Towards understanding and managing traceability data quality in high-tech and cyber physical systems

van Duin, B.

Award date:
2020

Master Thesis

Towards Understanding and Managing Traceability Data Quality in High-Tech and Cyber Physical Systems

A Philips Healthcare Case Study

Master Thesis in the department of Industrial Engineering and Innovation Sciences

Author:

B. van Duin

University Supervisors:

1st – dr. M. Razavian 2nd – dr. ir. R. Dijkman

Company supervisors:

1st – H. Verkerk 2nd – S. Backer

Abstract

With the technology readily available and the economic and environmental aspects of circular and sustainable economies becoming more important and attractive, material traceability is becoming increasingly important within high-tech and cyber physical systems. Legal obligations due to quality and safety concerns add another reason for companies to improve their traceability data. Measuring the quality of the current traceability data is a challenge due to the complex nature and the impact of the traceability data quality is hard to determine because most cost and benefits are indirect.

This master thesis provides a case study that measures the traceability data quality of an x-ray tube in Image Guided Therapy systems by defining material specific business rules that cover all the possible data problems that impact traceability. The results of the analysis are used to give an estimation of the costs and benefits of the current data quality, and of improving the data quality, by interviewing several data consumers within the company to eventually aid in the decision making process of whether to improve the current data quality or not. The methods and techniques used in this case study are eventually combined in a general framework that is applicable for any company to aid in deciding on investment in improving traceability data quality.

Management Summary

Introduction

Traceability is becoming more popular and more required by manufacturing industries, especially when quality or personal safety are at stake. Despite the many advantages traceability data also has to offer, the quality of the traceability data is not always as desired. The aim of this Master thesis is to determine how to measure and control the current level of traceability data quality at Philips. Based on these results, the costs and benefits of improving the data quality are required in order to aid in the decision making process of whether to improve the traceability data quality. These aims are translated into the following research question:

How can the current quality of installed base traceability data be measured and controlled and what are the costs and benefits of improving the quality?

This question is divided into several sub-questions in order to systematically answer the main research question:

1. *What is the current quality of traceability data?*
2. *What is the impact of the current and improved data quality, and what can be learnt from it?*
3. *How can understanding and managing data quality be guided?*

Literature

The literature section aims to answer the research questions “Which techniques exist for measuring data quality?” and “Which techniques exist for estimating the value of data quality?” The literature is gathered following a structured review protocol which allows reproduction of the results. Data quality can be measured in a subjective and objective aspect. For this study the objective quality of the data is desired. Since all company data is unique to the company, there is not one universal technique for objectively measuring the data quality, although there are general understandings of data quality problems and metrics. For instance completeness, Accuracy and Currency. Estimating the value of traceability data is very difficult due to the indirect nature of the effects of data quality. Therefore a cost-benefit analysis is proposed to indicate the value of the traceability data quality.

Data Quality Problems and Measurement

In this section the research question “What is the current quality of the traceability data at Philips?” is attempted to answer. In order to measure the data quality, metadata about the data structure and the dataflow is required to create a data landscape and eventually set up a measuring system. This is done by following the dataflow and interviewing employees that involved in the processes or knowledgeable of the data and the requirements that need to be met for traceability data. Based on this metadata a set of business rules is derived which are related to the data problems and metrics from literature. For these business rules the data is analyzed to generate an overview of the performance of the data quality per business rule. The different business rules are then combined in two analyses to determine the overall traceability data quality.

After analyzing the results, a new series of interviews is conducted in order to determine the costs and benefits of the found traceability data quality. The costs and benefits are grouped per category to provide a clear overview of the results the current data quality and improved data quality. Giving an indication of the monetary equivalent of the costs and benefits is impossible because of the size of the organization, the indirect nature of the consequences of data quality and the broad range of processes affected by poor quality data.

Framework

Based on the research conducted in this case study, a framework is developed that compares the interviews, the document analysis and the impact analysis with literature. This framework consists of six steps that guide researchers and practitioners in setting up a similar study to eventually contribute to the decision making process on improving the traceability data quality within a company.

Step	Task
1	Create a data landscape
2	Select data quality problems
3	Data quality problem understanding
4	Scoping and document analysis
5	Impact analysis
6	Go / No Go decision

Conclusion

After the introduction of the framework the study is concluded by answering all the research questions and suggesting further research. The framework serves as an answer to the main research question by supplying practitioners and researchers with a tool to structure the analysis of the current data quality and gather the possible costs and benefits of improving the data quality. However, there is one crucial distinction to be made. When a company is obliged by law to have optimal traceability data, the framework still helps in assessing the current data quality. If the quality of that traceability data is not high enough, the impact analysis can be used to determine what other advantages improving the data quality can bring. However, the choice whether to invest in improving the data quality for materials subject to regulation has already been determined by the law since the company is obliged to abide by the law. Not doing so will result in severe consequences by authorities that can block sales ore completely shut down the factory.

Further research should focus on testing and expanding this framework, also adding solutions for analyzing more complex traceability data problems with single and multiple relations in the data and data problems concerning multiple sources. Other topics of interest are the quantification of costs and benefits as the result of poor or improving data quality so the effects are more clear, resulting in more attention for data quality problems and well balanced decision making. Traceability data, especially in high-tech and cyber physical systems is underrepresented in literature, despite being crucial to many modern business models. More literature regarding the complexity and mechanism of traceability data would enhance the understanding of the importance and value of it.

Contents

List of Figures

List of Tables

1. Introduction

This research proposal is part of the thesis that completes the Master Innovation Management at Eindhoven University of Technology (further referred to as TU/e). The thesis is performed during a graduation internship at Philips Medical Systems International B.V. (further referred to as Philips) within the Installed Base Data Program of the Image Guided Therapy department. Aim of this proposal is to define the content of the thesis and project.

Existing literature focusses a lot on measuring and improving data quality in general (Sebastian-Coleman, 2013) , and agri-food or software requirements traceability (Bosona & Gebresenbet, 2013; Wohlrab, Steghofer, Knauss, Maro & Anjorin, 2016). Currently there is little research available on traceability data quality of hardware components or industrial parts, especially within the healthcare domain. The increasing complexity, partly due to regulations that require unique identification for medical devices and medical device components (European Union, 2016), might well be underestimated and therefore the area of traceability data quality in medical devices needs to be addressed. Furthermore, from environmental and economic perspectives, circular and sustainable economies are becoming more important and attractive (Stahel, 2016). Philips for instance, intends to buy back all systems that have reached the end of their life with the purpose of reusing and refurbishing as much as possible. For economic efficiency it is important know what parts are present in the systems that are offered, which requires reliable traceability data.

Guided by the design science methodology by Wieringa (2014), the study progresses through the first three steps of the design cycle: Problem Investigation, Artefact Design and Artefact Validation. Due to time and resource affiliated with Master thesis projects, the fourth step of Artefact Implementation is excluded and might later be executed on Philips' behalf, as a result of, but aside from this thesis.

Next up in section 1.1 is a short description of the company and the problem, followed by the used methodology research the research questions derived from the problem statement in section 1.2. Section 1.3 briefly states the deliverables for this Master thesis project.

1.1 Problem statement

This section provides a short introduction on Philips, starting with a summary of its history that leads to the current situation. The story flows over into section 1.1.2, where the problem of traceability data quality is further explained, mentioning the possible consequences. The scope of the Master thesis project is the determined in section 1.1.3, after which the most important stakeholders are identified to conclude the problem statement section.

1.1.1 Company description

Philips Medical Systems International B.V. originates from Philips & Co which was founded by Frederik and Gerard Philips in Eindhoven, 1891. It started as a light bulb factory grew rapidly through research and acquisitions. Their product range also expanded and Philips became a symbol in Eindhoven and the rest of the Netherlands. 'Philips & Co' became 'Koninklijke Philips' (Royal Philips) and took a prominent position in Dutch and international markets with a wide variety of products. Because of increasing international competition in the mid 70's, Philips began to cut costs and shrink. In 2016 the lighting

division of Philips spun off and changed its name to Signify in 2018 (Signify Holding, 2018), which means Philips now is a true healthcare company divided in four business areas: Precision Diagnosis, Image Guided Therapy, Connected Care and Personal Health

The Philips Image Guided Therapy (IGT) Fixed Systems main location is situated in Best, The Netherlands. Philips is leader in IGT, integrating best in class imaging solutions with specialized diagnostic and therapeutic devices that aid a range of complex interventional procedures. Examples of clinical areas where these systems are used for are spine surgery, oncology, neurology and electro-physiology. Additional information on IGT systems is found in the infographics in Appendix A.

Within the Installed Base (IB) Sustainment department of the IGT systems location in Best, the Installed Base program is concerned with identifying & driving improvements for the installed base IGT systems worldwide. However, there is no general and continuous measurement on the quality of traceability data and the potential costs or benefits. In other words, the question is raised what the current quality of data is, and whether the costs and benefits of improving it are worth investing in. This question needs answering on a continuous basis and as efficient as possible. Ideally the answer, with some modifications, is also applicable across other departments within Philips.

1.1.2 Problem context

Philips IGT systems is the leading innovator in image-guided interventional solutions. Seamlessly integrated systems, including interventional X-ray systems and software solutions enable personalized, minimally invasive treatment decisions for every patient and help guide minimally invasive procedures more effectively. Medical devices are a regulated business by several instances differing per international location, consequently to each of which Philips has to conform. In the USA for one, regulations are maintained by the Food and Drug Administration (FDA), whereas the member states of the European Union adhere to the regulations by the European Commission. Newly announced European Union Medical Device Regulation (EUMDR) applicable for medical devices sold and distributed in EU region, requires among others transparency on traceability through Unique Device Identification (European Union, 2017). Non-compliance to these regulations may result in hefty fines or complete shutdown of company locations and restriction of distributing and sales within the EU until compliance is restored.

Apart from that risk, incorrect or missing data due to maintenance or alterations that have not been logged (correctly) may also raise costs. Customer service for example might not provide the required help directly, if data is incorrect. Or planned maintenance at a hospital that appears to be unnecessary if parts are not registered (correctly), leading to undesirable downtime for the hospital and costs but also damages in customer experience. Employees may lose valuable time if data quality is limited and requires combining data from different sources as part of data validation checks.

Additionally material traceability is applied for business interest, providing information that can be used to optimize operational activities such as maintenance. The derived information can be used in calculating business cases that serve as input for (management) decision making.

Last but not least, high levels of data quality have proven to increase competitive advantage (Redman, 1995) by providing insight and discovering opportunities sooner.

Data quality is measured & controlled in (sub) processes while an End to End measurement is not widely in place. This means that at Philips there is a lack of End to End measurement about the current state of the traceability data quality and, with that, the impact thereof. This research is conducted to measure the traceability data quality and what the impact is on different business units within Philips.

1.1.3 Scope

Philips is divided into four business clusters: Precision Diagnosis, Image Guided Therapy, Connected Care and Personal Health. IGT Systems is part of the Image Guided Therapy cluster and develops integrated systems that include interventional X-ray and software solutions to enable personalized minimally invasive treatment decisions and help guide the procedure more effectively. In this study, an IGT System is often referred to as "system" and is defined as all hardware and software components combined that together make a fully functioning system, intended for performing Image Guided Therapy. This includes all materials needed to mount and secure the system in its place.

Subject for this study is one type of IGT systems; Allura. Every system is provided with a six digit numerical code (6NC) and serial number (SN) where the combination of those is unique to that system and can be seen as a Unique Device Identification (UDI). Not only the system itself, but also some of its parts, need to be traceable via a UDI in order to comply with announced EUMDR regulations. These parts have a twelve digit numerical code (12NC) and SN. Furthermore, the SN's provide the opportunity to trace individual parts through SAP and different databases. Philips has SN traceability for materials in place, which is also a regulatory requirement in the foreseeable future. Another reason for scoping to the traceability parts level is that the study is representative and reproducible for other medical systems within Philips. Software is decided out of scope due to the different nature and registration of it, combined with a running connectivity project that is expected to identify software versions directly through an online connection with the system.

IB traceability data concerns the processes and time period between installation and de-installation of an IGT system. During this period of time, several events are performed to an IGT system to maintain or improve it. These events can be divided into three main categories: maintenance, upgrade and recycle. Within these categories another division van be made as displayed in Figure 1. These events may occur unrelated to each other and with no sequential order, it is merely a classification of event types.

Of the three categories (Maintenance, Upgrade and Recycle), Maintenance is the most important one to address data quality issues, since 100% of the IGT systems are affected by (corrective and/or preventive) maintenance multiple times during their lifecycle. Furthermore, any mistakes in registering the right data in these processes may result in a snowball effect throughout the other proceedings. Whereas careful registrations during maintenance may correct errors made in one of the other categories.

Both Corrective and Preventive Maintenance as well as Field Change Orders are issued by and registered via Service Work Orders, of which the data is directly traceable to the system performed on, providing a general overview of the actions performed on one system concerning the material in question. Therefore the entire Maintenance section is considered as the horizontal scope, hence the dashed framework in Figure 1.

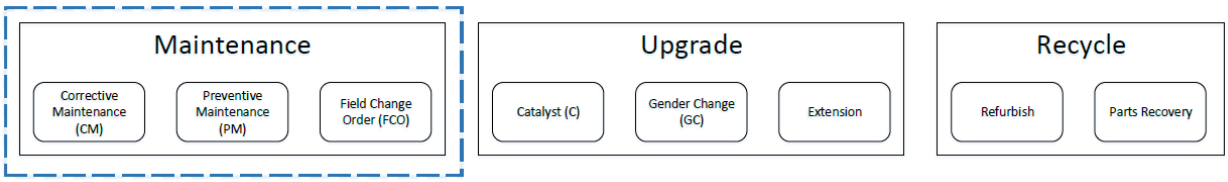


Figure 1 Horizontal scope

All the traceability data of the tubes contains information of the system and the tube combined. Among this data is also a user status that indicates whether it is operational, scrapped, ready to be installed, etc. This user status attribute is available for the material as well as the system. Based on the user status, the scope is further narrowed down to all the systems of which either the system user status and/or the material user status is 'operational'. Reasoning behind this scoping is that according to the data, these are all within our maintained data and operational, and should therefore be traceable. It furthermore eliminates the relevance discussion for systems that are not operational anymore. The consequences of basing the scope for data research on data that is perceived to be of poor quality are treated in the discussion.

1.1.4 Stakeholders

Initially five stakeholders were identified for this study. First stakeholder and problem owner of the project is the IGT Customer Service department, because this research is executed within their domain. The IB program within Customer Service was created to identify & drive improvements for the installed base IGT systems worldwide including assessment when to install material traceability for business interests. Furthermore the data, and thus the project, affect several other departments within Philips. The other departments considered as stakeholders are Quality & Regulatory, Service Parts Supply Chain and the Market Organizations that are responsible for service in specific areas worldwide. Apart from all internal stakeholders within Philips, hospitals are also a significant, though indirect, stakeholder since they are the customers for the IGT systems. All stakeholders are individually listed below Figure 2 with more detailed information about their relation to the project.

Through several interviews with employees from different departments, stakeholders are analyzed on power and interest for the subject of this study and their relation to it. Using the stakeholder matrix (Mendelow, 1981), three key stakeholders are identified by having the highest power and interest, as displayed within the dashed framework in Figure 2.

