of the model include:

- The addition of a configuration guide
- The addition of an extra validation loop
- Applying a reference model for selecting dimensions
- Including a method for metric normalization
- Inclusion of data collectors for goal formulation and problem identification

In addition, to further validate the usefulness of the model, it will need more validation. Currently, the model is demonstrated in a single case study, and evaluated based on the opinions and experiences of participants in this case study. This can cause the evaluation to be biased to the results of this single case. Think for example of the solution objective genericness (the process model should be applicable independent of context): it is hard for the participants of the case study to evaluate its ability to be applicable in another context. Therefore, applying this model in more cases, will result in a more complete evaluation.

# 9. References


Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2), i–xxxii. https://doi.org/10.1017/CBO9781107415324.004

Ahmed, H. H. (2018). Data quality assessment in the integration process of linked open data (LOD). In *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA* (Vol. 2017–Octob, pp. 1–6). IEEE. https://doi.org/10.1109/AICCSA.2017.178

Albala, M. (2011). Making Sense of Big Data in the Petabyte Age. *Cognizant 20-20 Insights Executive Summary*.

Aljumaili, M., Karim, R., & Tretten, P. (2016). Metadata-based data quality assessment. *VINE Journal of Information and Knowledge Management Systems*, *46*(2), 232–250. https://doi.org/10.1108/VJIKMS-11-2015-0059

ASML. (2017). ASML: Press - Fact Sheet.

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, *44*(4), 462–484. https://doi.org/10.1287/mnsc.44.4.462

Batini, C., Barone, D., Maurino, A., & Ruffini, C. (2007). *A FRAMEWORK AND A METHODOLOGY FOR DATA QUALITY ASSESSMENT AND MONITORING (Practice-Oriented)*. Retrieved from http://mitiq.mit.edu/iciq/pdf/a framework and a methodology for data quality assessment and monitoring.pdf

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2006). *Data Quality Consepts, Methodologies and Techniques*.

Bicalho, T., Sauer, I., Rambaud, A., & Altukhova, Y. (2017). LCA data quality: A management science perspective. *Journal of Cleaner Production*, *156*, 888–898. https://doi.org/10.1016/j.jclepro.2017.03.229

Brown, J. S., Kahn, M., & Toh, D. (2013). Data quality assessment for comparative effectiveness research in distributed data networks. *Medical Care*, *51*(8 SUPPL.3), S22–S29. https://doi.org/10.1097/MLR.0b013e31829b1e2c

Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Proceeding of the 28th international conference on Software engineering - ICSE '06* (p. 1051). New York, New York, USA: ACM Press. https://doi.org/10.1145/1134285.1134500

Caballero, I., Verbo, E., Calero, C., & Piattini, M. (2007). A data quality measurement information model based on ISO/IEC 15939. In *Proceedings of the 12th International Conference on Information Quality*. MIT, Cambridge, MA, USA.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2. https://doi.org/10.5334/dsj-2015-002

Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, 68–73. https://doi.org/10.1145/1012453.1012465

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88. https://doi.org/10.1145/1978542.1978562

Dama International. (2009). *The DAMA Guide to The Data Management Body of Knowledge*. Technics Publications.

De Amicis, F., & Batini, C. (2004). A Methodology for Data Quality Assessment on Financial Data. *Studies in Communication Sciences*.

del Pilar Angeles, M., & García-Ugalde, F. J. (2009). A Data Quality Practical Approach. *International Journal on Advances in Software*, *2*(3).

Deming, W. E. (1986). *Out of The Crisis*. *Massachusetts Institute of Technology. Center for advanced engineering study*. MIT Press.

Eckerson, W. W. (2002). Data quality and the bottom line. *The Data Warehouse Institute*, 1–32. https://doi.org/10.1038/nprot.2010.116

Eppler, M. J. (2006). *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. Springer. https://doi.org/10.1007/3-540-32225-6

Eppler, M. J., & Muenzenmayer, P. (2002). Measuring Information Quality in The Web Context: A surve of State-of-the-Art Instruments and an Application Methodology. *Proceedings of the Seventh International Conference of Information Quality*, 187–196. https://doi.org/10.1.1.477.4680

Friedman, T., & Smith, M. (2011). Measuring the Business Value of Data Quality. *Gartner*, (G00218962).

Heinrich, B., Klier, M., & Kaiser, M. (2009). A Procedure to Develop Metrics for Currency and its Application in CRM. *Journal of Data and Information Quality*, *1*(1), 1–28. https://doi.org/10.1145/1515693.1515697

Hevner, A. R., & Chatterjee, S. (2010). Introduction to Design Science Research. In *Design research in information systems: Theory and practice* (Vol. 28, pp. 1–8). https://doi.org/10.1007/978-1-4419-5653-8

Huang, K.-T., Wang, R. Y., & Lee, Y. W. (1998). *Quality Information and Knowledge*. *Computerworld* (Vol. 33). Prentice Hall PTR. https://doi.org/10.1186/1756-8935-2-7

Hüner, K. M., Ofner, M., & Otto, B. (2009). Towards a maturity model for corporate data quality management. In *Proceedings of the 2009 ACM symposium on Applied Computing - SAC ´09* (p. 231). New York, New York, USA: ACM Press. https://doi.org/10.1145/1529282.1529334

ISO 8000-61. (2016). Data quality -- part 61: Data quality management: Process reference model. *International Organization for Standardization, Geneva, Switzerland*.

IT Governance Institute. (2007). *Cobit 4.1*. *Governance An International Journal Of Policy And Administration*. https://doi.org/10.1016/S0167-4048(97)84675-5

Jaya, M. I., Sidi, F., Ishak, I., Suriani Affendey, L., & Jabar, M. A. (2017). A REVIEW OF DATA QUALITY RESEARCH IN ACHIEVING HIGH DATA QUALITY WITHIN ORGANIZATION. *Journal of Theoretical and Applied Information Technology*, *30*(12).

Jeusfeld, M. A., Quix, C., & Jarke, M. (1998). Design and analysis of quality information for data warehouses. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1507, pp. 349–362). https://doi.org/10.1007/978-3-540-49524-6_28

Juran, J. M., & Godfrey, A. B. (1998). *Juran's Quality Handbook*. *McGrawHill*. McGraw Hill. https://doi.org/10.1108/09684879310045286

Kagermann, H., Österle, H., & Jordan, J. M. (2011). *IT-Driven Business Models: Global Case Studies in Tranformations*. John Wiley & Sons.

Kahn, B., Strong, D., & Wang, R. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, *45*(4), 184–192. https://doi.org/10.1145/505999.506007

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *2013 46th Hawaii International Conference on System Sciences* (pp. 995–1004). IEEE. https://doi.org/10.1109/HICSS.2013.645

Karr, A. F., Sanil, A. P., & Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, *3*(2),

137–173. https://doi.org/10.1016/j.stamet.2005.08.005

Kovac, R., Kovac, R., Lee, Y. W., Lee, Y. W., Pipino, L. L., & Pipino, L. L. (1997). Total data quality management: the case of IRI. In *Conference on Information Quality* (pp. 63–79). https://doi.org/10.1145/1141277.1141634

LeCompte, M. D. (2000). Analyzing Qualitative Data. *Theory Into Practice*, *39*(3), 146–154. https://doi.org/10.1207/s15430421tip3903_5

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to data quality*. *Computer* (Vol. 1). MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management*, *40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Madhikermi, M., Kubler, S., Robert, J., Buda, A., & Främling, K. (2016, November 30). Data quality assessment of maintenance reporting procedures. *Expert Systems with Applications*. Pergamon. https://doi.org/10.1016/j.eswa.2016.06.043

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality*, *8*(1), 1–29. https://doi.org/10.1145/2964909

Oakland, J. S. (1989). Total quality management. *Quality and Reliability Engineering International*, *5*(4), 339–339. https://doi.org/10.1002/qre.4680050414

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211. https://doi.org/10.1145/505248.506010

Redman, T. C. (1996). *Data Quality for the Information Age*. Artech House.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82. https://doi.org/10.1145/269012.269025

Reid, A., & Catterall, M. (2005). Invisible data quality issues in a CRM implementation. *Journal of Database Marketing & Customer Strategy Management*, *12*(4), 305–314. https://doi.org/10.1057/palgrave.dbm.3240267

Scannapieco, M., & Catarci, T. (2002). Data Quality under the Computer Science perspective. *Computer Engineering*, *2*(2), 1–12.

Sebastian-coleman, L. (2013). Measuring Data Quality for Ongoing Improvement : A Data Quality Assessment Framework, *1*.

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), 14–32. https://doi.org/10.4018/jdm.2003100102

Shardt, Y. A. W., & Huang, B. (2013). Data quality assessment of routine operating data for process identification. *Computers and Chemical Engineering*, *55*, 19–27. https://doi.org/10.1016/j.compchemeng.2013.03.029

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733. https://doi.org/10.1002/asi.20652

Su, Z., & Jin, Z. (2004). A Methodology for Information Quality Assessment in the Designing and Manufacturing Processes of Mechanical Products. In *Information Quality Management* (pp. 190–220).

https://doi.org/10.4018/978-1-59904-024-0.ch009

van Wierst, J. (2018). Organizational data quality assessment and improvement: a Literature Review. *Unpublished Manuscript*, 34.

Wan, Y., Shi, W., Gao, L., Chen, P., & Hua, Y. (2015). A general framework for spatial data inspection and assessment. *Earth Science Informatics*, *8*(4), 919–935. https://doi.org/10.1007/s12145-014-0196-9

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

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

Wang, R. Y., & Strong, D. M. (1996a). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Wang, R. Y., & Strong, D. M. (1996b). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Weidema, B. P., & Wesnæs, M. S. (1996). Data quality management for life cycle inventories-an example of using data quality indicators. *Journal of Cleaner Production*, *4*(3–4), 167–174. https://doi.org/10.1016/S0959-6526(96)00043-1

Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: The Hybrid Approach. *Information and Management*, *50*(7), 369–382. https://doi.org/10.1016/j.im.2013.05.009

Yin, R. K. (2003). *Introduction and designing case study*. *Case study research Design and methods*. https://doi.org/10.1017/CBO9780511803123.001

# 10. Appendix

## Appendix I: 70 data quality dimensions provided by Eppler (2006)

70 data quality dimensions provided by Eppler, (2006)