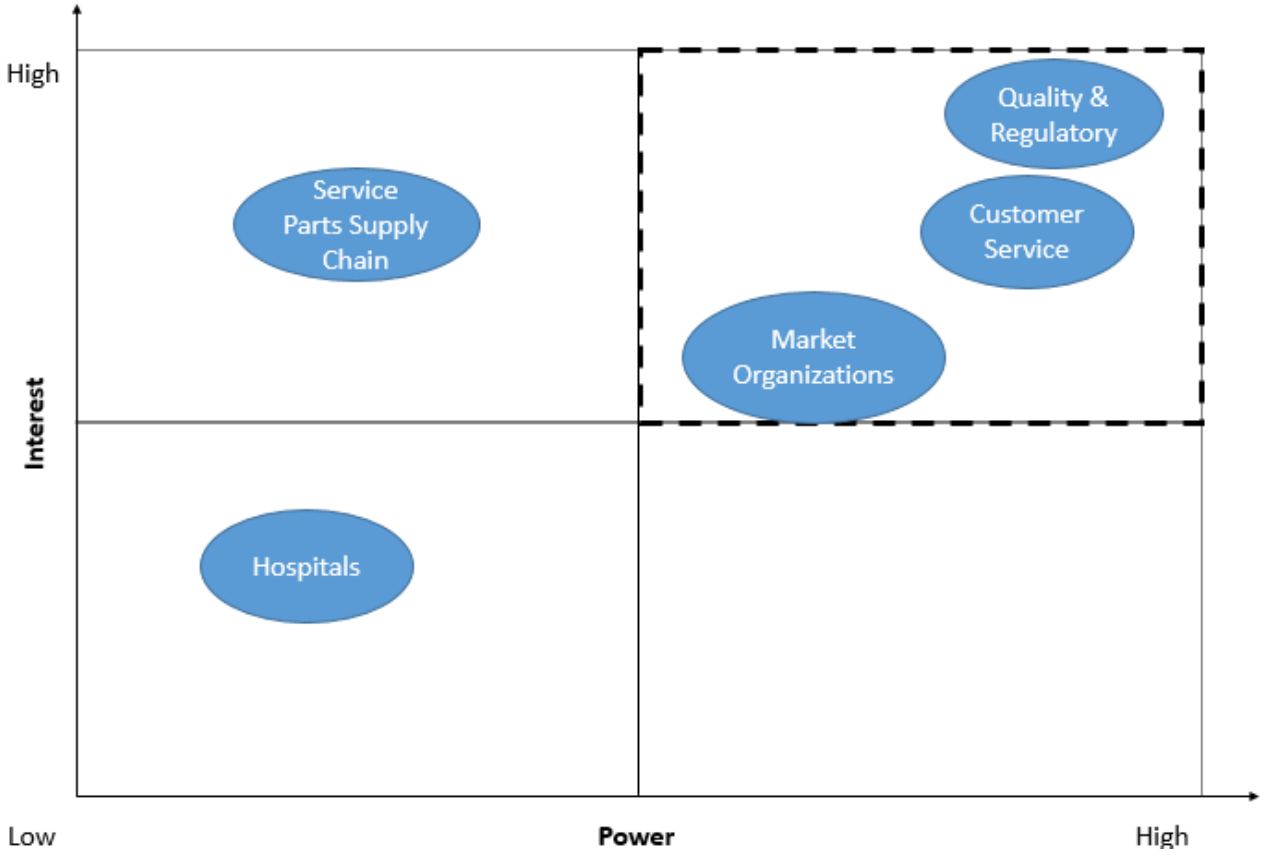


Figure 2 Stakeholder matrix

Customer service – Key stakeholder and problem owner, the Customer Service department, is concerned with, as the name entails, the part of the business that aims to maintain and improve the services for Philips' customers. Within the department, the Installed Base Pogram that includes assessment the use of material traceability for business interests which also contains the traceability data that is topic of this research. Other programs within Customer Service also rely on and use the data in order to provide the best possible service for the customers.

Quality & Regulatory – The second key stakeholder has quality and regulations as their main concerns. They are closely related to the traceability data quality topic due to the already existing and upcoming EUMDR regulations, as it is their job to translate the regulations and come up with what is required by Philips to fully comply. These regulations now require Unique Device Identification (UDI) and that each unique medical device or medical device component can be traced. If the regulations are not met when enforced in May 2020, healthcare companies like Philips risk not being able to sell their medical devices anymore or even have to recall products that are already out in the market (European Union, 2016), which will have serious consequences for the company. Apart from EUMDR, there are several other

authorities worldwide, for instance the Food and Drug Administration (FDA) in the USA, that each have their own regulations regarding medical devices and their traceability.

Market Organization – Market organizations are responsible for certain parts of international market, for example the Benelux, Asia or Latin America. They are closest to the customer and also manage the Field Service Engineers (FSE). These FSE's perform the actual maintenance, installations and FCO's and have the responsibility to initiate the update for exchanged and installed parts in the data. Poor quality data might lead to extra time and can affect first time fix rates due to not being able to prepare optimally for the service event. Additionally correcting traceability data issues in the system cost extra time and effort on top of normal activities.

Service Parts Supply Chain – For many events performed on the systems, spare parts are needed. Service Parts Supply Chain manages, indeed, the supply chain of service parts. Parts need to be exchanged when maintenance or FCO's have to be performed. They manage their supply chain partly based on the IB (trace) data and therefore are codependent on it to perform their tasks well and provide optimal stock solutions. For example, when according to the data 50 machines need the same part exchanged because of a known failure possibility, but the data is incorrect and 20 of them have already been replaced, it could result in overstock.

Hospitals – Hospitals are the customers and therefore very important for any business. Their equipment not working optimally results in treatment of less patients, the same holds for obsolete scheduled maintenance that translate to unnecessary down time. Even without the ethical perspective, every untreated patient is also loss in revenue and possibly affects their relationship with Philips.

1.2 Research design

First, the goal of this research is explained with the corresponding research questions. After that the research method is described and then this chapter is concluded by the outline of the thesis.

1.2.1 Research goal

This study aims to solve a business problem. The problem in question is an operational problem concerning data quality that withholds stakeholders from functioning optimally and possibly impacting the company financially. Below, the design problem is further described by design science methodology guidelines (Wieringa, 2014).

Design problem:

Problem context:	Low quality of material trace data quality has a direct or indirect negative influence on operational processes.
(Re) designed artefact:	Data quality measurement and control tool.
Requirements:	It should be able to identify problems in data quality and indicate the impact thereof.
Stakeholder goals:	Give valuable insights in data quality to improve data control and revenues.

The design problem can be stated as follows in Wieringa's (2014) formatting:

Improve the ability to measure installed base traceability data quality that has a direct or indirect negative influence on operational processes

By designing a data quality measurement and control tool

Such that identifies problems in data quality and indicates their impact

In order to give valuable insights in data quality on how to improve data control and revenues.

Based on this statement and the design problem, the following research question is formed:

"How can we guide understanding and managing data quality?"

In order to answer the research question, it is divided into several sub questions as formulated below:

What is the current quality of traceability data?

What is the impact of the current and improved traceability data quality, and what can be learnt from it?

1.2.2 Research method

As previously mentioned in the introduction, this research follows the design science methodology without the artifact implementation. The artifact implementation part is out of scope for this Master thesis as visualized in Figure 3. The remaining three steps of the design cycle are described below. The project will continuously cycle through these steps in order to successfully develop the final artifact.

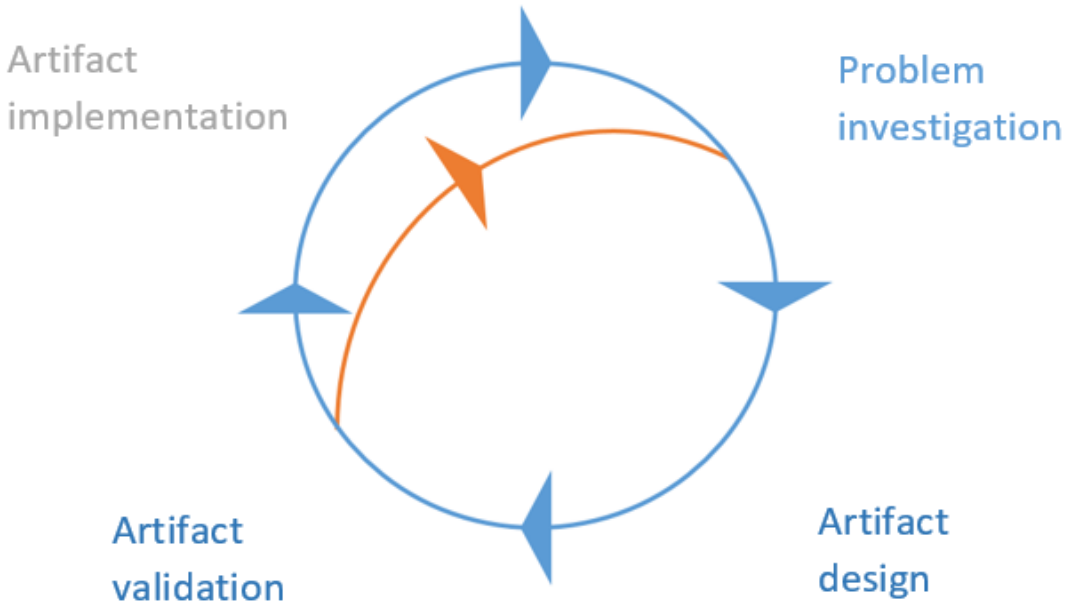


Figure 3 Design science cycle for this Master thesis (Wieringa, 2014).

1.2.2.1 *Problem investigation*

The first phase of the design cycle is the problem investigation, which kicked off with a systematic literature review on the topic of data quality and the issues and challenges that come with it. The goal of the literature review is to generate a theoretical background where the thesis can be built on. The literature review attempts to answer the first section of sub questions that belong to the overall research question. Further investigation of the problem is performed at Philips by having semi-structured interviews with multiple people from different positions that are closely related to the data and the supposed quality issues. During these interviews, scenarios are outlined wherein the data quality errors are discussed. The participant will be questioned about the data quality and errors he or she is familiar with and through story telling the causes and effects of the different traceability data issues are discovered. Next to the interviews, data analysis is performed to evaluate the actual data quality. Guided by the found data quality, interviews are again performed to determine the consequences of the found data quality.

1.2.2.2 Artifact design

Based on the information from the interviews and the traceability data, data quality measurement criteria are developed. Furthermore, evaluating interview information is compared with the data in order to evaluate the effects of possible errors and do a cost benefit analysis by analyzing the possible costs related to faulty data versus the costs of addressing these errors compared to the benefits it could bring. Based on the measurement criteria a tool is developed to continuously measure the data quality from which estimates can be made what financial and operational effects the current traceability data quality has and what impact improvement of the data quality has. It is an iterative process that requires a step by step approach to ultimately design a detailed tool that outputs concrete information.

1.2.2.3 Artifact validation

During the iterations of the process, all measures and estimations are performed by using samples of the data to validate the tool's effectiveness. It is assessed whether the tool correctly identifies the data quality issues and estimates the consequences appropriately by demonstrating to, and discussing with, several professionals from the key stakeholder groups. Afterwards a focus group is created for a final overall validation of the tool.

2. Literature Review

Although literature specifically covering material traceability data concerning medical devices is scarce to non-existent, a considerable amount has been published on data quality in general and the possible effects it has on businesses. Different definitions of data quality can be found in literature (Arts, de Keizer & Scheffer, 2002; Watts, Shankaranarayanan & Even, 2009; Sadiq, 2013), but ISO defines data quality as the "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions" (ISO/IEC 25012:2008) which corresponds with the more common and in research widely accepted term "fitness for use", used by Wang & Strong (1996) where they take the data consumer's viewpoint for assessing data quality: "data that are fit for use by data consumers".

As the expression "from the data customer's viewpoint" already indicates, the way data quality is measured differs with the purpose and viewpoint of the study. In this literature review, an insight is gained into the different ways data quality is measured by answering the following research question:

Which techniques exist for measuring data quality?

Measuring data quality is done by measuring the amount of problems or issues are found in the data. A data quality problem can be seen as something that impacts the previously defined "fitness for use" in a negative way. So before the above research question is answered, data quality problems are first further explained and later linked to the different measurement techniques. A selection will be made for the techniques used in the succeeding parts of this study.

After doing a successful assessment of the data quality, one could wonder "so what?" which is a valid question. Would a lower level of data quality really affect my business and if so, is that impact big enough to justify addressing those issues. In other words, what is data quality actually worth? That will be the second research question addressed in this literature review:

Which techniques exist for estimating the value of data quality?

Both research questions are answered in this literature review by gathering information from existing literature. These scientific pieces of literature are collected using a systematic literature review protocol that carefully explains every step of the information gathering process, providing validation of the review by allowing it to be replicated.

2.1 Review Protocol

In order to find literature that answers these questions, keywords and relevant terms are identified before and during the literature study. To ensure not to miss anything, synonyms for the search terms are also used as input. These search terms are combined in logical combinations and are entered as strings (" …") and/or combined with commands like AND, OR, etc. The keywords, search terms and synonyms are displayed in the table below:

Table 1 Literature Study Search Terms

Index	Keyword	Synonyms and variants	Source
1	traceability		n.a.
2	information	data, intelligence, knowledge	Thesaurus.com
	healthcare		n.a.
	healthtech		n.a.
	medical devices		n.a.
3	automotive		n.a.
4	aerospace		n.a.
5	aviation		n.a.
6	assess	assessment, assessing, estimate, determine, appraise, evaluate	Thesaurus.com
7	measure	measuring	Thesaurus.com, L. Sebastian-Coleman (2013),
	software		n.a.
	electronics		n.a.
	components		n.a.
	hardware		n.a.
8	spare parts		n.a.
9	framework		L. Sebastian-Coleman (2013)
10	incorrect	erroneous, false, faulty, flawed, inaccurate, unreliable	Thesaurus.com, L. Sebastian-Coleman (2013),
11	costs	loss, expenditure	Thesaurus.com
12	value	price, profit, costs	Thesaurus.com
13	profit	benefit, earnings, income, proceeds	Thesaurus.com
14	effect	Impact, consequence,	Thesaurus.com
15	cost benefit analysis		Eppler, M., & Helfert, M. (2004)
16	problem	Issue, complication	Thesaurus.com

These search terms are used in several databases, of which Google Scholar is the main choice due to its connection to almost all other databases and objectivity. Other databases used are the Eindhoven University of Technology library, Scopus, PiCarta and JSTOR. After finding an initial set of approximately 10 relevant articles or books, the snowballing technique (Wohlin, 2014) is used to further explore related articles that are referenced within the source or in which the source is cited. Snowballing is an efficient technique to efficiently gather related literature, by digging through the citations and reference of the already discovered books and articles. This way related articles are easily found and renowned researchers can be identified that have multiple publications in the field of research. Snowballing is an iterative process, used again for every new article found, to make sure sufficient literature on the topic is covered and the most important sources are identified.

In order to decide what literature is suited for the study, all search hits are initially filtered on title and abstract to see if they are relevant to the topic and possibly able to answer (one of) the research questions. The source must be written in either Dutch or English, available online or as hardcopy in the TU/e library. Furthermore the research domain must be comparable or have comparable elements / characteristics with data, traceability and Philips's medical systems, for example traceability in general, industrial parts, automotive, aerospace, software or even agri-food. To ensure they are not outdated, sources older than 25 years are not included in the literature study, unless they are related to fundamental understandings of data quality. Papers are apart from relevance also selected on credibility, by comparing the journals they are published in and the number of citations. Books are also evaluated and both are assessed by the number of citations in order to see which are the main sources of literature in their specific fields. The selection of literature used for this review is provided in

Table 2.

Table 2 List of literature for review

Paper / Book	Year	Author(s)	Publication Outlet
Defining and improving data quality in medical registries: A literature review, case study, and generic framework	2002	Arts, de Keizer & Scheffer	Journal of the American Medical Informatics Association
Data quality assessment in context: A cognitive perspective	2009	Watts, Shankaranarayanan & Even.	Decision Support Systems
Handbook of data quality: Research and practice	2013	Sadiq	Springer
Criticality of data quality as exemplified in two disasters	2001	Fisher & Kingma	Information & Management
A data preparation framework based on a multidatabase language	2001	Sattler & Schallehn	Proceedings of the International Database Engineering and Applications Symposium, IDEAS
Data and Information Quality	2016	Batini & Scannapieco	Springer
Towards Quantifying Data Quality Costs	2003	Won & Choi	Journal of Object Technology
The impact of poor data quality on the typical enterprise	1998	Redman	Communications of the ACM
Data Quality Assessment	2002	Pipino, Lee & Yang	Communications of the ACM
Declarative Data Cleaning : Language, Model, and Algorithms	2001	Galhardas, Florescu, Shasha, Simon & Saita	HAL-Inria
A formal definition of data quality problems	2005	Oliveira, Rodrigues & Henriques	Proceedings of the 2005 International Conference on Information Quality
A Taxonomy of Dirty Data	2003	Kim, Choi, Hong, Kim & Lee	Data Mining and Knowledge Discovery
Problems, Methods, and Challenges in Comprehensive Data Cleansing	2003	Müller & Freytag	Humboldt-Universität zu Berlin
AIMQ: a methodology for information quality assessment	2002	Lee, Strong, Kahn & Wang	Information & Management
Overview and Framework for Data and Information Quality Research	2009	Madnick, Wang, Lee & Zhu	Journal of Data and Information Quality
Measuring Data Quality for Ongoing Improvement	2013	Sebastian-Coleman	Morgan Kaufmann
A classification and analysis of data quality costs	2004	Eppler & Helfert	Proceedings of the Ninth International Conference on Information Quality
A Survey on Data Quality: Classifying Poor Data	2015	Laranjeiro, Soydemir & Bernardino	2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing
The costs of poor data quality	2011	Haug, Zachariassen & van Liempd	Journal of Industrial Engineering and Management

Data Quality for the Information Age	1998	Redman	Communications of the ACM
The Story of Information Sprawl	2012	Glazer & Henry	Gartner
Tailoring Traceability Information to Business Needs	2006	Arkley & Riddle	Proceedings of the IEEE International Requirements Engineering Conference
Motivation matters in the traceability trenches	2009	Mader, Gotel & Philippow	Proceedings of the IEEE International Conference on Requirements Engineering
The Barriers to Traceability and their Potential Solutions: Towards a Reference Framework	2012	Regan, McCaffery, McDaid & Flood	Proceedings of the 8th Euromicro Conference on Software Engineering and Advanced Applications
Collaborative Traceability Management: Challenges and Opportunities	2016	Wohlrab, Steghofer, Knauss, Maro & Anjorin	Proceedings of the IEEE 24th International Requirements Engineering Conference
Perspectives on traceability in food manufacture	1998	Moe	Trends in Food Science & Technology
Assessing the value and role of seafood traceability from an entire value-chain perspective	2015	Sterling, Gooch, Dent, Marenick, Miller & Sylvia	Comprehensive Reviews in Food Science and Food Safety
Food traceability as an integral part of logistics management in food and agricultural supply chain	2013	Bosona & Gebresenbet	Food Control
Enterprise Knowledge Management: The Data Quality Approach	2001	Loshin	Morgan Kaufmann
Executing Data Quality Projects	2008	McGilvray	Morgan Kaufmann

The relevant information from the books and papers is extracted and summarized by the core elements of the research questions, as demonstrated in the example in Figure 4. This provides an overview of the available information and identifies information that is still missing or required. The summarization of information is not only used for the literature review, but can be consulted throughout the entire thesis.