1. Comprehensiveness
2. Accuracy
3. Clarity
4. Applicability
5. Conciseness
6. Consistency
7. Correctness
8. Currency
9. Convenience
10. Timeliness
11. Traceability
12. Interactivity
13. Accessibility
14. Security
15. Maintainability
16. Speed
17. Objectivity
18. Attributability
19. Value-added
20. Reputation (source)
21. Ease-of-use
22. Precision
23. Comprehensibility
24. Trustworthiness (source)
25. Reliability
26. Price
27. Verifiability
28. Testability
29. Provability
30. Performance
31. Ethics/ ethical
32. Privacy
33. Helpfulness
34. Neutrality
35. Ease of Manipulation
36. Validity
37. Relevance
38. Coherence
39. Interpretability
40. Completeness
41. Learnability
42. Exclusivity
43. Right Amount
44. Existence of meta information
45. Appropriateness of meta information
46. Target group orientation
47. Reduction of complexity
48. Response time
49. Believability
50. Availability
51. Consistent Representation
52. Ability to represent null values
53. Semantic Consistency
54. Concise Representation
55. Obtainability
56. Stimulating
57. Attribute granularity
58. Flexibility
59. Reflexivity
60. Robustness
61. Equivalence of redundant or distributed data
62. Concurrency of redundant or distributed data
63. Nonduplication
64. Essentialness
65. Rightness
66. Usability
67. Cost
68. Ordering
69. Browsing
70. Error rate

## Appendix II: Collection of data quality dimensions and metrics from different methodologies (Batini et al., 2009)

| Dimensions | Name | Metrics Definition |
|---|---|---|
| Accuracy | Acc1 | Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one<br>Syntactic Accuracy=Number of correct values/number of total values |
| | Acc2 | Number of delivered accurate tuples |
| | Acc3 | User Survey - Questionnaire |
| Completeness | Compl1 | Completeness = Number of not null values/total number of values |
| | Compl2 | Completeness = Number of tuples delivered/Expected number |
| | Compl3 | Completeness of Web data = $(T_{max} - T_{current})^*$ $(Completeness_{Max} - Completeness_{Current})/2$ |
| | Compl4 | User Survey - Questionnaire |
| Consistency | Cons1 | Consistency = Number of consistent values/number of total values |
| | Cons2 | Number of tuples violating constraints, number of coding differences |
| | Cons3 | Number of pages with style guide deviation |
| | Cons4 | User Survey - Questionnaire |
| Timeliness | Time1 | Timeliness = $(\max(0; 1\text{-Currency/Volatility}))^s$ |
| | Time2 | Percentage of process executions able to be performed within the required time frame |
| | Time3 | User Survey - Questionnaire |
| Currency | Curr1 | Currency = Time in which data are stored in the system - time in which data are updated in the real world |
| | Curr2 | Time of last update |
| | Curr3 | Currency = Request time- last update |
| | Curr4 | Currency = Age + (Delivery time- Input time) |
| | Curr5 | User Survey - Questionnaire |
| Volatility | Vol1 | Time length for which data remain valid |
| Uniqueness | Uni1 | Number of duplicates |
| Appropriate amount of data | Appr1 | Appropriate Amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided)) |
| | Appr2 | User Survey - Questionnaire |
| Accessibility | Access1 | Accessibility = max (0; 1-(Delivery time - Request time)/(Deadline time - Request time)) |
| | Access2 | Number of broken links - Number of broken anchors |
| | Access3 | User Survey - Questionnaire |
| Credibility | Cred1 | Number of tuples with default values |
| | Cred2 | User Survey - Questionnaire |
| Interpretability | Inter1 | Number of tuples with interpretable data, documentation for key values |
| | Inter2 | User Survey - Questionnaire |
| Usability | Usa1 | User Survey - Questionnaire |
| Derivation Integrity | Integr1 | Percentage of correct calculations of derived data according to the derivation formula or calculation definition |
| Conciseness | Conc1 | Number of deep (highly hierarchic) pages |
| | Conc2 | User Survey - Questionnaire |
| Maintainability | Main1 | Number of pages with missing meta-information |
| Applicability | App1 | Number of orphaned pages |
| | App2 | User Survey - Questionnaire |
| Convenience | Conv1 | Difficult navigation paths: number of lost/interrupted navigation trails |
| Speed | Speed1 | Server and network response time |
| Comprehensiveness | Comp1 | User Survey - Questionnaire |
| Clarity | Clar1 | User Survey - Questionnaire |
| Traceability | Trac1 | Number of pages without author or source |
| Security | Sec1 | Number of weak log-ins |
| | Sec2 | User Survey - Questionnaire |
| Correctness | Corr1 | User Survey - Questionnaire |
| Objectivity | Obj1 | User Survey - Questionnaire |
| Relevancy | Rel1 | User Survey - Questionnaire |
| Reputation | Rep1 | User Survey - Questionnaire |
| Ease of operation | Ease1 | User Survey - Questionnaire |
| Interactivity | Interact1 | Number of forms - Number of personalizable pages |

## Appendix III: Databases searched for literature review

WorldCat.org
ScienceDirect (0)
Walter de Gruyter eBooks
ABI/INFORM Collection
JSTOR Health & General Sciences Collection
Oxford Reference Online
SAGE Journals
Annual Reviews
ACM Digital Library
MEDLINE
Academia
SPIE Digital Library
Wiley/IEEE Press Books
JSTOR Business I Collection
IEEE Publications Database
World Scientific Journals
SpringerLink
Royal Society of Chemistry Books
BMJ Journals
Mary Ann Liebert Online
Oxford Journals
JSTOR Mathematics & Statistics Legacy Collection
Nexis Uni
World Scientific eBooks
Royal Society of Chemistry Journals
JSTOR Arts & Sciences VII Collection
Walter de Gruyter eJournals
Brill Journals
Thieme Connect
INFORMS Journals
Emerald Group Publishing Limited
Wiley Online Library
JSTOR Arts & Sciences I Collection
Taylor and Francis Journals
SAE Technical Papers and Journal Articles
Hindawi eJournals
Elgaronline
IET Publications Database
Directory of Open Access Journals
Institute of Physics eJournals and Archive
BioOne
NARCIS

## Appendix IV: Search words used for literature review

Data
Information
Quality
Assessment
Methodology
Method
Approach
Process
Framework
Guide
Tasks
Activities
Roles
Organization
Organizational
Business
Company
Measurement
Measures
Metrics
Dimensions

## Appendix V: Detailed descriptions of activities and data objects

All activities and data objects presented in the process model are described in detail in this section to ensure that they are understandable and applicable:

**Activity descriptions**

Define business process(es): Clearly define the business processes that are related (i.e. create, modify or consume) to the data to be assessed. Provide a visual presentation of the process in BPMN, and a textual description of this model.

Define data objects and relations: Clearly define the data that is to be assessed. Create an UML class diagram and present this model in a UML class diagram to show all data objects, attributes and relations. Provide a textual description of this model

Map activities to data objects: To obtain a greater understanding and a clear definition of the context, map the activities in the process models to the data objects in the data object model: for each activity describe what data and how this data is created, modified or consumed.

Perform stakeholder analysis: Identify the stakeholders that are in any way involved with the data to be assessed. For each stakeholder, define their interests related to the data.

Assign roles to stakeholders: Assign the roles (data experts, data consumers) to the identified stakeholders and select individuals to participate in the process.

Review assessment scope: Using the knowledge of the data expert(s), check whether the defined business processes, data object models, activity mapping and stakeholder analysis is correct and complete. If, not redefine based on the obtained feedback, and review again.

Define rules: based on the obtained knowledge of the business processes and data objects, define logical rules based on functional dependencies (e.g. if attribute “marital status” has value “YES”, then attribute “Married to” must have a value), attribute analysis (e.g. attribute “gender” can only have two values) and referential integrity (e.g. every value of attribute “employee ID” in the manager database must appear in the employee database).

Conduct interviews: Conduct interviews with the selected data consumers. The interviews aim to identify the data quality goals and requirements (i.e. what should this data do?) and the experienced data problems (i.e. what is going wrong?). Semi-structured interviews allow for asking standardized questions to all consumers, and for going into more depth on specific goals or problems.

Identify data quality dimensions: translate the data quality goals identified in the interviews into data quality dimensions (e.g. data quality goal: “we want to enable a fast reporting of production progress” can be translated to timeliness)

Create metrics: Create metrics for the identified dimensions. Metrics can be either subjective in the form of questionnaire items, or objective, using a calculation.

Create metrics for errors and rule compliance: besides creating metrics based on data quality goals and requirements, metrics should also be created for experienced data problems, and rule violations (the

bottom-up approach). Translate the perceived data issues (identified from the interviews) and data rules into questionnaire items and objective metrics.

Define adequate dimensions: the metrics for errors and rule compliance can directly be created from the perceived errors and data rules. However, for the final reporting, metrics should be assigned to dimensions. Therefore, adequate dimensions should be assigned to these metrics.

Select objects: Based on the dimensions that need to be measured and the defined metrics, select the data objects (information systems, tables, attributes, history etc.) and attributes from the data object model that are of interest for these dimensions and underlying quality goals.

Review metrics: after metrics are created by the data quality expert, they need to be reviewed. Data experts ensure that the defined metrics based on rules are valid and reflect the rule violation. Data consumers ensure that the defined metrics accurately reflect their goals, requirements and experienced problems. Criteria for metrics are defined by RUMBA; metrics should be Reasonable, Understandable, Measurable, Believable and Achievable (see Kovac et al., 1997) for developing RUMBA data quality metrics).

Assign weights to metrics: data consumers and data experts assign weights to the metrics to indicate how well they reflect the intended dimensions.

Conduct Questionnaire: obtain the experience of the data consumers by obtaining their answers to the questionnaire items. An agree-disagree Likert scale can be used to measure data consumers experience. Besides measuring the questionnaire items, the questionnaire is used to obtain the perceived importance of the identified dimensions (for example using a 100-dollar test). These can be translated into dimensions weights which are needed to obtain a final data quality score.

Obtain objective measures: obtain the defined objective metrics. Depending on the metrics, this may include software computations, obtaining reference data, or creating/collecting metadata.

Analysis: Combine the results of the questionnaire and the objective measures. Average the answers of the questionnaire items and translate to a score that is comparable to the objective measures. Using the metric weights and dimensions weights, obtain final dimension scores and a final data quality score.

Reporting: report the scores of data quality dimensions and final data quality score.

**Data objects**

Business process model(s): BPMN presentations of the business processes that create, modify or use the data to be assessed.