Research Question	Core element	information
Which techniques exist for estimating the value of accurate data in operational processes?	Data value	- we define "data quality" as data that are fit for use by data consumers. (Wang & Strong, 1996)
		- This article would be enhanced with an estimate of the total cost of poor data quality, but studies to produce such estimates have proven difficult to perform. I am aware of three proprietary studies that yielded estimates in the 8–12% of revenue range. More informally, 40–60% of a service organization's expense may be consumed as a result of poor data. These ranges are "good working estimates" of the cost of poor data quality. (Redman, 1998-a)
		- Value of data is maximized by people, in completely understanding the data they are using (Sebastian-Coleman, 2013).
		- six ways in which poor quality data can affect an organization's financial performance (Redman, 1998-b)
		- Operational, Tactical and Strategic impact. (Redman, 1998-a)
		- Direct versus hidden costs and costs on operational or strategic level. (Haug, Zachariassen & van Liempd, 2011)
		- Even when data quality is high and well protected and loss is impossible, the value is zero if the business user can't access the data to make decisions. (Glazer & Henry, 2012)
	How to measure/estimate	- This article would be enhanced with an estimate of the total cost of poor data quality, but studies to produce such estimates have proven difficult to perform. I am aware of three proprietary studies that yielded estimates in the 8–12% of revenue range. More informally, 40–60% of a service organization's expense may be consumed as a result of poor data. These ranges are "good working estimates" of the cost of poor data quality. (Redman, 1998-a)
		- Quantifying the measurement of data quality by means of a cost-benefit analysis. This analysis is done by relating the impacts of high versus low levels of data quality into eight categories (Loshin 2001)
		- These measures need to be evaluated for all three domains as defined before by Redman (1998-a): Strategic, operational and tactical (Loshin, 2001)
		- Describe eight steps to identify business impact, ranging from least time consuming and complex (McGilvray,)

Figure 4 Example of information summarization and extraction table

2.2 Importance of Data Quality

First of all, the understanding of the importance of data quality is essential. As illustrated in the research by Fisher and Kingma (2001), data quality is one of the problems organizations face at some point in time. Although the research is published almost twenty years ago, the extent to which management makes data-based decisions only increased over the years. The considerably extreme examples in Fisher and Kingma's (2001) article of the exploding space shuttle and the shot down passenger plane demonstrate the criticality of data quality. Although data quality is not the only factor responsible for the disasters, it sure played an important role in the run up and the decision-making processes by basing decisions on flawed data. The quality of the input data strongly influences the analysis results and the decisions based on that analysis, which is known as the "garbage in, garbage out" principle (Sattler & Schallehn, 2001).

Apart from these extraordinary risks paired with data quality issues, several studies emphasize on the consequences experienced of poor data quality, often without making the explicit connection to their causes (Batini & Scannapieco, 2016). Kim & Choi (2003) identify seven types of damage/costs for organizations: Loss of revenue, Waste of money, Lost opportunity, Tarnished image, Invasion of privacy and civil liberties, Personal injury and death of people and Lawsuits. Research carried out by Redman (1998-a) identifies several typical data issues and the impacts thereof on operational, tactical and strategic level. Typical data issues are inaccurate data (1-5% of the data fields erred), inconsistencies across databases and necessary data being unavailable for certain operations or decisions. Operational impacts as a result of these issues are lowered customer satisfaction and lowered employee satisfaction. For the few, carefully studied cases, the increased costs due to these data issues were an 8-12% of revenue, while for service organizations the costs increased amounted 40-60% of expenses. Apart from these operational impacts, there were also several tactical impacts. Decisions were poorer and took longer to make, data warehouses were more difficult to implement and reengineering became more difficult. Overall there was an increase in organizational mistrust. The impact on the strategic aspect of the organization is that it is more difficult to set and execute strategy, while compromising the ability to align the organization. Furthermore it contributes to issues of data ownership and it diverts the attention of management.

Although data quality issues are as old as data itself, nowadays data quality affects business strategies through Business Intelligence systems, raising the stakes in data quality for any organization (Sadiq, 2013). Sadiq further explains that data quality issues and management challenges relate to the increasing volumes of data produced by organizations, while diversity of data increases as well as the number of sources by using various technologies.

2.3 Data Quality Problems

Data quality is studied in both the management and database community. Where management studies it from several aspects like relevancy, objectivity, accessibility, etc. (Pipino, Lee & Wang, 2002), the database research community studies data quality from a technical point of view (Galhardas et al., 2001). Because of the technical nature of the data and the identified issued at Philips, this research is in the perspective of the database community, focusing on the quality of data values, or instances. Overall, very few studies explicitly identify and define data quality problems, while they are often implicitly included in the data quality metrics used for the measurement of data quality. Therefore the data quality

problems found in two studies are explained here, while the metrics will be explained in the next section. Oliveira, Rodrigues & Henriques (2005) describe several data quality problems over different levels. The first level is the level of attribute and tuple, which can be divided into three groups: single attribute and single tuple, single attribute of multiple tuples (known as a column), or multiple attributes of a single tuple (known as a row). The next level is that of a single relation where multiple attributes of multiple tuples are analyzed of a relation, followed by the level of multiple relation, concluding with the level of multiple data sources. The data problems in this research are limited to the first level, concerning the attribute or tuple. The data problems in the study of Oliveira, Henriques & Rodrigues (2005) are supported by "A Taxonomy of Dirty Data" of (Kim, Choi, Hong, Kim, & Lee, 2003) and report the following data quality problems on the attribute/tuple level:

Data quality problems concerning a single attribute of a single tuple are:

Missing value – There is no value in the attribute, which is not optional.

Syntax violation – For example the date is in the wrong format.

Incorrect value – The updated value is incorrect.

Domain violation – The value is not within the domain, for example the quantity is a negative number.

Violation of business domain constraint – Constraints imposed by business rules, for instance minimum or maximum number of characters input.

Invalid substring – Part of the string value is invalid for that attribute.

Misspelling error – Spelling errors might occur in name or location attributes.

Imprecise value – Values are imprecise and can be interpreted in multiple ways.

Data quality problems that concern a single attribute and multiple tuples (column)

Unique value violation – The same value occurs for different tuples while not allowed.

Existence of synonyms – Different words used for the same intended value.

Violation of business domain constraint - rules as applied by the organization are not adhered to, for instance wrong ordering of the data.

Data problems that concern multiple attributes of a single tuple:

Semi-empty tuple – The defined amount of maximum empty attributes for one tuple is violated

Violation of functional dependency – If the value of one tuple's attribute is functionally dependent of another attribute, but this dependency is violated.

Other research by (Müller & Freytag, 2003) classifies data problems, or data anomalies as they call it, in three different types: Syntactical Anomalies, Semantic Anomalies and Coverage Anomalies. Within the syntactical anomalies they define lexical errors, domain format errors and irregularities. The semantic anomalies consist of integrity constraint violations, contradictions, duplicates and invalid tuples. Where missing values and missing tuples belong to the coverage anomalies. Of these problems, duplicates and invalid tuples are not considered in this Master thesis project, since the tuple (equipment numbers) are automatically generated by SAP and error-free. The same holds for Missing tuples. Although the terms used are slightly different, the meaning and taxonomy of the data problems are comparable as displayed in Table 3.

Table 3. Comparison of Data Quality Problems in Literature.

Oliveira, Henriques & Rodrigues [2005]	**Müller & Freytag [2003]**
Combination of Missing value and Violation of business domain constraint	Lexical errors
Syntax Violation	Domain format errors
Imprecise value, Existence of synonyms	Irregularities
Domain violation	Integrity constraint violations
Violation of functional dependency	Contradictions
Missing value	Missing values

The data quality problems as described by Oliveira, Henriques & Rodrigues (2005) is more specific and complete than the ones by Müller & Freitag (2001). For that reason a selection is made of those problems that apply to the traceability data quality as studied in this Master thesis project. The problems relevant for measuring the traceability data quality analysis in this study are summarized in Table 4.

Table 4 Overview of traceability data quality problems and definitions

Traceability Data Quality Problems *(Oliveira, Henriques & Rodrigues, 2005; Müller & Freytag, 2003)*	Definition
Missing Value	Empty tuple, although a value is required.
Syntax violation	Information in a tuple is not adhering to the syntax rules required for the attribute in question.
Incorrect value	The value of the tuple is not a correct representation of reality.
Domain violation	The tuple value is outside the range of realistic and logical values for that specific attribute.
Violation of business domain constraint	The value is not within the domain that is set by specific business rules for that attribute.
Unique value violation	The tuple contains a duplicate value when not allowed.
Violation of functional dependency	The value of the tuple is dependent on another attribute, but the dependency is violated.

2.4 Measuring Data Quality

Several researchers attempt to highlight the importance of data quality and provide ways of measurement and assessment (Pipino, Lee & Yang, 2002; Lee, Strong, Kahn & Wang, 2002; Madnick, Wang, Lee & Zhu, 2009; Watts et al., 2009), with the most recent one the book Measuring Data Quality for Ongoing Improvement by Sebastian-Coleman (2013) that builds on all the existing literature of this topic and presents a Data Quality Assessment Framework (DQAF). Information quality is measured along a variety of dimensions such as accuracy, completeness, relevance, timeliness etc. Of these dimensions completeness and accuracy are among the most returning metrics and lend themselves to objective measurement (Watts et al., 2009). Another well-known method for measuring data quality is AIMQ, which stands for "A Methodology for Information Quality Measurement"(Lee et al., 2002) which presents questionnaires about the data quality to the data consumers and producers and statistically compares the answers among companies as a benchmark and within the company to analyze gaps in perception of the data quality. However, this measure is based on subjective data, and therefore not sufficient to fulfill the need of quantitative results at Philips.

This illustrates that there is no explicit technique of measuring data quality, but each and every case needs to be assessed individually to determine a framework with metrics relevant for the specific set of data and the needs of the organization.

Building on, among other, Eppler & Helfert's (2004), Sebastian-Coleman's (2013) and Sadiq's (2013) work, Laranjeiro, Soydemir & Bernardino (2015) again acknowledge and underpin the importance and potentially high impact of poor data quality in organizations. Next to that, they surveyed data quality classification research, while emphasizing on defining data quality dimensions and classification of poor data quality. Based on data quality research in several domains, including data-science, cyber-physical systems and healthcare, the authors provide a standardized procedure of mapping data quality problems

over the metrics Accessibility, Accuracy, Completeness, Consistency and Currency (Timeliness) which, according to their research, prove to be the most important metrics. Although the technique of the AIMQ method by (Lee et al., 2002) does not fit this research, they did gather a large amount of metrics based on a multitude of studies to measure data quality in their surveys: Accessibility, appropriate amount, believability, completeness, concise representation, consistent representation, ease of operation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness and understandability. Of all these metrics, the ones required for this study are the ones that correspond with the database community of research as previously explained. The ones remaining are then compared to the metrics as provided by Laranjeiro et al. (2015).

Table 5 Comparison of Technical Database Oriented Data Quality Metrics

Lee et al. (2002)	**Laranjeiro et al. (2015)**
Completeness	Completeness
Free-of-error	Accuracy
Timeliness	Currency

Here we see that the four metrics from Lee et al. (2002) can be summarized into three metrics, with the more general metric accuracy, since concise representation can be seen as an accurate value and free-of-error is essentially also saying that the data is correct, or accurate. Lee at al (2002) do not provide their own definitions of the metrics, but rather use the definitions from other papers:

Completeness: As an intrinsic dimension, completeness is defined in terms of any missing value. As a contextual dimen- sion, completeness is also defined in terms of missing values, but only for those values used or needed by information consumers.

Free-of-error: This information is correct.

Timeliness: n.d.

The definitions of the metrics by Larajeiro et al. (2015) are:

Completeness: The degree to which an entity has values for all expected attributes and related entity instances.

Accuracy: Degree to which data's attributes correctly represent the true value of the intended object.

Currency: The extent to which data holds attributes of the right age.

With this information it is possible to create a mapping of the data quality problems for this study of Table 4 across the different metrics, similar as in the study by Laranjeiro et al. (2015). In this case the metrics required are completeness, accuracy and currency.

Table 6 Mapping of Traceability Data Problems into Metrics

Data quality problems	Accuracy	Completeness	Currency
Missing Value	x	x	
Syntax violation	x		
Incorrect value	x		
Domain violation	x		
Violation of business domain constraint	x		
Unique value violation	x		
Violation of functional dependency	x		

Missing values scores on completeness, since the data would not be complete when missing a value. However, when a value is missing, it is also not the right value and therefore scoring on accuracy as well. Syntax violations are obviously score on accuracy, since the value cannot be correct if the syntax is wrong. The same goes for every other data quality problem that is a violation. If it violates any (logical) domain, constraint or functional dependency, it scores on accuracy. What stands out in Table 6 is that none of the problems scores on currency. That is since this study does not take currency, or timeliness if you like, into account for the analysis due to not being able to measure that with the composed dataset and boundary setting due to time restrictions, for the problems listed in the table were the most pressing issues for the Philips Case.

2.4 Value of Data Quality

As earlier indicated by Redman (1998-a) there are many tactical and strategical implications of low quality data which are present but hard to measure. The financial implications do not differ with respect to the ease of measurement. In their paper Eppler & Helfert (2004) state that it is particularly difficult to calculate costs of insufficient data quality levels, because most of them are indirect costs. Indirect costs are hard to identify, let alone to quantify in monetary results. Nevertheless, their research focusses on identifying, categorizing and measuring the costs of missing data quality and how to establish causal links between data quality effects and monetary effects. Consequently, they present a framework that can be used in various business and research scenarios, specifically for risk assessment, business case, program assessment and benchmarking of data quality. Haug, Zachariassen & van Liempd (2011) identify two types of costs for faulty data. One being the costs for cleaning and ensuring high quality data and the other being the costs of not cleaned data that lead to faulty decision making. The cleaning and maintaining of high quality data becomes less profitable at some point, and is complex to measure but relatively easy with respect to estimate the costs poor quality data has, due to the indirect and intangible effects associated with it. Luckily they simplify the estimation of the inflicted costs, by breaking them down in two dichotomies: direct versus hidden costs and costs on operational or strategic level. Hidden costs are the costs that the company is incurring, but management is not aware of and direct costs are

the costs visible and immediately present to management. On the operational level, the data is used for making decisions and performing tasks with a relatively short time span, where the strategical level of costs is related to the decisions based on the data that have a relatively long time span compared to the operational level.

Eppler and Helfert (2004) provide and extensive classification and analysis of the different types of costs of low data quality. An taxonomy of all the costs related to low data quality is displayed in Table 7. Although they don't provide definitions for the taxonomy they made, the terms are more or less self-explanatory. An important distinction they make in their taxonomy is between the costs of having low data quality and the costs improving or assuring the data quality. The costs caused by low data quality are then divided into direct costs and the indirect costs that are harder to identify and measure. The costs of improving or assuring data quality are costs concerning the prevention of diminishing data quality, the detection of data quality problems and costs for repairing or restoring the data quality.

Table 7 Taxonomy of data quality costs (Eppler & Helfert, 2004)

Data Quality Costs				
Costs **caused** by low data quality		Costs of **improving** or **assuring** data quality		
Direct costs	Indirect costs	Prevention costs	Detection costs	Repair costs
Verification costs	*Costs based on lower reputation*	*Training costs*	*Analysis costs*	*Repair planning costs*
Re-entry costs	*Costs based on wrong decisions*	*Monitoring costs*	*Reporting costs*	*Repair implementation costs*
Compensation costs	*Sunk investment costs*	*Standard development and deployment costs (system and process setup)*		

In many organizations however, data is not valued in monetary terms at all. The only way people can maximize the value of the data, is by completely understanding the data they are using (Sebastian-Coleman, 2013). In Data Quality for the Information Age, Redman (1998-b) identifies six ways in which poor quality data can affect an organization's financial performance: impacting decision making, impeding reengineering efforts, lowering job satisfaction and breeding organizational mistrust, introducing unnecessary costs and impeding long-term business strategy. Furthermore he points out that producing and maintaining high data quality can be a unique source of competitive advantage, since it is unique to the organization and a reflection of its history. It cannot be purchased, replaced or substituted, which allows it to be an extremely unique asset, for which it should be treated like that.

Glazer and Henry (2012) emphasize that the value can be high when it is well protected and loss is impossible, but if the business user cannot access it in order to make decisions, the value is zero. On the

other hand, when the business user has access but the information is replicated with variations in each replica and impacted on quality and integrity, the information value is again very low. All in all, the value of data can be very high, but is hard to translate to monetary terms. Although everything depends on the quality of the data and how it is utilized.

Although high quality data is valuable, the exact measurement of the value remains an issue. Loshin (2001) provides a method of quantifying the measurement of data quality by means of a cost-benefit analysis. This analysis is done by relating the impacts of high versus low levels of data quality into eight categories as seen in Table 8:

Table 8 Categorization of costs and benefits by Loshin (2001)

Category	Definition
Cost increase	The degree of which poor quality data increases the cost of doing business.
Revenue decrease	The effect of low data quality on current revenue.
Cost decrease	The extent wo which improvement in data quality can reduce costs
Revenue increase	The extent to which improvement in data quality increases revenues.
Delay	Measures the degree to which there is a slowdown in productivity.
Speed up	Measures the degree to which the cycle time of a process can be reduced.
Increased satisfaction	Measures the increase in level of satisfaction of the employee, shareholder or customer.
Decreased satisfaction	Measures the decrease in level of satisfaction of the employee, shareholder or customer.

The first category is cost increase, which measures the degree of which poor quality data increases the cost of doing business. Second is revenue decrease, which measures the effect of low data quality on current revenue. Cost decrease measures how an improvement in data quality can reduce costs, while revenue increase measures how improving data quality increases revenues. The delay measures whether there is a slowdown in productivity, where speedup measures the degree to which the cycle time of a process can be reduced. The final two categories are increased and decreased satisfaction, which measure the change in level of satisfaction of the employee, shareholder or customer. Loshin states that these measures need to be evaluated for all three domains as defined before by Redman (1998-a): Strategic, operational and tactical domain. Apart from specifying the categories into which the costs and benefits of data quality can be classified, Redman (1998-a) does not further explain how to exactly quantify those costs.

In the book of McGilvray (2008) however, she does describe eight steps to identify business impact, ranging from least time consuming and complex to more time consuming and complex. Those steps do not need to be done in consecutive order, but can be used in combination based on available time,

resources and experience. Step one for instance is using anecdotes to estimate the impact of the data quality, which is the least time consuming. Step two is determining the various business processes, people, and/or applications that use the data. Step three is the five “Whys”, step four a benefit versus cost matrix and step five is ranking or prioritization to determine on which data quality efforts needs to be focused most. Step six visualizes the business impact for one or two specific processes, on which step seven can be used to identify the costs of low-quality data. Step eight is a complete cost-benefit analysis of the data quality, which is the most complex and time consuming.

2.5 Causes of Traceability Data Quality Problems

Not only is it important to understand to identify and classify the current state of the data and the organizational, tactical or strategical impact it has on the organization, but also the reason of (poor) data needs to be understood (Sebastian-Coleman, 2013). For a large part, this becomes clear during the identification and classification of the data quality problems. When human error is the cause it could be a simple mistake, which humans tend to make. However, when it comes to traceability data Arkley & Riddle (2005) and Mader, Gotel & Philippow (2009) found that lack of motivation is one of the key issues due to the fact that no direct benefit is perceived by the engineers that have to register the traceability information. Although their research is on requirements engineering for software, the same might be applicable within other domains.

Concerning material or industrial parts traceability, there is very little to be found in literature but the main principle of traceability is the same across fields and sectors like the tracing of software or even food and other agricultural products: you need to know at every point of the material’s cycle where it is, came from and everything in between and therefore the problems in these fields might be comparable. Where (Regan, McCaffery, McDaid & Flood, 2012) identify costs and complexity associated issues as barriers for software companies to implement traceability. Other barriers in the software requirements traceability are lack of guidance, political issues and issues with tools for implementing software traceability. Other research by Wohlrab, Steghofer, Knauss, Maro & Anjorin (2016) supports these barriers and stresses that practitioners have to see the usefulness and benefits of traceability to overcome them. A crucial tool in persuading practitioners turns out to be the cost-benefit analysis. Apart from financial benefits, traceability can enable collaboration, facilitate knowledge management and support coordination across disciplines in systems engineering contexts.

Another domain in which traceability research is performed is the agri-food sector. Although substantially different and possibly less complicated from material traceability due to the food agri-food products going in one direction, without being replaced, repaired or recycled, there are also similarities. For instance, it enables improved process control and management systems, optimizes processes and planning and meets current and future government requirements (Moe, 1998). Sterling et al. (2015) add the consumer’s attitude towards sustainability and addressing market demands as drivers for food traceability. These drivers correspond with the drivers found by Bosona & Gebresenbet (2013) and consist of regulatory concern, safety and quality concern, social concern, economic concern and technological concern.

All of these drivers are applicable to the Philips case. Whereas in the agri-food sector legislations need to be abided in order to stay in business and maintain market power, for medical devices it is no different, especially with the upcoming EUMDR legislation in May 2020 that requires optimal traceability of

medical devices and components. Apart from that, there can be no ownership disputes when the materials are traceable, the same as it prevents animals from theft in the case of animal production in the agri-food sector. One very important driver for both the agri-food and medical sector is the safety and quality concern. In either of the sectors, when safety and quality cannot be guaranteed, it can seriously impact the business. Traceability helps maintaining safety and quality standards and quickly identifies source and location in case one of them is negatively affected in order to take appropriate measures instantly. Socially the customer confidence and trust in the organization grows when safety, quality and regulations are assured and incidents are adequately handled. Both sectors have economical drivers, although the reason differ. The agri-food sector profits from better market access and better prices, while being able to receive funding. In Philips' perspective the economic driver is the cost reductions and competitive advantage traceability brings. The technological driver for both the medical sector as the agri-food sector is that technology nowadays has all the possible solutions to enable traceability, ranging from barcodes to RFID tags and comprehensive ERP systems.

Bosona & Gebresenbet (2013) provide several advantages of food traceability. Firstly, it reduced supply-related costs and logistic costs due to effective management of resources. Furthermore, it helped develop better quality products, which led to competent market performance. Additionally it contributes to scientific and technological research in food traceability systems due to the demand for effective traceability systems while contributing to agricultural sustainability by creating awareness through data capturing and reducing food loss by promoting effective packaging technologies. All potential advantages that could be applicable in the medical device industry.

These advantages come with their own barriers, significantly different from the barriers in software requirements engineering. One of the first barriers is the difference in international standards for food traceability (Sterling et al., 2015). Another barrier is the fact that for successful traceability systems all partners in the supply chain need to cooperate, which will cost each of them considerable investment of money and effort. Agriculture is inherent to uncertainty, making it difficult to collect certain and timely data throughout all stages of the supply chain. Additionally there is lack of standardization in capturing data and companies might not have the capacity to implement traceability along the entire supply chain, and ultimately a familiar barrier for traceability as well as data quality presents itself again, the costs versus the benefits of implementing and maintaining traceability (Sterling et al., 2015). For Philips some of these barriers don't even apply. Since Philips depends on materials and these are less sensitive than food or animals, there is less uncertainty. Furthermore, there is already some standardization in place concerning the capturing traceability data and a lot of the traceability along the supply chain is already in place. This allows for a relatively low costs versus the benefits traceability has to offer.

2.6 Conclusion

In the end, the purpose of the literature review is to find answers to the questions "which techniques exist for measuring data quality?" and "which techniques exist for estimating the value of data quality?" There is no "off the shelf" solution for data quality, since every database is company specific and unique. Therefore a specific framework is required for data quality assessment, consisting of the data quality problems specified for that specific company's data. For these problems the corresponding metrics are then identified on which the data quality can then be assessed. Estimating the value of any level of data quality however, is far more complicated. The main cause the fact that the effects of poor data quality

are almost all indirect and with that, not only hard to identify but also hard to quantify into monetary effects. Also, data quality comes with two types of costs. One being the costs of dirty data that leads to bad decision making, while the other costs concern the cleaning and ensuring high quality data. At one point, high quality data becomes less profitable to clean and maintain. Estimating the costs of cleaning and maintaining data quality is relatively easy compared to estimating the costs of non-quality data, again because of the indirect effects within the entire organization on both strategic as operational level. As for the techniques mentioned to estimate the value of data quality, it all comes down to a cost-benefits analysis. However, apart from categorizing, Loshin (2001) and Redman (1998-a) succeed in categorizing the costs and benefits of data quality, but lack techniques to actually quantify the value. McGilvray (2008) manages to provide several steps that can be taken in order to quantify the data, of which step one, anecdotes, is the most feasible for this case study, regarding time and resource restrictions combined with the size of Philips and the large number of people and processes using the data all around the world.

3. Data Quality Problems and Measurement

In order to determine the costs and benefits of improving the traceability data quality, first there needs to be an understanding of the level of data quality. In its turn, assessing the current level of data quality requires understanding of what data is required, how it is structured and collected to ultimately be measured. The goal of this section of the thesis is to answer the research question:

What is the current quality of the traceability data at Philips?

For this research question data quality is defined as by ISO/IEC (25012:2008), the "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions". The specified condition as required by the definition of data quality, is in this case tracing back all unique copies of a specified material from the location and system they are in now to the moment they were shipped, with every event performed on them in between. The stated and implied needs need to be determined by interviewing Philips employees that are experts of the data and material, and initial, manual analysis of the traceability data. Because the participants are not expected to have sufficient specific knowledge about data quality and data quality problems, the interviews are used to identify business rules to which the data must adhere. The business rules are then labeled on the data quality problems they indicate. The data is then analyzed to how well it scores on each business rule. After that, an overall analysis is performed to measure the number of systems that follow or violate any of the business rules.

3.1 Research method: case study

The primary objective of this study is exploratory since it is used to gain a deeper understanding of the traceability data quality and the consequences thereof, using qualitative data. However the design is flexible, meaning that both qualitative and quantitative data are used. These are the typical characteristics of a case study as described by Runeson & Höst (2009). Within a case study, triangulation is an important tool to enhance the empirical research precision. Triangulation means providing a broader perspective on the studied object by taking different angles towards it. In this specific case study methodological triangulation is used, as explained by Runeson & Höst (2009) here different types of data collection (qualitative and quantitative) are used. The study starts with qualitative research by having semi-structured interviews concerning the data structures and the perceived issues and how to collect this data. By the information extracted from those interviews the traceability data is collected and analyzed to conclude this chapter of the thesis. The next chapter uses the results of the problem understanding and measurement for a new series of interviews to determine the impact of the found traceability data quality.

3.1.1 Qualitative analysis

Quantitative analysis is used in an explorative way to understand what are the key sources and attributes for traceability data analysis and what business rules are applicable to the data. Later on in the same method is used to analyze the impact of the found data quality. The data for the qualitative analysis is gathered by conducting interviews with Philips employees. Of these interviews recordings are made and notes are taken. The recordings are worked oud on paper and combined with the already present notes. The information is from the first series of interviews is then coded on “source”, “attribute” and different data problems, whereas the second series of interviews concerning the impact of the data quality is coded on the impact classification as described by Loshin (2001): cost increase, revenue decrease, cost decrease, revenue increase, delay, increased satisfaction and decreased satisfaction. Once the data is coded, the information of each theme used for codification is combined to extract the required information per theme, and find the required sources, attributes and business rules for the first series of interviews. For the second set of interviews the impact is mapped on the eight different classifications.

3.1.2 Quantitative analysis

After collecting the traceability data from the different sources and determining which attributes are necessary for traceability data for the determined scope, quantitative analysis takes place on the dataset. This is performed by statistically measuring the data quality over the different business rules identified in the interviews and by initial qualitatively going through the data to see what data problems occur. For each business rule, the quantitative analysis is performed in excel using hard coding and pivot tables in order to score it on to which extent the data adheres to the business rule.

3.2 Interviews

During the interviews the goal is to get a view of the perceived problems and understand the nature of them, together with an understanding of the data structure. Semi-structured interviews are held to determine what data is important for material traceability purposes and information on the related processes is gathered to identify how the data can be affected during the lifetime of a system. Additional interviews had the goal of understanding where to find the relevant data and how to extract them from the SAP environment and use it for analysis.

Among the interviewees are several Philips IGT Systems employees related to, and specialized in, traceability data. Other participants are material specialists and program managers in order to grasp a broad understanding of the data structure, quality, processes and issues. During the interviews notes are taken and audio is recorded as way of collecting data. After gathering and analyzing the data from the interviews, the extraction of business rules started to which the traceability data must adhere. During the process, it became clear that several business rules are material specific, and therefore the decision is made to select a specific material for analysis as a case study.

Table 9 interview participants, roles and findings

Participant	Role	Reason for interviewing
1A	Program manager IGT installed base	Head of the installed base program, the person that has the responsibility of driving improvements for the installed base
1B	Service Program Manager, IGT Systems Management	Expert in Philips' data, working at Philips Healthcare for over seven years. Before that time also working at Philips, related to data. Also works on remote connectivity program, that focusses on connecting all components of an installed system to one hub and connects to Philips, able to retrieve information about those components in a safe way.
1C	Business Analyst Customer Service	Business analyst that works with traceability data on daily basis. Knows how to extract and merge the right data and what things are important for traceability purposes.
1D	Director IB Sustainment	Has the responsibility over the entire IB sustainment department, which includes the Installed Base program this study is performed in. Long history within Philips, highly involved and interested with the traceability data.
1E	Quality Management System Manager	Part of the Quality and Regulatory group. Is involved in ensuring that Philips abides by the necessary regulations.
1F	Global Installed Base BI Program Manager	Is involved in the 'close the loop' program that uses traceability data. Also multiple years of experience within Philips.
1G	Sr Business Process Analyst, Service Parts Supply Chain Architecture	Expert on SWO data.
1H	Productivity Project Manager	Bases his proposals and opportunities concerning x-ray tubes on traceability data. Is seen as expert of all the different kinds of tubes.

3.2.1 Interview results

During the interviews the method for measuring the traceability data quality becomes clear. There are certain attributes our data must contain, that are required and intended for traceability purposes according to participants 1A, 1B, 1C, 1D and 1E. Also, it is important to know what correct source is to collect the data from, since there are various data sources used for different purposes within Philips (Participants 1A, 1B, 1D & 1F). The sources need to contain the master data and contain the right attributes. During the interviews, these attributes and sources are identified in order to analyze the traceability data registration per system. This means a historical overview of the material replacements is visible, by collecting the data from different sources: the data as the system and materials were shipped, the SWO data that represent all the inbound, outbound an unused returned materials and the Maintained data that represents the configuration of the system in the field.

3.2.1.1 Sources

In order to answer the research question what the current quality of traceability data is in this case study, specific data is necessary for analysis. The data is collected by running several queries in SAP MP1 and SAP MBP databases. The data required for this traceability case study is divided in three different kinds. The AsShipped data is required in order to have an overview of how the system and its materials are originally shipped to the customer. The Maintained data is the representation of how the actual configuration of the system and materials in the field is registered. Due to maintenance and repairs executed through a Service Work Order (SWO), the configuration is changed over time, so the Maintained data is updated and will differ from the AsShipped data. All those changes are registered in the SWO data (Participants 1A, 1B, 1C & 1D). The data of these three sources together form a complete overview of what tubes have been installed and deinstalled over time and are therefore required to identify and trace back the materials. Any errors in these data impact the quality and are therefore measured to determine the current traceability data quality.