Data object model: A UML Class diagram presenting the data objects, their attributes and relations.

Activity mapping: A description of the relation between each activity in the process model and each data object/attribute in the data object model, explaining how data is created, modified or consumed.

Stakeholders: a description of the identified stakeholders for the given context.

Scope: Consists of the business process models, data object model and stakeholder descriptions.

DQA role definitions: the definition of the roles that are to be assigned in a data quality assessment, as identified by this research in section 4.5.2

Data rules: Rules that result from the definition of data objects, based on referential integrity, functional dependencies and attribute analysis.

Data quality goals: Description of the goals of the tasks of data consumers (i.e. what do we do with this data?), identified by the interviews

Perceived data issues: Description of a set of perceived data issues by data consumers, related to the quality of this data. Identified by the interviews.

Questionnaire items: Questions to be asked in a questionnaire that can be rated on a 0-10 scale. Created for subjectively measuring dimensions.

Objective measures: Definition of calculations to be performed on the data to be assessed that provides a metric for the given dimensions / perceived quality problems and rules.

Data quality dimensions: the identified dimensions that follow from the data quality goals, perceived quality issues, and the data rules.

Data quality metric: the combination of both the questionnaire items and objective metrics.

Subjective measurement: The results of the questionnaire items, filled out by the data consumers.

Objective measurement: The results of performing the calculations of the objective metrics.

Metric weights: the importance of metrics (scaled between 0 and 1) for a dimension, rated by data consumers.

Dimensions weights: the importance of dimensions (scaled between 0 and 1), rated by data consumers in the questionnaire.

Data quality measurement: The combination of subjective and objective measurement results and their weights, dimensions scores, and data quality score.

## Appendix VI: Organizational and departmental background of ASML EUV

ASML (Advanced Semiconductors Manufacturing Lithography) is the global market leader in supplying lithography machines that are critical for the production of microchips. ASML started in 1984 as a joint venture between Philips and ASMI (Advanced Semiconductor Materials International). Currently ASML is active in over 16 countries worldwide and employee over 16.500 people. The net sales over 2017 are 9 billion euro, with a profit of 2.1 billion euro (ASML, 2017).

Currently, AMSL produces two system generations: the DUV systems and EUV systems. The latter is the newest system, applying Extreme Ultraviolet (EUV) Litography. To produce these machines, ASML has two large cleanrooms in Veldhoven; the Twinscan Factory (TF) and the EUV factory (EF). The EUV factory is supported by a variety of departments. One of these departments is Facility Management (see Figure 10.1). The case study will be conducted within EF Business Engineering (BE), which is part of Facility Management. Currently, this BE team is largely occupied with cycle time reduction projects, as they wish to speed up the production of EUV machines.

Extreme Ultraviolet Lithography is a promising innovation that is being introduced for volume chip manufacturing. The small wavelength of EUV (13.6 nanometer light) enables chipmakers to create smaller structures on a chip, and thus provide more functionality on a smaller surface. This makes chips faster and more powerful. The challenge of EUV light is that it is absorbed by everything, including air. Therefore, an EUV system contains a large high-vacuum chamber in which the light can travel far enough to land on the wafer (a circular slice of semi-conducting material that serves as a substrate for the manufacturing of chips). Recently, EUV systems are applied for mass production by ASML's customers (see for example https://news.samsung.com/global/samsung-electronics-starts-production-of-euv-based-7nm-lpp-process). The EUV Factory of ASML in Veldhoven is where these EUV machine are made. An EUV system contains over 100,000 parts, 3,000 cables, 40,000 bolts and 2 kilometers of hosing. The production of these machines is highly complex: it consists of many steps, dependencies between these steps and many exceptions and deviations from standard procedures, frequent job reworks, machine failures, repairs and tests and other uncertainties induced by the complexity of the end product.