There is not one overview from which all the required data can be extracted at once, the data comes from different locations within SAP. First, the AsShipped data is collected from SAP MBP, the kernel for all the factory data. The Maintained and SWO data are collected from SAP MP1, the kernel where all the data for the markets is stored (Participants 1B, 1C & 1D). More details on which queries to run to collect the data Is found in Appendix C: Data gathering queries in SAP MBP and MP1.

3.2.1.2 Attributes

Participants 1A, 1B, 1C, 1D, 1E and 1F all agreed that for traceability purposes, and to fulfill quality and regulatory demands, at least the 6NC and serial number of the system the material is installed in must be correctly registered, as the 12NC and serial number. Additionally the parent equipment number is required to group the data together per system (participant 1C). In order to determine what the latest registration update was and to determine the age of the system and the moment certain data problems occurred, the date is also a crucial attribute in the traceability data as participant D said: “If possible it is interesting to see when which data problems occur, so that we might be able to discover a specific period the errors occurred in and to have an overview over the performance over the years. If the amount of data errors is declining, it shows a positive trend. If the amount of data errors is increasing over time it is important to have that info as well and find out where improvements can be made”. With that last sentence, another important attribute is identified, the country code. This indicates in which country the system is installed and therefore which market is responsible for the maintenance and registration of the system. This code allows us to later analyze the performance per market and identify possible high or low performing markets that can lead as an example or need extra attention (participant 1D). The hospital name attribute is added to get a better understanding of the reason why some data might be flawed. For instance, some systems might be used for Philips internally as demo material or is sold to customers that don’t disclose and do not allow storage of any (material trace) data for instance.

Apart from attributes that are available from all data sources, there are also attributes that are limited to one data source. For the Maintained data the system status and material status are used in order to see whether the system or material is operational and whether those two statuses correspond. The status attribute values were necessary for scoping, but also provide an indication of what happened in the material registration (participant 1C). For the SWO data, the attributes movement type and Batch are

added, to identify the type of movement is represent for the tuple and whether the material is unused (U) or defect (D) respectively.

3.2.1.3 Data model

For a better understanding of the structure and relations of the data, it is presented in a UML class model. The data model is based on the information gained from SAP and the input of an experienced business analyst that works fulltime on these databases. IGT Systems are made to order and customized to the customer's requirements. After ordering the specific composition of the system, a shell record is created in the Maintained data that only contains an equipment number and customer data (SAP MP1). As soon as the system is finished and shipped, the data of the system and all its traceable parts and components, the Bill of Material (BOM), is stored in SAP MBP - AsShipped data. After shipping, the system with the complete BOM data is directly inherited by the pre-existing Maintained shell record in SAP MP1 where the SWO data is also generated. A UML class diagram is presented in Figure 5, where the AsShipped data is displayed as one, grey-colored class. In fact it consists of the same classes and entities as all the green classes together that make up the Maintained data of the system, but for readability purposes it is simplified into one. The SWO data and its entities are shown in the blue classes below the Maintained.

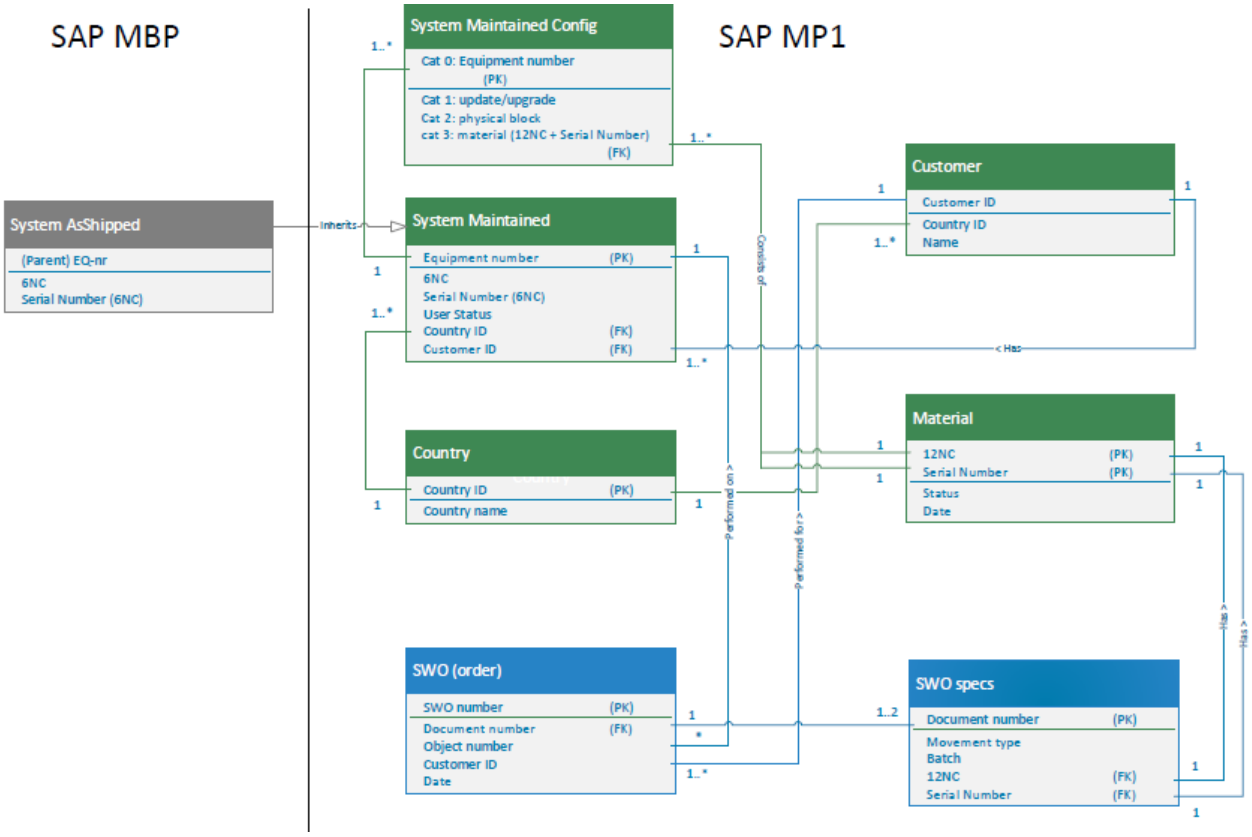


Figure 5 UML Class Diagram of material traceability data displaying the relations between data entities.

Below, the UML class diagram and the entities and relations will be explained in detail. First the Maintained data will be elaborated on, up until the country and customer part. Afterwards the SWO data will be treated, where the AsShipped data does not require any further explanation on entities and relations since it is exactly the same as the Maintained data in that respect.

Material

Within the Maintained data, several entities are important and are discussed here. Since Material traceability is the core of the study, the material class is treated first and then the whole class diagram is worked through from there. To begin with, a material has status which indicates whether the material is operational (OPER), available (AVLB), to be installed (TBIN), scrapped (SCRA), in storage (STOR), lost (LOST) or the status is unknown (UNKN). This study does not go into detail on the statuses, but only scoped on either systems and/or tubes with the status operational and reports any inconsistencies between system and material statuses when found. After all, a system with status operational and a tube in it that is not operational, or vice versa, must not be possible and would be a reason to raise an eyebrow. Next to the user status, a material has a date. This date represents the day it was registered as part of that system, it can either be the date the system with materials was shipped or the date it was replaced during maintenance or repairs and is useful in data analysis to see when actual changes were made to the maintained data.

As explained before, every material has a 12NC that indicates what type of material it is, but to be able to trace individual materials, this 12NC is combined with a serial number (SN) which is unique for that 12NC. That way the combination of 12NC and SN is unique for that specific material and therefore it is required to identify and trace every copy of the material. Of that specific combination of 12NC + SN there can only be one at any moment in a system. However, it can be seen in multiple systems over time when a tube is replaced and the old one gets repaired and reused again.

System Maintained Config

Tubes are always a category three material. Categories are a way of categorizing the data and materials in a system, displaying them in a physical view. The categories range from zero to three, with in category three the material. Category two is the physical block in which materials are grouped together. Category one is where upgrades or updates to systems are placed in the data and category zero is the parent equipment number level, which is the equipment number of the entire system. The top parent equipment number is used to identify and trace the systems and connect the traceable materials to them. All materials have an equipment number as well, visible at their own category but these are not used for this study since it is more practical to do a search query for 12NC since it says what type of material it is.

System Maintained

One parent equipment number stands for one unique system in the Maintained class. Every system / equipment number has a 6NC which indicates what type of system it is and therefore what rules could apply to those systems. In this study for example, most systems have only one example of this tube, where one type of system has two of the same type tube. This 6NC also has its own serial number, where

the combination of 6NC + SN is unique and together are linked to the equipment number. Furthermore, similar to and previously explained in the material section, the system has a status to represent the state.

Customer and Country

A system can only be in one country, which is indicated by a country ID, where a country can have multiple systems. Logically this country has a name that the country ID stands for. Within a country there can be multiple customers that each have their own ID. Customers can in principle be active in multiple countries, but for this case they are regarded as the physical customer at the physical location, for which there can only be one country ID per customer. The country ID is used to identify performance on data quality per market, for Philips to know where potential improvements are required. For a customer to be a customer, it would need to have at least one of our systems, but a multitude of systems is possible and not exceptional. For every customer the name is also available but not needed for analysis, it gives more information if outliers are detected what the reason could be for that outlier.

SWO (order)

Every SWO has a unique SWO number, to identify that specific work order, and a date on which the work order is registered. This date is key to see what SWO is the most recent, so what should have been the latest changes in the Maintained data. The SWO is linked to one specific system, where a system can have many SWO's or none at all (yet). This data is needed in order to link the SWO data to the system and be able to compare the SWO's performed per system. Also the SWO is linked to the customer for who the SWO is performed, which again can only be one, while vice versa there can be none or many SWO's for a customer. Although these data is not strictly required for traceability, it can be useful for background information in the case of outliers. The document number directly linked to the SWO specs class, which is a document with the further details of the SWO, concerning the movement and materials in the SWO.

SWO specs

This class is a further specification of the SWO (order) class, with the document number as direct link between them. First of all there is a movement type, which as discussed earlier can be 261 for inbound, 262 for outbound and 632 for unused returned materials. Batch is used to identify whether it is an unused material (U) that is installed in the system (mvt 261) or a defective material (D) that is removed from the system (mvt 262). In case of an unused return (mvt 632), the value of batch must be "U". Furthermore the inbound and outbound material are always defined in the SWO specs by 12NC + SN. This information is crucial for the traceability concept, when analyzing the flow of the materials. In case of a replacement, there are always two movements, the 261 and 262 for the inbound and outbound material respectively. Both movements are registered under the same document number, where the document number relates back to the SWO (order) class.

3.2.1.3 Case Selection

As explained before, during interviews it became clear that it is necessary to specify the right material for traceability data analysis that has to fulfill a set of requirements. First and foremost, it should be a traceable material, which means it is classified within Philips as a traceable material and has its own serial number. The traceability dataset should be large and complex enough for meaningful analysis, to an extent that processing and analyzing remains impossible. For instance, a material with a very low replacement rate (due to defects being exceptions) results in barely any flow in data. In this case, the chances of data being changed would be very low, leaving very little room for error while not accurately representing the process, and also resulting in a small dataset. On the other hand, the opposite holds. High replacement rates of materials lead to enormous amounts of data, which overcomplicates and clouds initial analysis and is less feasible for the timespan of a master thesis project.

With this in mind, the chosen material is a specific type of X-ray tube, MRC 200 0407 ROT-GS 1004, that is used in IGT Systems. The reasoning behind this choice is that the tube is a crucial part for the systems' functioning and has an average replacement rate of approximately once per five years, depending on the individual situation of every system and the market it is in. The value of such tubes is significant, providing clear insight in the effects of potential data quality issues for Philips. The tube in question is used for frontal x-ray imaging and can therefore be used in both monoplane and biplane IGT systems. The reason it can only be used for frontal scanning is the strength of x-ray emissions, which in principle means there can only be one tube of this type in the maintained data per system. One exception exists for a single type of system that uses the same tube for lateral imaging and thus can have two of these tubes in its maintained data. The tube in question can have one of five 12NC's (twelve digit numeric code), depending on which version of the tube type it is and whether it is shipped with the system from the factory or whether it is a replacement part. The form, fit and function of the tube stays the same across the allowed 12NC's. The variations for the tubes shipped with the system from the factory are 989000086501 and 989000086502, while the tubes sent as replacement parts can have one of these three 12NC's: 989000085101, 989000085102 or 989000085103. Between the 12NC in the replacement category (or factory) the first 11 numbers are the same. The variation in the last number stands for the "point number", or version number. A higher last number means a newer version of the same material with minor changes. Differences between factory installed parts and replacement parts are also very small.

3.2.2 Business Rules

The goal of this chapter is to determine what the current data quality is. As indicated before, this is done by measuring how well the data abides by a set of business rules that are applicable for the material selected in this case study. To begin with, the possible data problems derived from the interviews and initial data analysis are listed and classified by the data problems and metric that resulted from the literature review in Table 10. For every problem is indicated which participants mentioned the problems and in to which stakeholder group they belong: Market Organization (MO), Customer Service (CS) or Quality & Regulatory (Q&R).

Although some problem classification problems are very straightforward, others might need some more explaining. For instance, the amount of inbound movement differing for the amount of outbound movements, means that there is no inbound record found at all, while an outbound record is present or vice versa. Both movements are supposed to be logged in one SWO document. However, if for instance the movement type or the 12NC is not registered in the SWO, the data of that movement is not retrieved with the SAP queries. The next one, lifecycle of a tube is for that reason classified as a business domain constraint violation and missing data. It could well be that due to a missing value, the SWO data of a replaced tube is not retrieved, for which it seems that no replacement has taken place, while in reality there has but does not show up in the data. Though, if the data is complete and the expected lifecycle of a tube is well surpassed, it becomes a business domain violation, because Philips expects it to last for approximately five years. Another example of problems falling within different classifications are the systems (6NC) and material (12NC) with serial numbers (SN). For a material where it would occur that a serial number simply is missing, this situation makes it a missing value problem. It can also be a domain violation. Serial numbers of systems and some materials start at 1 and the 'n'th copy of it has serial number 'n'. If a serial number has a higher number than the number of systems or materials of that type ever produced, it is seen as a domain violation, but also an incorrect value. Furthermore, syntax restrictions apply to serial numbers. Systems only have plain numbers starting from 1, while material serial numbers can have letters and numbers. For the material in this case study, the serial number consists of five digits, then the letter "M" followed by six digits. Any deviance from this format is a syntax violation, but also incorrect values for the serial numbers can be entered. Accordingly, a wrong type of 6NC is a business domain constraint violation, since the 6NC's represent types of system as assigned by Philips.

Unique value violations is in this case considered in two aspects. For an outbound or unused return movement, the 12NC+SN of the SN should only be visible per system twice, in the outbound or unused return movement type and in the inbound movement or AsShipped data. It must not be visible in the Maintained data for that system or in any other outbound movement. In this aspect, there is one (or more) too many duplicate values of the 12NC+SN combinations. In the other aspect, the 12NC+SN of the latest inbound material, via SWO or AsShipped, is expected to be in Maintained, but not in any other inbound, outbound or unused return movement. Also, the 12NC+SN of the outbound material must be visible as a previous inbound material, but not in the Maintained data anymore. So this problem is not typically a unique value violation, but more a maximum amount of duplicate values problem, based on a set of constraints, making it one of the most complicated problems.

Functional dependency violations are more straightforward, for when a system status changes, the material user status is expected to change to the same status. For example, it cannot be that a system is scrapped, but the tube that is assigned to the system is still operational, simply due to the fact that the tube cannot be operated without a system. Furthermore, a system's equipment number is fixed and does not change over time. The 6NC+SN indicates the type and copy of the system and therefore is directly linked to the equipment number. The only way the 6NC can be changed for an equipment number is when a large upgrade is performed, in which case this is also visible in the data. However, it can also occur that an update is not registered to the system it is performed on. This means that it is floating. This phenomenon is known due to different regulations in the past, but due to the scoping in this study, that data does not appear in the dataset. The last data problem in the list concerns which 12NC is used, based on whether it is a replacement part or factory installed part. This distinction was

created in 2014 because of RoHS (Registration of Hazardous Substances) legislation, which means that since then there are specific 12NC's allowed for factory installed parts or field replacement units.

Table 10 Classification of possible data quality problems in Philips data

Problem	Metric	Mentioned by participants	Stakeholder
Missing Value	Completeness		
Data from one or multiple sources is missing for a system.	& Accuracy	A, B, C	MO / CS / Q&R
Amount of inbound movements differs from the amount of outbound movements.		-	MO / CS / Q&R
Lifecycle of a tube is on average 5 years.		H	MO / CS
Tuple does not have a valid 12NC+SN registration.		A, B, C, D, F	MO / CS / Q&R
Tuple is not registered to a valid 6NC or 6NC+SN.		A, B, C, D, F	MO / CS / Q&R
Syntax violation	Accuracy		
Serial number is in the wrong format.		C, H	MO / CS / Q&R
Tuple does not have a valid 12NC+SN registration.		A, B, C, D, F, H	MO / CS / Q&R
Tuple is not registered to a valid 6NC or 6NC+SN.		A, B, C, D, F	MO / CS / Q&R
Incorrect value	Accuracy		
Attribute value is incorrect.		A, B, C, D, E, F, H	MO / CS / Q&R
Tuple does not have a valid 12NC+SN registration.		A, B, C, D, F	MO / CS / Q&R
Tuple is not registered to a valid 6NC or 6NC+SN.		A, B, C, D, F	MO / CS / Q&R
Domain violation	Accuracy		
Tuple is not registered to a valid 6NC or 6NC+SN.		A, B, C, D, F	MO / CS / Q&R
Tuple does not have a valid 12NC+SN registration		A, B, C, D, F	MO / CS / Q&R
Violation of business domain constraint	Accuracy		
Lifecycle of a tube is on average 5 years.		H	MO / CS
Equipment number has wrong 6NC.		A, C, D	MO / CS / Q&R
Unique value violation	Accuracy		
Outbound or unused return 12NC+SN is also present in Maintained data.		C	MO / CS / Q&R
12NC+SN of latest inbound part is not visible in Maintained data.		C	MO / CS / Q&R
12NC+SN of the outbound part, is not registered as an inbound material, either via AsShipped or SWO.		C	MO / CS / Q&R
Violation of functional dependency	Accuracy		
System and material status are not the same.		C, D	MO / CS
Equipment number has wrong 6NC.		A, C, D	MO / CS / Q&R
Wrong 12NC for factory installed or field replacement unit.		C, H	Q&R

Based on the identified possible data quality problems, a first draft of the business rules to which the data must adhere is set up. Then they are refined and confirmed by manually going through the dataset and through discussions with business analysts and product experts. It is in iterative process that ultimately extends throughout the entire research due to possible new findings and information. The business rules aim to measure the data quality on specific aspects over the metrics completeness and accuracy, and therefore are crucial in answering the question "what is the current quality of traceability data at Philips?"

To be able to answer that question, thorough data analysis is required. The analysis is performed using excel, by means of formula's, pivot tables and charts. Before starting the analysis, it is important to set up business rules to which the process and data should adhere. These business rules are derived from findings in the data and open interviews with several employees and data specialists within Philips concerning the requirements and processes through open coding. The coding is performed based on the framework created in the literature review, classifying the business rules in the different data quality problems and metrics as shown in Table 10. Subsequently, the business rules are also labeled on the quality problems from Table 10, which is between brackets after each business rule.

Based on the number of materials that adhere to, or violate these rules, the data quality is evaluated. The number of non-violations over the totals is expressed in percentages as KPI's to how good the data quality is. The overview of the 14 business rules is company sensitive information and therefore only disclosed to Philips in Appendix E. To give an idea of what these business rules could look like, few fictional examples based on cars are provided. Additional to the business rule example, overall assessment is added to indicate the overall data quality.

Example 1: A car must have exactly one steering wheel in the Maintained data. *[Unique value violation/Accuracy]*

A car can must have one steering wheel. Two or more would be pointless, while no steering wheel at all makes it impossible to steer the car. Therefore in the Maintained data of the car, there must be exactly one steering wheel present. If not, it is seen as a data quality problem.

Example 2: The Unique Device Identifier (UDI) of the last inbound steering wheel must be the same as in the Maintained data.

If the steering wheel is replaced, the new replacement part is installed. This means that the latest inbound material needs to be updated in the Maintained data and replaces the previous one. If the Maintained data is not updated correctly, or the inbound material is not registered correctly, it will result in a violation of this business rule.

Example 3: The Maintained data of every car needs to be correct.

This business rules checks for all cars whether any of the business rules that concern the Maintained data are violated. This way, it provides an overview of all cars of which the Maintained data is affected.

The same analysis can be done for any car at any point in time to give insight for which mistakes have been made in registration at any point in time.

3.3 Document Analysis

After a complete set of business rules is defined and is verified whether all possible data quality problems are covered, the actual data analysis is executed in order to answer the research question: "What is the current quality of traceability data at Philips". The analysis is performed for each of the business rules separately in Excel, using hard coding and pivot tables. The individual results are combined and aggregated to final result.

3.4 Impact Analysis

After estimating the current level of data quality at Philips, this section aims to answer the questions "what are the costs and benefits of the current traceability data quality?" and "What are the costs and benefits of improving the traceability data quality?" The costs and consequences of data quality are indirect, as previously stated by Eppler & Helfert (2004). That indicates that there is no straightforward solution to measuring the impact the costs and benefits of low or high data quality. As McGilvray (2008) showed, there are eight methods to analyze the impact of data quality, ranging from least to most time and resource consuming. Since a Master thesis project is very limited in time and resources and the impact analysis is just a part of the entire project, the impact analysis here is restricted by the first method of analyzing the impact: anecdotes. These anecdotes are collected by means of interviews, to gain an understanding of how the data quality may affect the different stakeholders in their business and thereby answer the research question "*What is the impact of the current and improved traceability data quality, and what can be learnt from it?*". Table 11 provides a list of all the roles of the participants, which stakeholder group they belong to and why they are interviewed.

Table 11 Roles of participants of the impact analysis interviews with their respective stakeholder groups (CS = Customer Service, Market, SPS = Service Parts Supply Chain).

Participant	Role	Stakeholder group	Reason for Interview
2A	Business Analyst 1	CS	Works with traceability data on a daily basis.
2B	Business Analyst 2	CS	Works with traceability data on a daily basis.
2C	Business Analyst 3	CS	Works with traceability data on a daily basis.
2D	Productivity Project Manager	CS	Specialist on the case material, uses traceability data to identify & quantify productivity opportunities
2E	Operations Manager FSE BeLux	Market	Leads a team of Field Service Engineers, can provide their perspective on the data registration.
2F	Modality Manager UK & Ireland	Market	Former Field Service Engineer and at that time was directly involved with the traceability data quality and registration. Hence has valuable insights in how the data is registered.
2G	FCO Manager	CS	Uses the traceability data to identify the systems on which Field Change Orders have to be issued on.
2H	Service Program Manager	CS	Has a lot of knowledge of and experience with the data in SAP.
2I	Product Support Manager	CS	Leads the team that provides the last line of support globally. For their job they use the traceability data to analyze what system and materials they are dealing with before they arrive on site.
2J	Services & IB Marketing Manager	Market	Looks at the traceability data from a sales perspective and has been cleaning up a lot of the scrapped systems in the data for the past year.
2K	Global Planning & Supply Manager	SPS	Concerned with the service parts supply chain and are looking to base their estimates and prediction on traceability data.
2L	Director Customer Service	CS	Has the responsibility over the entire IB sustainment department, which includes the Installed Base program this study is performed in. Long history within Philips, highly involved and interested in the traceability data.

Worth noting is that Quality & Regulatory had already been interviewed in the first series of interviews. From the information in that interview it was very clear what the consequences of flawed data are for their department. If the traceability data quality is not 100%, as the regulatory bodies expect it to be, there is a chance on heavy fines or even inability to sell products in the market the regulatory body is active, based on how serious they think the violation is. Quality & Regulatory was not able to give estimates on when which measures will be taken, because that is something the regulatory bodies determine for every violation individually and the regulations concerning the EUMDR are still being redeveloped and defined. Additional to the Market and Customer Service stakeholder groups, a participant of Service Parts Supply Chain is included in the interview series. Although they don't belong to the key stakeholders as identified in the introduction, their input can add valuable information towards the impact analysis.

During the semi-structured interviews the findings of the traceability data analysis are shown and explained. Afterwards, the participants are asked what the impact of the found data quality is or can be for their personal professional environment, as well as for Philips in general. Additionally they are asked what the impact would be if the data quality was less than 100% and what according to them potential reasons for errors are in order to map the costs and benefits. Furthermore participants received the question what aspect of the data they think is most important, trying to discover what data is most valuable and has the highest impact on Philips' business. During the interviews notes are taken and the audio is recorded. After the interview, the notes are complemented with the information from the interview audio to have all information on paper. The interview notes are then analyzed using codification in order to analyze all aspects of the impact the found data quality.

3.4.1 Codification

In literature, eight categories of a cost-benefit analysis to determine the value of data quality are determined Loshin (2001). These categories are “cost increase”, “revenue decrease”, “cost decrease, “revenue increase”, “delay”, “speed up”, “increased satisfaction” and “decreased satisfaction”. The codification process started with those codes, while paying attention to what information is also valuable for the impact analysis and is missed by the initial codification. This resulted in discovering five other relevant themes in the data. One important aspect of the cost benefit analysis is the cost involved with correcting the data (“Costs of correcting”), since that is an investment that needs to be made in order to reach high quality data. Additionally, the “most important data” is used for codification since that is one of the questions the participants were asked and might help uncover the most important or worrisome data and problems. Furthermore, in some interviews the “cause” or possible “solution” are mentioned during the conversation, providing insights in what could be costs related to correcting the data quality problems. Apart from information related to these codes, there is also information retrieved that is not directly relevant to the costs or benefits of the data quality, but provide background information or surprising findings that can contribute to a deeper understanding of the data quality problems. Those pieces of information are coded with “extra”. The information relating to each of the codes is put together for all the different interviews to create an overview of all relevant information per aspect of the costs and benefits in order to analyze the impact over all the aspects of the cost-benefit analysis. Table 12 provides an overview of all categories and their definitions.

Table 12 Definitions of codes used for impact analysis.

Code	Definition
Cost increase	Cost increase of doing business caused by poor data quality.
Revenue decrease	Decrease in revenue caused by poor data quality.
Cost decrease	Possible cost decrease as result of improved data quality.
Revenue increase	Possible revenue increase as a result of improved data quality.
Delay	Slowdown in productivity.
Speed up	Cycle time reduction of a process.
Increased satisfaction	Increase in customer, employee or shareholder satisfaction.
Decreased satisfaction	Decrease in customer, employee or shareholder satisfaction.
Costs of correcting	The costs that come with improving the data quality.
Cause	Possible causes for poor data quality.

3.4.2 Interview Results

A summary of the results of the qualitative impact analysis, using codification techniques, is presented in Table 13. The findings are presented per category to provide a comprehensive overview of all the impact of poor versus improved data quality, as identified in the interviews with the different participants. Afterwards, the findings are explained and referenced to the role of the participants that provided the insights.

Table 13 Business Impact of poor data quality per category

Category	Impact
Cost increase	Risk of fines by authorities. Risk of being blocked from sales by authorities. Risk of factory shut down by authorities. Unnecessary extra work due to rework related to reporting activities Flawed decision making - "garbage in is garbage out". False opportunity discovery. Lost opportunities. Material costs for returns of wrongly ordered materials. Stockout, resulting in compensation for customers. Overstock, leading to excessive costs for warehousing and scrapping.
Revenue decrease	Risk of being blocked from sales by authorities. Risk of factory shut down by authorities. Loss of market share due to exposure and lack of trust. Flawed decision making - "garbage in is garbage out". New initiatives built on garbage. Innovations are held back. Loss of sales opportunities (sell-ups and service contracts).
Cost decrease	Decrease in FTE's concerned with rework. Some data cleansing projects can be scrapped entirely. First time right reporting
Revenue increase	Better decisions based on traceability data. Better, more efficient use of resources. More resources focused on adding value to the company. More efficient service organization. Increased commercial growth. Waste reduction (lean. Better margins on service contracts. Increased sales opportunities.
Delay	Time lost on traceability data validation. Long throughput time for reporting due to sanity check & rework of data sets Reduced number of first time fix, resulting in revisits and downtime. Longer throughput time for FCO implementation

	Revisits for service activities.
Speed up	Short throughput time for report generation First Time Fix, resulting in one visit and downtime reduction Shortened FCO implementation time. Eventually time saving for SPS after mechanism is set up.
Increased satisfaction	Customer experience through first time fix, reducing down time of the systems. Customer experience through better quality and service. Philips employees spend no time on reworking & correcting mistakes and consequently can focus on other tasks.
Decreased satisfaction	Delays and quality issues impact customer experience. Authorities are dissatisfied with non-compliance. FSE's have to revisit customers for the same problem. Wrongly ordered materials cost to return. Dissatisfaction amongst employees because of rework and time loss. Mistrust in data. Data source of frustration.
Costs of correcting	Many different departments and people involved. Data, material and process specialists Education and behavioral change First step is turn the tap on dirty data, second step is correcting. Use of remote connectivity for traceability data. Manual audits by FSE's
Cause	Processes - definition, knowledge & understanding, implementation. Manual registration increases risk for human error. Data systems that are not linked increase risk on inconsistencies.
Requirements	Turn the tap on dirty data. Quality Culture Automation. System integration. Training and education. Remote connectivity and manual audits.

Cost increase of doing business caused by poor data quality

One of the main costs concerning the quality of traceability data is the risk of fines by regulatory bodies like the FDA or any other national organization. These fines can be the results of findings during audits and range in amount based of the size and severity of the problems they find and are therefore hard to estimate. In extreme cases however, these organizations can prohibit Philips from selling their product or product ranges in their country, of which the consequences are much more severe and it will lead to a revenue decrease (Participants 2B, 2I, 2L). The reason why these organizations care that much about traceability data is because of quality and safety issues. If a series of a material is unsafe or not performing, this potentially damages the business of customer, or worse, the health of customer or their patients. So if a quality or safety issue occurs, it is important to eliminate the possibility of reoccurrence

as quickly as possible to identify which specific systems are affected by this risk (Units Affected List: UAL) (2C, 2E, 2F, 2G, 2L).

Next to these compliance and safety issues, other costs within the operational and tactical domain (Redman, 1998-a) are created for basing decisions on faulty data. “Garbage in is garbage out” is a principle that is familiar when basing business decisions on flawed data (2F). As a response to the question what are the costs of the current data quality, participant 2D answered “You don’t know what you don’t know”, because to them the traceability data is the truth and therefore has no clue of what dirty data would cost Philips.

The decisions made based on the data vary a lot. An engineer could order the wrong replacement material because the data said another material is in the system than there actually is, which results in having to revisit the customer, losing approximately half a day of his and his customer’s time because of traveling and working on the system, while the market also has to pay return costs (2D, 2F). On the other hand, the productivity manager tries to find new productivity opportunities by means of a business case based on the data. However, if the data is flawed, what seems to be an opportunity could be nothing at all, while true opportunities are never discovered (2D).

Another important example of business decisions based on flawed data, is the Last Time Buy process, which is executed when a material is reaching end of production, while the systems in which these materials are used still need to be maintained for several years. This means that Philips has to decide on how many items of this material they will buy in order to fulfill the demand of replacement material for the coming period of time, without having stockout or too much overstock (2K, 2L). When too much of a material is bought, of course the purchase costs cause an increase in costs, but afterwards the scrapping of those excess materials will drive that up even further. In case of stockout, another option has to be found in order to address the customer needs. This could range from developing a new product, which is a time consuming and expensive process, to providing customers with free upgrades to their system in order to keep them up and running (2K, 2L). According to the director of customer services the lack of ability to relate the demand to traceability data, because of IB inaccuracy, is likely to result in unexpected costs related to unexpected need for additional new production or alternatively costs for scrapping unexpected excess stock.

Revenue decrease caused by poor data quality

As mentioned before in the cost increase, faulty data could lead to serious risks for customers and/or patients. Furthermore, it can result in a serious loss of trust with loss of market share because Philips equipment could be perceived as unreliable (2K), which lead to a significant decrease in revenue (2K).

Apart from these extreme consequences, daily business processes are also impacted that affect revenue in a negative way. Business analyses are used to make decisions (2C). If based on faulty data, this may result in opportunities that are lost, or what seems to be an opportunity turns out not to be and return on investment calculations are flawed, since a UAL is also generated to support the case of productivity ideas (2D). Additionally, not having correct data may result in loss of sales opportunities: Correct data are required to identify customers who could benefit from a service contract or sell-ups to their existing systems (2K).