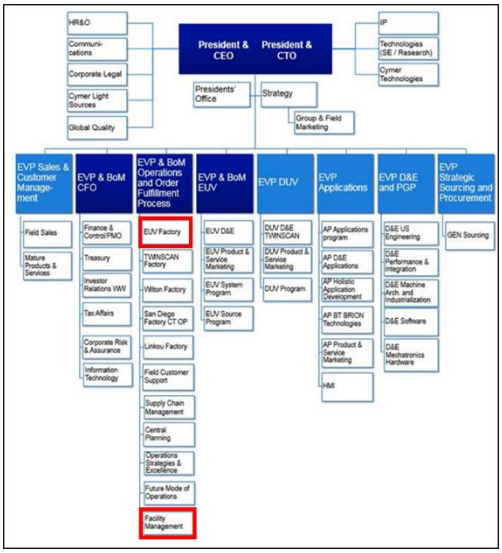
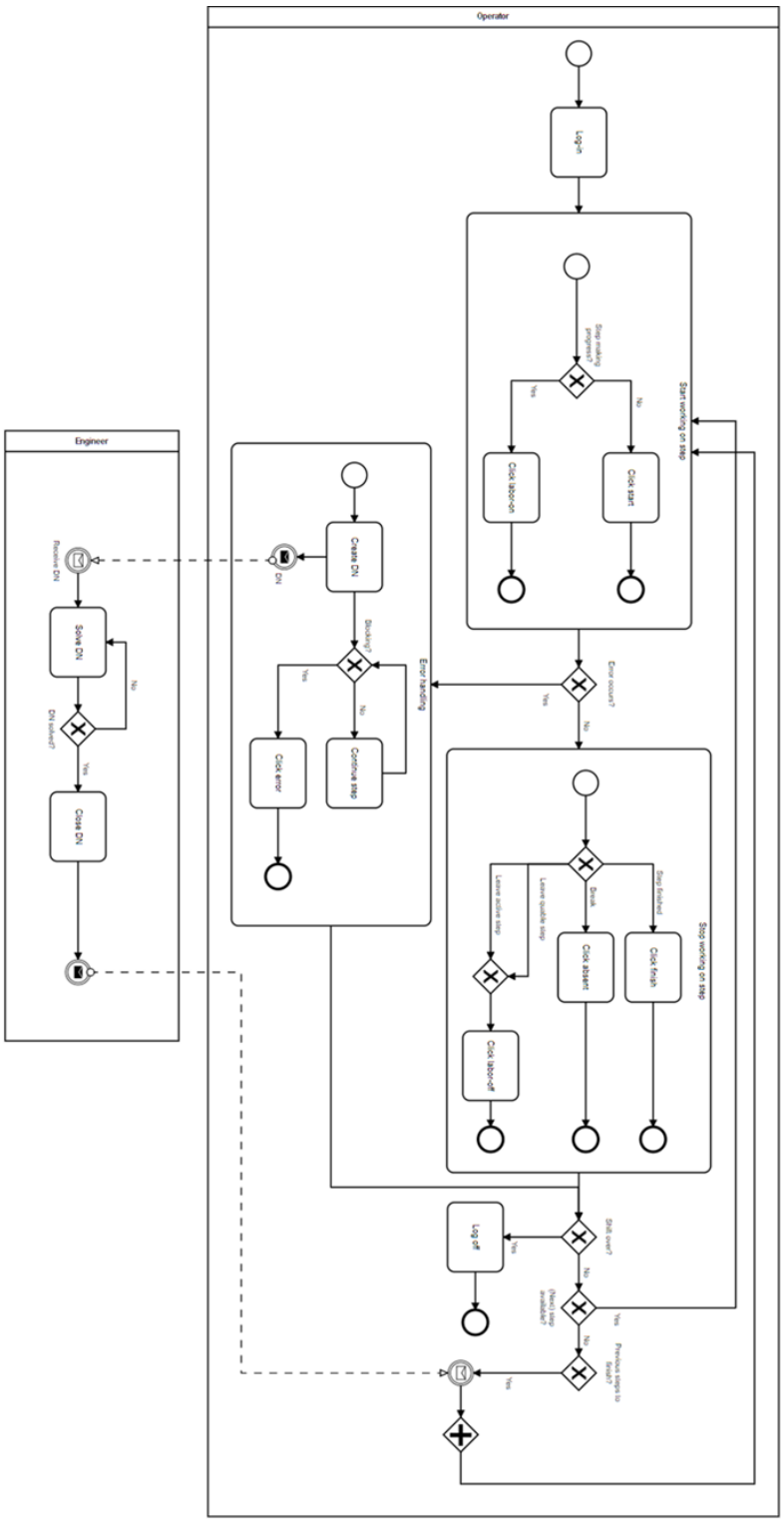