Possible costs decrease as a result of improved data quality

With regards to decreasing costs, the business analysts unanimously agree that data quality is directly related to time spent in business analytics. The operations manager explains that data quality is positively contributing to first time fix rate. The supply chain is benefited by good traceability data because if they have a good overview of where which system is, combined with failure rates of the materials, they can provide more accurate demand estimations that ultimately lead to reduced stocks and prevent stock outs (2K).

Possible revenue increase as a result of improved data quality

As repeatedly mentioned before, higher data quality leads to better decisions which in their turn can lead to increased revenue or other advantages (2C). Apart from making better decisions, people that are concerned with the traceability data can make full use of their skills as they do not need to rework and resolve mistakes in the data. Hence high data quality is required to achieve a high efficient service organization, allows increased commercial growth and enables more efficient use of resources (2D). The services and IB marketing manager shares that data quality has a direct relation with revenue and financial results: Higher data quality contributes to more accurate calculation of margins on service contracts and positively contributes to optimization of service offerings.

Slowdown in productivity caused by poor data quality

For the business analysts the impact of data quality on efficiency became clear in the previous sections. Lower data quality is likely to require several iterations of checking and/or merging data from several sources. Furthermore the FSE's are delayed if the data is not correct by not being able to fix the customer's problem the first time (2E, 2F, 2I).

Possible cycle time reduction as a result of increased data quality

Higher quality data prevents all the delays as mentioned before. The business analysts' work becomes faster and better (2A, 2B, 2C) as does the process of the UAL creation and FCO implementation (2G). The director confirms the impact of data quality on the cycle time in these processes. Lower data quality slows down the organization and is likely to limit capacity available for continuous improvements. For SPS processes currently generic trending information is used as input. Changing the inputs to being based on detailed traceability data would require initially spending more time and resources on setting up the processes to use material specific traceability data as input.

Increase in customer, employee or shareholder satisfaction due to improved data quality

The main increase in satisfaction that is mentioned by the participants is the customer experience that goes up with the higher data quality. Reasons given for this satisfaction increase are the increased first time fix rate (2D, 2E, 2I) and with that the lower down time. Other increases in satisfaction come from increased safety and quality, which impacts the customers as well as the patients that are treated by the customer (2I). Not only the customers' and their patients' satisfaction is increased, also employee

satisfaction is impacted since higher data quality eliminates (or limits) the work required for correcting mistakes (2D).

Decrease in customer, employee or shareholder satisfaction due to poor data quality

Naturally the returning topics of customer experience due to delays and safety and quality issues return under the decreased satisfaction classification as a result of low quality data (2c, 2E, 2F, 2G, 2I, 2K, 2L). This decreased dissatisfaction however is not only present at customer, patients and authorities, it also concerns the satisfaction of Philips' own employees. Losing time due to inaccurate data that leads to ordering the wrong materials is frustrating (2E).

Extra costs of correcting and improving the data

Between the costs of the low data quality and the benefits of high data quality, there is one more very important factor, the costs required for the journey from low to high data quality. There are several reasons for the high costs involved of the cleaning of the data. First of all a lot of people need to be involved in the process. All people that are involved with the data need to understand the need and cooperate, especially the people that populate the data, like the market organizations (2B, 2C, 2D). Apart from just adjusting the data, first education and behavioral change are required to prevent new data problems to occur.

The root causes of the data quality problems need to be identified and addressed in order to prevent new dirty data contaminating the traceability data again (2L). One option mentioned by participant 2C and 2H is using remote connectivity. The other option to make sure all the data is correct mentioned by several participants, is sending FSE's to completely audit all systems. Although this is an expensive and time consuming option, 2C, 2G, 2I and 2K estimate that performing these audits during already planned maintenance limit the costs and the impact for the customers.

Possible causes of poor quality data

The possible causes that participant mentioned in their answers vary. Lack of process understanding, to not implementing the process or lack of process are mentioned by participant 2C and access to the master data without proper knowledge and training is named as a reason by participant 2B. An often mentioned potential cause of data quality problems is human error.

Requirements for data quality

Apart from preventing or eliminating the causes above, requirements were also mentioned during the interviews. As mentioned in the costs of correcting, turning the tap is the first thing to do, because cleaning the data from one end, while on the other end data problems are still entering the data is a task without end (2B, 2H). Automating processes and linking systems is required to eliminate the chances of human error contaminating the data (2B, 2D, 2G) while the few people remain able to manually alter the data need proper education. (2B, 2D).

After those systems and mechanisms are in place, the cleansing of the data can commence. Part of the data can be solved from a distance by business analysts or the use of remote connectivity (2C, 2H). For actively correcting the data in the field, FSE's need to be scheduled to audit systems during other maintenance activities to be as efficient as possible (2C, 2G)

Aspects of the data seen as most important for business purposes

A remarkable finding is that few of the participants mentioned the AsShipped data as the most important data, except for participant 2G that requires the AsShipped data for UAL creation in the FCO process and participant 2H that only needs the link between AsShipped and Maintained data on the system level. The AsShipped data is not often required for business decisions. Most participant see the Maintained data as the most important data there is for their activities (2A, 2C, 2D, 2K). Furthermore participants 2J and 2K indicate that the consumption and demand are the most important features for their purposes, which means that not just the Maintained but most certainly the SWO data are crucial for success in their areas.

4. Framework

This section provides a framework by comparing and identifying gaps between literature and practice in traceability data quality analysis as an answer to the research question: “How can we guide understanding and managing data quality?” The comparison provides insight that might be valuable for researchers and practitioners interested in traceability data quality analysis an improvement, and suggestions are made that fill the identified gaps. Those suggestions are, combined with the existing literature, incorporated in a cost-benefit analysis framework for improving traceability data quality that provides a step-by-step process that assists in making the decision whether to invest in improving the traceability data quality or not. The steps of the framework are explained below.

4.1 Step 1: Create data landscape

Before assessing the data quality, either as an outsider or insider to the company, it is important to get a broad understanding of the data, the structure and the perceived problems. Existing literature does not cover the process of how to gain an understanding of the data, structure and its problems, which results in the first gap identified in literature. Although it seems not directly relevant, it is an important process. Especially in larger corporations it can be difficult to gather the right information and find the right sources.

Good understanding of the data begins with a good understanding of where to find the information about the data, which is overlooked in data quality research and therefore identified as a gap. As Sebastian-Coleman (2013) explains on metadata, “knowledge of business concepts is the foundation for measurement”. It is essential in understanding the concepts of, and the relationships between data. The data modal and metadata together are required for understanding how the data is stored, maintained and updated and therefore determine how the data needs to be measured. She states that knowledge of five subjects is required to understand the data chain:

- Business concepts represented by the data.
- Business and technical processes that create data.
- Business and technical processes that maintain, update, or delete data.
- The data model in the target system (i.e., where data will be measured).
- Data processing rules for the target system.

Sebastian-Coleman (2013) also mentions that few companies have such metadata documented and available, so then it has to be created as part of the assessment. What is not mentioned in her book “Data Quality Assessment Framework for Ongoing Improvement”, nor in any other literature on data quality, is how to do that. For this study this is done by identifying the start and/or endpoint of the data and following the dataflow. Document all sources there are in the data and which events occur that create, change or delete data. By following the flow of the data, the concepts, processes, data model and the people that are involved become clear. Those people can also be interviewed to provide further information to explain anomalies in the data and confirm the findings about the data landscape.

4.2 Step 2: Select data quality problems.

Step two is a checklist that helps narrowing down to the data quality problems that are relevant for the research that is being performed. The first question of this checklist is "from which perspective do you want to study the data measure the data quality?"

1. *Which aspect of the data quality is of interest, the management aspect, or the more technical database aspect?*

- *Interested in objective measures of the quality of data: Continue*
- *Interested in subjective measures of data quality: AIMQ (Lee, Strong, Kahn & Wang, 2002), Pipino, Lee & Wang (2002)*

Management studies primarily look into subjective measures by providing questionnaires to measure the data stakeholders' perception of the data quality (Pipino, Lee & Wang, 2002), while the studies focused on the technical, database aspect focuses on the quality of data values. This case study focusses on the technical, database perspective and will therefore focus on the objective quality of the data values. Researches or practitioners interested in the more technical and objective measurement can continue to the next step of this checklist, whereas if you are interested in the more subjective, management perspective of data quality, I would recommend the Literature of Pipino, Lee and Wang (2002) and (Lee et al., 2002) since it is not covered in this study.

2. *Which level of data quality problems is of interest?*

- *Single attribute/tuple: Continue*
- *DQ Problems at the Level of a Single Relation: Oliveira, Rodrigues & Henriques (2005).*

Table 14 Data quality problems for objective analysis at the single attribute/tuple level.

Traceability Data Quality Problems *(Oliveira, Henriques & Rodrigues, 2005; Müller & Freytag, 2003)*	Definition
Missing Value	Empty tuple, although a value is required.
Syntax violation	Information in a tuple is not adhering to the syntax rules required for the attribute in question.
Incorrect value	The value of the tuple is not a correct representation of reality.
Domain violation	The tuple value is outside the range of realistic and logical values for that specific attribute.
Violation of business domain constraint	The value is not within the domain that is set by specific business rules for that attribute.
Unique value violation	The tuple contains a duplicate value when not allowed.
Violation of functional dependency	The value of the tuple is dependent on another attribute, but the dependency is violated.

4.3 Step 3: Data quality problem understanding.

Several studies mention the data quality problems and metrics in the technical terms, but there is a gap in literature on how to understand these problems. Not every employee can be expected to be familiar with the data quality problems in literature or is able to identify them as such. Therefore it is important to do interviews, combined with initial data analysis to identify business rules that the data must adhere to. Metadata about the analyzed material(s), the database characteristics and the dataflow is crucial in setting up these business rules. Important to determine is whether these business rules apply to all the data or are specific to parts of the data, which could mean the data needs to be analyzed in separate parts. These business rules can then be labeled on data quality problems and metrics to get an understanding of the data quality status at the company. Even if a data quality problem may not occur at the company at that moment, it needs to be confirmed that the data is correct. For that reason, the set of business rules must be validated by the list of data quality problems to ensure that the data is measured for all possible data quality problems.

4.4 Step 4: Scoping and Analysis

Literature on traceability data quality is scarce. Although there are studies in the fields of Agri-food and software requirements engineering, there is nothing to be found on medical systems, automotive or even aerospace industries, despite the fact that material traceability is required there. Possible reason for that might be confidentiality. Sharing company unique information is always a delicate subject, especially when it can have competitive and compliancy risks. Due to this lack of literature for material traceability specifically, the complexity of the traceability remains unaddressed. High-tech and cyber-physical systems consist of many materials and components in layered structures, and due to this structure and the different nature of these materials have different requirements. This also leads to different requirements for the data of the materials. Therefore it is impossible to analyze the entire dataset at once.

For that reason, it is necessary to address the data of every traceable material separately and customizing the business rules based on the material's features. Some business rules are general and can be reused for every material, but others are material specific. Analyzing the entire database means combining the separate analyses of every material one by one, until all materials are covered. Although this is a very complicated and time consuming process, it is the only way to do it right. After the analysis is performed, the problematic areas can be identified, based on lowest scores or potential impact.

4.5 Step 5: Impact analysis

Since the effects of data quality are mostly indirect, these can't be easily calculated and quantified into monetary terms. Also, they are more difficult to identify. For a complete overview of the impact, semi-structured interviews need to be conducted to determine the costs and benefits of poor quality data and improving the data quality. The costs of poor quality data and improving and assuring data quality are specified by Eppler and Helfert (2004). However, there is one important gap in their taxonomy. One type of direct cost of poor quality traceability data is the compliancy risk for regulated industries like aerospace, agri-food or healthcare. Therefore the risk of fines can be calculated as a cost in the cost-benefit analysis.

Table 15 Taxonomy of data quality costs with addition of compliancy risk..

Data Quality Costs				
Costs **caused** by low data quality		Costs of **improving** or **assuring** data quality		
Direct costs	Indirect costs	Prevention costs	Detection costs	Repair costs
Verification costs	*Costs based on lower reputation*	*Training costs*	*Analysis costs*	*Repair planning costs*
Re-entry costs	*Costs based on wrong decisions*	*Monitoring costs*	*Reporting costs*	*Repair implementation costs*
Compensation costs	*Sunk investment costs*	*Standard development and deployment costs (system and process setup)*		
Compliancy risk				

Loshin (2001) present presents eight categories for the cost-benefit analysis, but forgets one crucial factor in the cost-benefit analysis: the costs of improving and assuring the data quality, which Eppler & Helfert (2004) do categorize in their paper. This cost is the most important one that needs to be outweighed by the benefits before a business is willing to invest in improving data quality.

Adding this to the categories provides a more complete set of categories for the cost-benefit analysis (

Table 16). These different categories can be used to base the questions on for the semi-structured interviews. Make sure that the interview participants consist of people with different professional backgrounds that use different aspects of the data for a wide variation of business processes.

This provides a broad view of the use of the data and therefore the different impacts the data quality can have. After gathering the information from the al participants and writing down the findings, combine the answers from all participants per category to have a more complete view of the areas each category is represented.

Table 16 Cost-benefit analysis categories by Lonshin (2001) complemented with the cost increase for improving or assuring improved data quality.

Category	Definition
Cost increase	The degree of which poor quality data increases the cost of doing business.
Revenue decrease	The effect of low data quality on current revenue.
Cost decrease	The extent wo which improvement in data quality can reduce costs
Revenue increase	The extent to which improvement in data quality increases revenues.
Delay	Measures the degree to which there is a slowdown in productivity.
Speed up	Measures the degree to which the cycle time of a process can be reduced.
Increased satisfaction	Measures the increase in level of satisfaction of the employee, shareholder or customer.
Decreased satisfaction	Measures the decrease in level of satisfaction of the employee, shareholder or customer.
Improving or assuring cost increase	The degree of which improving or ensuring improving the data quality increases costs.

4.6 Step 6: GO / NO GO decision

Now there is a complete overview of all the costs and benefits, it is up to the company to decide whether they think the benefits of improving traceability data quality outweigh the costs. However, for some industries there is one more thing to consider than the costs and benefits: the law. Although the risk of fines is often seen as a calculated risk and therefore can be categorized in costs of poor quality data, mainly in industries where poor data quality can impact the safety of people, the law may demand companies to improve the data quality to the highest possible standard. In that case there is no choice but to comply, since non-compliance can result in authorities shutting down the entire company or blocking sales within countries or continents. This consequence is not documented in literature but is most decisive in making the decision whether to improve the data quality or not. The impact analysis can then aid to determine what other advantages can be gained from improving data quality to maximize profitability or counter the costs.

4.7 Validation of the framework

The framework created in this study is validated by a focus group. The group consists of five data professionals of Philips. During the validation the framework is presented, after which each participant is asked to share his or her opinion on what is missing in the framework or what could be improved. Also, possible implications of the framework are discussed. The results of the validation sessions are presented per session.

4.7.1 Focus group

The participants (

Table 17) of the practical validation sessions are asked to not think as Philips employee during the session, but as an objective data professional and to critically analyze the framework on universal applicability in any company.

Table 17 Participants in validation session.

Participant	Role
V1A	Director Customer Service
V1B	Senior FCO Manager
V1C	Business Analyst 1
V1D	Business Analyst 2
V1E	Business Analyst 3

The group agreed with the framework as proposed, although V1A and V1C and V1D did have a few things to add. Participant V1C had the remark that one could start what is to be expected from the data quality. This is a valid point and is added to the framework in order to see how far the current data quality is of the target level if there is such target available. The participant also made the remark that it must be clear that the business rules are related to metadata, which was also an important finding in this study and is therefore stated more clearly in the framework.

Participant V1A had several remarks which are all included in the framework. The first one being that the data landscape is created by following the data flow. This is how it was done in this study, but in the first version of the framework it was not yet defined properly. Another action that was also taken in the original study, but not well represented in the framework is the assessment whether all data problems are covered after the business rules are created in step 3. This acts as a validation that even when data problems do not occur at this point, they do need to be part of the analysis in case they start occurring in the data. It is also important to measure the things that are going well.

Another comment of the customer service director was that some of the categories in the impact analysis seem like two sides of the same coin and might be condensed into half of the categories. Although it is true that these are often two sides of the same coin, for the purpose of keeping a clear overview, I favor to keep the categories the way they are.