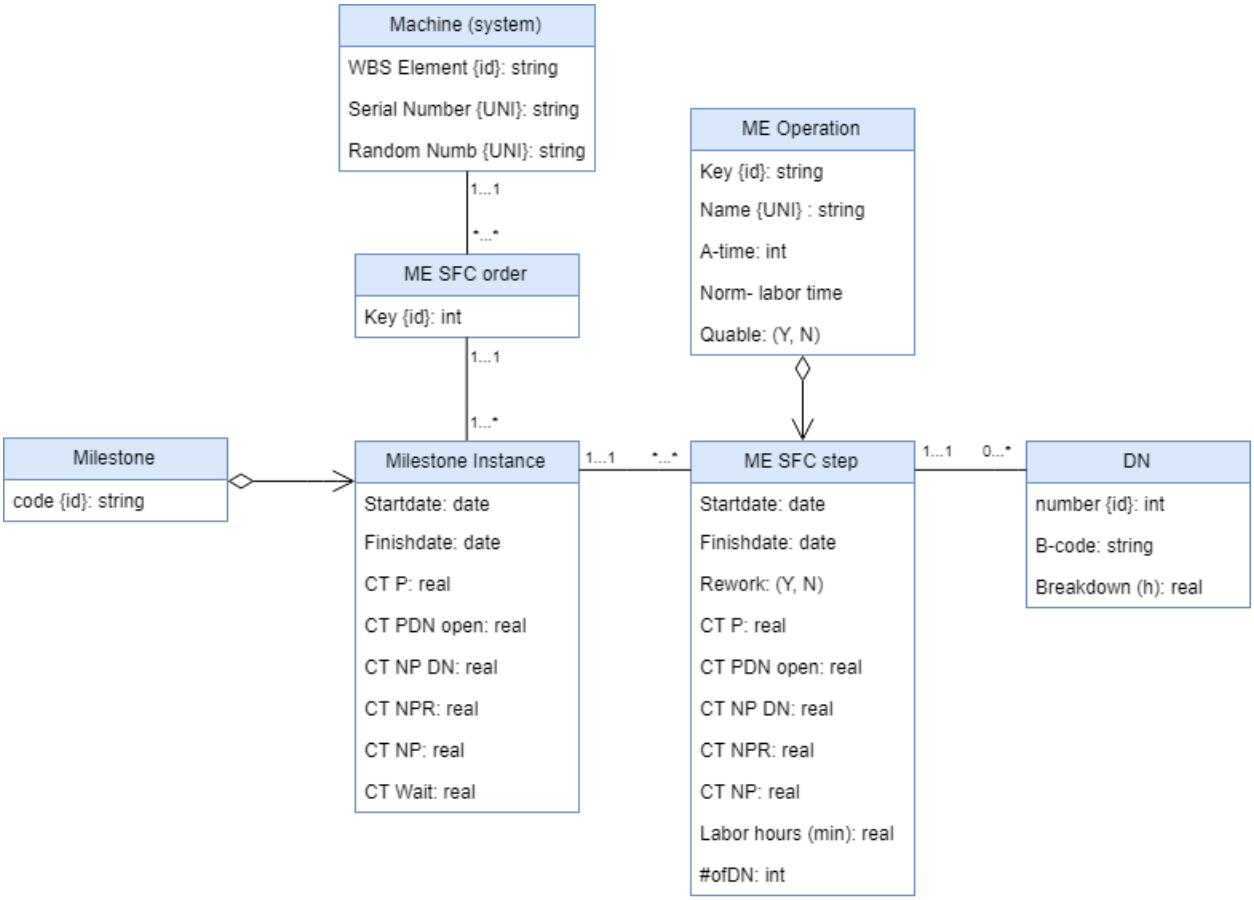

*Figure 10.1: Organization overview*

## Appendix VII: Case Study Results of the process model

1. Operator process



2. UML Class diagram



3. Activity Mapping

| Activity | Data object | Attribute | Result |
|---|---|---|---|
| Click Start | ME SFC Step | CT P | Start measuring time for CT P, stop measuring for other CT categories |
| Click Start | ME SFC Step | CT PDN | Start measuring time for CT PDN (if open DN), stop measuring for other CT categories |
| Click Start | ME SFC Step | CT NPR | Start measuring time for CT NPR if step is rework |
| Click Start | ME SFC Step | Labor hours | Start adding labor time for a single operator |
| Click Start | ME SFC Step | Start date | Record start date for the started step |
| Click labor on | ME SFC Step | Labor hours | Start adding labor time for (another) single operator |
| Create DN | ME SFC Step | #ofDN | Add 1 to the number of DN's |
| Create DN | ME SFC Step | CT PDN | Start measuring CT PDN, stop measuring for other CT categories |
| Create DN | DN | Number | Create DN number |
| Create DN | DN | B-code | Assign B-code |
| Click error | ME SFC Step | CT NPDN | Start measuring time for CT NPDN, stop measuring for other CT categories |
| Click error | DN | Breakdown (h) | Start measuring breakdown time |
| Close DN | DN | Breadown (h) | Stop adding breakdown time |
| Click Finish | ME SFC Step | CT P, CT PDN | Stop measuring CT P or CT PDN |
| Click Finish | ME SFC Step | Labor hours | Stop measuring labor hours for all active operators |
| Click Absent | ME SFC Step | Labor hours | Stop adding labor time for single operator to this step |
| Click Absent | ME SFC Step | CT NP | If only 1 operator active and step is not quable, start measuring CT NP |
| Click labor off | ME SFC Step | Labor hours | Stop adding labor time for single operator to this step |
| Click labor off | ME SFC Step | CT NP | If only 1 operator active and step is not quable, start measuring CT NP |

4. Goal identification from interviews

| Goal | Times mentioned | Data requirement(s) | Related Dimension |
|---|---|---|---|
| Accurately predict milestone cycle times for accurate planning | 3 | Accurate milestone cycle times, complete DN data | Accuracy, Completeness |
| Identify problematic milestones through comparison with targets | 5 | Accurate milestone cycle times | Accuracy |
| Identify bottlenecks in milestones that do not meet targets | 4 | Accurate step cycle times, complete DN data | Accuracy, Completeness |
| Analyze required labor per step and per milestone | 3 | Accurate step loggings | Accuracy |
| Obtain transparency in time spend by operators | 6 | Comprehensive labor data | Comprehensiveness |

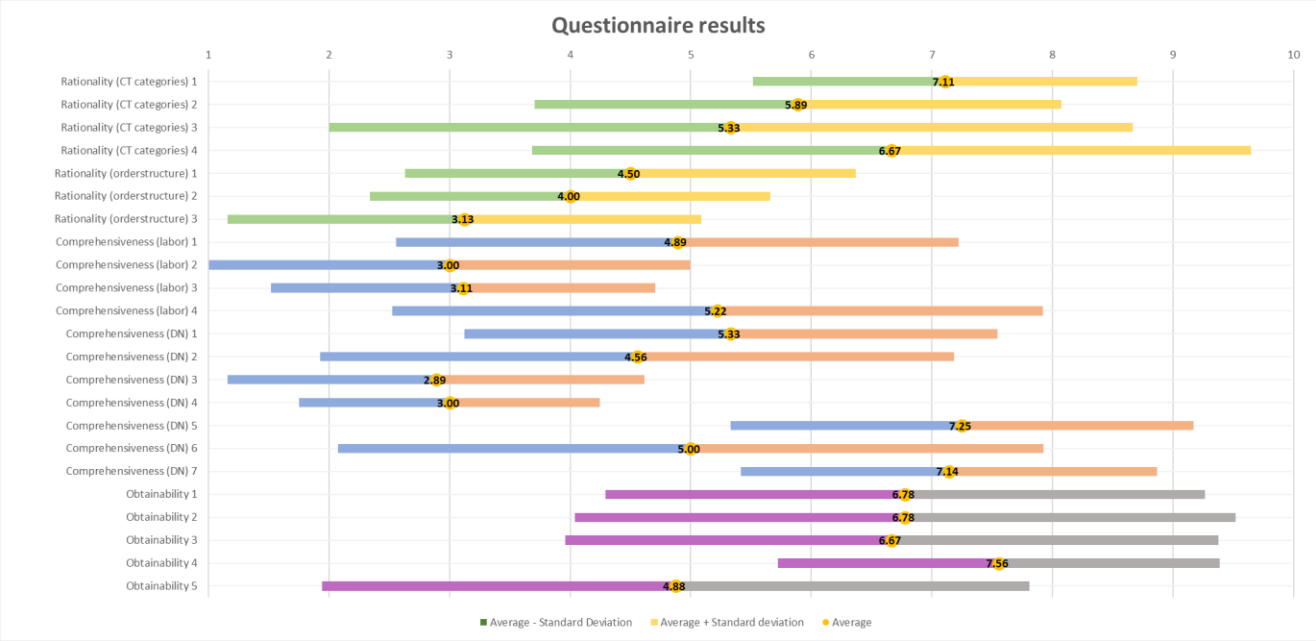
## 5. Problem identification from interviews

| Problem | Times mentioned |
|---|---|
| Start Stop logging | 7 |
| Cycle time categories are not logical | 2 |
| Orderstructure is not logical | 4 |
| SAP ME system down causes data loss | 4 |
| DN's are not logged | 3 |
| Steps are not closed | 2 |
| Milestones have ambiguous names | 3 |
| Operator idle time is not measured | 3 |
| Changes in DN priority over time is not recorded | 2 |
| Labor spent on solving DN's is not recorded | 2 |
| Not active enough in changing the sequence | 2 |
| Logging on step-level is highly time consuming | 5 |
| Measured labor differs greatly from theoretical measure (CT*MMR), a lot of hours are missing | 3 |

## 6. Final set of dimensions and metrics

| Dimensions | Dimension weight | Metric | Metric Weight | Result |
|---|---|---|---|---|
| **Integrity** | 0.08 | I1 | | 0,94 |
| | | I2 | 0,3 | 1,00 |
| | | I3 | | 1,00 |
| | | I4 | | 1,00 |
| | | I5 | 0,3 | 1,00 |
| | | I6 | | 1,00 |
| | | I7 | 0,1 | 1,00 |
| | | I8 | 0,3 | 0,88 |
| | | I9 | | 1,00 |
| | | I10 | | 1,00 |
| | | I11 | | 1,00 |
| **Consistency** | 0.18 | C1 | | 1,00 |
| | | C2 | 0,4 | 0,87 |
| | | C3 | 0,2 | 0,96 |
| | | C4 | 0,2 | 0,97 |
| | | C5 | | 1,00 |
| | | C6 | 0,2 | 0,94 |
| | | C7 | | 1,00 |
| **Validity** | 0.06 | V1 | 0,2 | 1,00 |
| | | V2 | | 0,83 |
| | | V3 | 0,4 | 1,00 |
| | | V4 | 0,2 | 0,46 |
| | | V5 | 0,2 | 0,82 |
| **Accuracy** | 0.24 | A1 | 0,6 | 0,77 |
| | | A2 | 0,2 | 0,98 |
| | | A3 | 0,1 | 0,83 |
| | | A4 | 0,1 | 0,32 |
| **Completeness** | 0.13 | Comp1 | 0,5 | 0,99 |
| | | Comp2 | 0,5 | 0,59 |
| **Rationality** | 0.06 | Questionnaire items: 4 questions | 0,5 | 0,63 |
| | | Questionnaire items: 3 questions | 0,5 | 0,39 |
| **Comprehensiveness** | 0.12 | Questionnaire items: 4 questions | 0,25 | 0,41 |
| | | Questionnaire items: 2 questions | 0,25 | 0,49 |
| | | Questionnaire items: 2 questions | 0,25 | 0,29 |
| | | Questionnaire items: 3 questions | 0,25 | 0,65 |
| **Obtainability** | 0.13 | Questionnaire items: 5 questions | 1 | 0,65 |

7. Questionnaire results




8. Final Data quality score

| Final DQ Score | | | |
|---|---|---|---|
| Dimension | Weight | Score | Final Score |
| Integrity | 0.08 | 96.45% | **75.32%** |
| Consistency | 0.18 | 92.20% | |
| Validity | 0.06 | 85.46% | |
| Accuracy | 0.24 | 77.29% | |
| (Record) Completeness | 0.13 | 79.06% | |
| Rationality | 0.06 | 50.63% | |
| Comprehensiveness | 0.12 | 46.02% | |
| Obtainability | 0.13 | 65.31% | |