Concerning the Go / No Go decision, both participants V1A and V1D asked what the actual reason was to go, since there will never be a clear ROI. The second question was do we choose to do it, or do we need to do it? Hinting towards the cases were regulations require to have high quality traceability data. Originally, non-compliance was categorized as a direct cost of poor quality data due to the risk of fines. Although that is also a valid classification, since risks are often calculated as costs in business, in some markets traceability is, or will soon be, a hard requirement. For instance the aerospace industry, but also automotive and healthcare industry. Therefore I decided that legislation has a decisive role in the Go / No Go decision, whereas the impact analysis is then a way of identifying the possible opportunities to benefit as much as possible from the investment in improving the traceability data quality.

5. Conclusion

The aim of this study was to provide guidance in managing and understanding data quality and the impact thereof in order to be able to help make the decision whether to invest on data quality. To do so, I started off with creating a data landscape by following the data flow, after which the data quality problems could be determined. Based on interviews and initial data analysis, a set of business rules was created to be able to apply measurements for all the identified data quality problems. Based on the results of the measurements, the impact of the current and improved data quality is analyzed as an aid in making a decision about whether to invest in improving the traceability data quality or not. Afterwards, a framework was created based on these steps in order to provide guidance for researchers and practitioners that need it to gain an understanding of data quality and how to manage it. With this framework the goal of the study is achieved.

In the next sections the research questions for this study are answered, the limitations of the research are discussed and future research is suggested.

5.1 What is the current quality of traceability data?

Since the research project was limited to a pilot which has served as a carrier to define a framework for measurement, the exact current traceability data quality cannot be defined. Taking a subset of the data as a pilot was necessary for several reasons. The complete set of traceability data is too extensive and complex, while every material has its own specific set of business rules which would require too much time to identify and analyze for all materials to fit within this Master thesis project.

5.2 What is the impact of the current and improved traceability data quality, and what can be learnt from it?

The costs and benefits of the current traceability data and improving the data quality are extremely difficult to determine, especially since the current data quality is not determined End to End. Furthermore according to the impact analysis and interviews a corporate Quality Culture is required, while the costs related to establishing and maintaining a Quality Culture has not been identified during the interviews.

The actual costs of improving data could be estimated by calculating how many systems are in the field. Fact is that depending on the actual level of data quality improvement is likely to be expensive, but it is not even possible to give a range with so many uncertain factors. This study confirms findings in existing literature related to the nature of costs often being hidden (Haug, Zachariassen & van Liempd (2011) and the difficulty to estimate of the total cost of poor data quality Redman (1998-a).

5.3 How can understanding and managing data quality be guided?

With this being the main research question, it is answered by providing the framework for measuring and estimating the impact of the data quality. The framework describes how to measure the data quality by creating a set of business rules that allow data analysis for each and every material specifically. Costs and benefits in the impact analysis are not quantified into monetary results, but do serve the purpose of

providing an overview of all possible costs related to poor data quality and the benefits improving the data quality can bring.

5.4 Limitations of this study

The biggest limitation in this study is that the analysis of traceability data quality is only performed for one material. Analyzing multiple materials would provide a better indication of the current data quality. Also, due to necessary scoping to even be able to analyze the data, the results vary from the actual level of traceability data.

Another limitation is that the framework is not tested under different conditions. Every company's data is unique, for which there might be challenges or complications when trying to utilize this framework in different companies within the same or different markets. Lastly, the framework is now only designed for data quality problems related to a single attribute or tuple as described by Oliveira et al. (2005). The other, more complicated are not covered in this study but could be used to extend the framework, adding to the functionality and usefulness of it for other studies.

5.7 Further research

This research is only limited to the data problems of a single tuple or attribute. Oliveira, Henriques & Rodrigues (2005) also identify data quality problems with single relations, multiple relations and data sources. Although these were not required for this study and are even more complex, the framework could be extended with the analysis possibilities for these data problems

Furthermore the quantification of the effects of data quality problems is still a big challenge, especially in relation with the level of data quality. Extensive research on this topic could provide some standards or more accurate guidelines as to what size of effects belong to certain levels of data quality. This could be very valuable for companies by assisting in analysis and decision making with respect to data quality improvement efforts.

Last but not least, the amount of literature available for traceability data in high-tech or cyber physical systems is very scarce. Possible reason for that is confidentiality of company data. More published studies on this aspect would enhance the general understanding of the complexity and mechanisms of material traceability data.

References

Arkley, P. & Riddle, S. (2005). Overcoming the traceability benefit problem. In *13th IEEE International Conference on Requirements Engineering (RE'05)* (pp. 385–389). IEEE. https://doi.org/10.1109/RE.2005.49

Arts, D. G. T., de Keizer, N. F. & Scheffer, G.-J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association*, *9*(6), 600–611. https://doi.org/10.1197/jamia.M1087

Batini, C. & Scannapieco, M. (2016). *Data and information quality - Dimensions, Principles and Techniques*. (M. J. Carey & S. Ceri, Eds.), *Springer*. Springer.

Bosona, T. & Gebresenbet, G. (2013). Food traceability as an integral part of logistics management in food and agricultural supply chain. *Food Control*, *33*(1), 32–48. https://doi.org/10.1016/j.foodcont.2013.02.004

Eppler, M., & Helfert, M. (2004). A classification and analysis of data quality costs. In *International Conference on Information Quality* (pp. 311-325).

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*.

Fisher, C. W. & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management*, *39*(2), 109–116. https://doi.org/10.1016/S0378-7206(01)00083-0

Galhardas, H., Florescu, D., Shasha, D., Simon, E. & Saita, C. (2001). *Declarative Data Cleaning : Language, Model, and Algorithms*.

Haug, A., Zachariassen, F. & van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, *4*(2), 168–193. https://doi.org/10.3926/jiem.2011.v4n2.p168-193

Henry, T. & Glazer, I. (2012). *The Story of Information Sprawl*.

Kim, W. & Choi, B. (2003). Towards quantifying data quality costs. *Journal of Object Technology*, *2*(4), 69–76. https://doi.org/10.5381/jot.2003.2.4.c6

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K. & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, *7*(1), 81–99. https://doi.org/10.1023/A:1021564703268

Laranjeiro, N., Soydemir, S. N. & Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. In *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)* (pp. 179–188). IEEE. https://doi.org/10.1109/PRDC.2015.41

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

Loshin, D. (2001). *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann.

Mader, P., Gotel, O., & Philippow, I. (2009). Motivation matters in the traceability trenches. In *2009 17th IEEE International Requirements Engineering Conference* (pp. 143-148). IEEE.

Madnick, S. E., Wang, R. Y., Lee, Y. W. & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, *1*(1), 1–22. https://doi.org/10.1145/1515693.1516680

McGilvray, D. (2008). *Executing Data Quality Projects*. *Executing Data Quality Projects*. Elsevier Inc.

Mendelow, A. L. (1981). Environmental Scanning--The Impact of the Stakeholder Concept. *ICIS 1981 Proceedings*. Association for Information Systems.

Moe, T. (1998). Perspectives on traceability in food manufacture. *Trends in Food Science & Technology*, *9*, 211–214.

Müller, H. & Freytag, J.-C. (2003). *Problems, Methods, and Challenges in Comprehensive Data Cleansing*.

Oliveira, P., Rodrigues, F. & Henriques, P. (2005). A formal definition of data quality problems. In *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005*.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.Redman, T.C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM - Citeseer*.

Redman, Thomas C. (1995). Improve Data Quality for Competitive Advantage. *Sloan Management Review*, *36*(2), 99.

Redman, Thomas C. (1998). Data Quality for the Information Age. *Communications of the ACM*, *41*(2), 79–82.

Regan, G., McCaffery, F., McDaid, K. & Flood, D. (2012). The Barriers to Traceability and their Potential Solutions: Towards a Reference Framework. In *2012 38th Euromicro Conference on Software Engineering and Advanced Applications* (pp. 319–322). IEEE. https://doi.org/10.1109/SEAA.2012.80

European Union. (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. *Official Journal of the European Union*.

Runeson, P. & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering, 131–164. https://doi.org/10.1007/s10664-008-9102-8

S Sadiq, S. (Ed.). (2013). *Handbook of data quality: Research and practice*. Springer Science & Business Media.

Sattler, K. U. & Schallehn, E. (2001). A data preparation framework based on a multidatabase language. In *Proceedings of the International Database Engineering and Applications Symposium, IDEAS* (pp. 219–228). https://doi.org/10.1109/ideas.2001.938088

Sebastian-Coleman, L. (2013). Measuring Data Quality for Ongoing Improvement. *Measuring Data Quality for Ongoing Improvement (pp. 3-91)*. https://doi.org/10.1016/C2011-0-07321-0

Signify Holding. (2018). Philips Lighting is now Signify. Retrieved April 3, 2019, from https://www.signify.com/en-gb/our-company/news/press-releases/2018/20180516-philips-lighting-is-now-signify

Stahel, W. R. (2016). The circular economy. *Nature News*, *531*(7595).

Sterling, B., Gooch, M., Dent, B., Marenick, N., Miller, A. & Sylvia, G. (2015). Assessing the Value and Role of Seafood Traceability from an Entire Value-Chain Perspective. *Comprehensive Reviews in Food Science and Food Safety*, *14*(3), 205–268. https://doi.org/10.1111/1541-4337.12130

Watts, S., Shankaranarayanan, G. & Even, A. (2009). Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, *48*(1), 202–211. https://doi.org/10.1016/J.DSS.2009.07.012

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering, 1–10. https://doi.org/10.1145/2601248.2601268

Wohlrab, R., Steghofer, J.-P., Knauss, E., Maro, S. & Anjorin, A. (2016). Collaborative Traceability Management: Challenges and Opportunities. In *2016 IEEE 24th International Requirements Engineering Conference (RE)* (pp. 216–225). IEEE. https://doi.org/10.1109/RE.2016.17

Appendices

Appendix A: infographics on Image Guided Therapy Systems

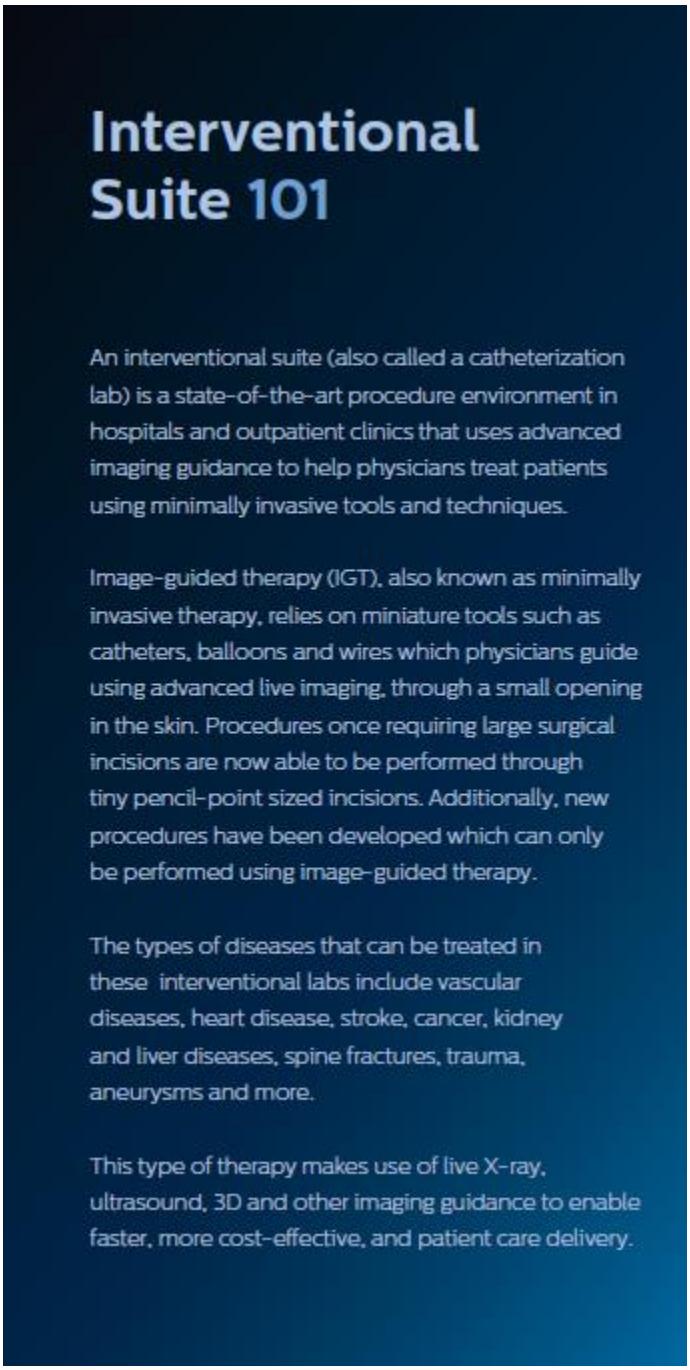


Figure 6 Infographic IGT System (1).

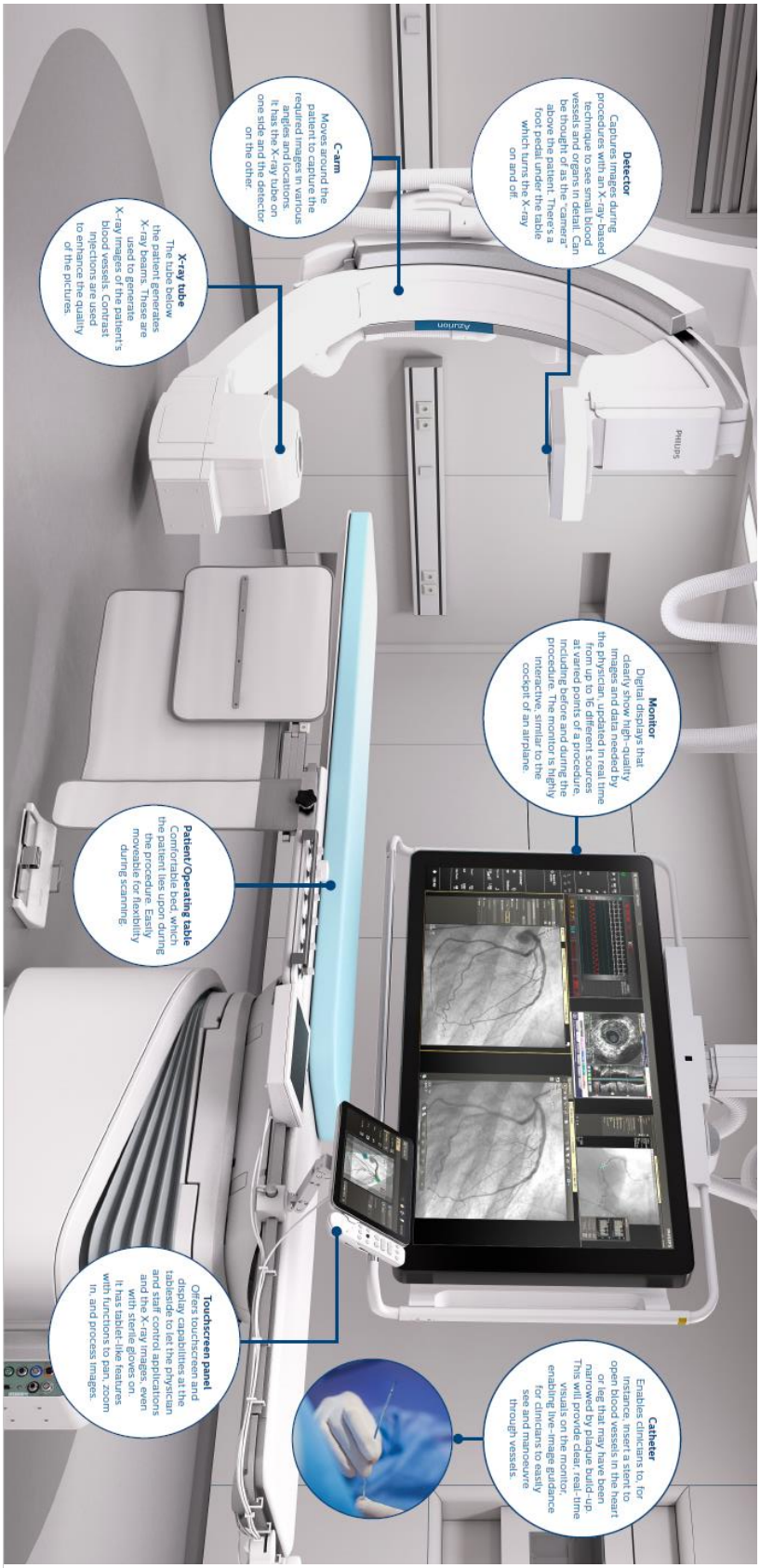


Figure 7 Infographic of IGT System (2).

Figure 8 Infographic IGT System (3).

Appendix B: Merging and Sorting of Traceability Data

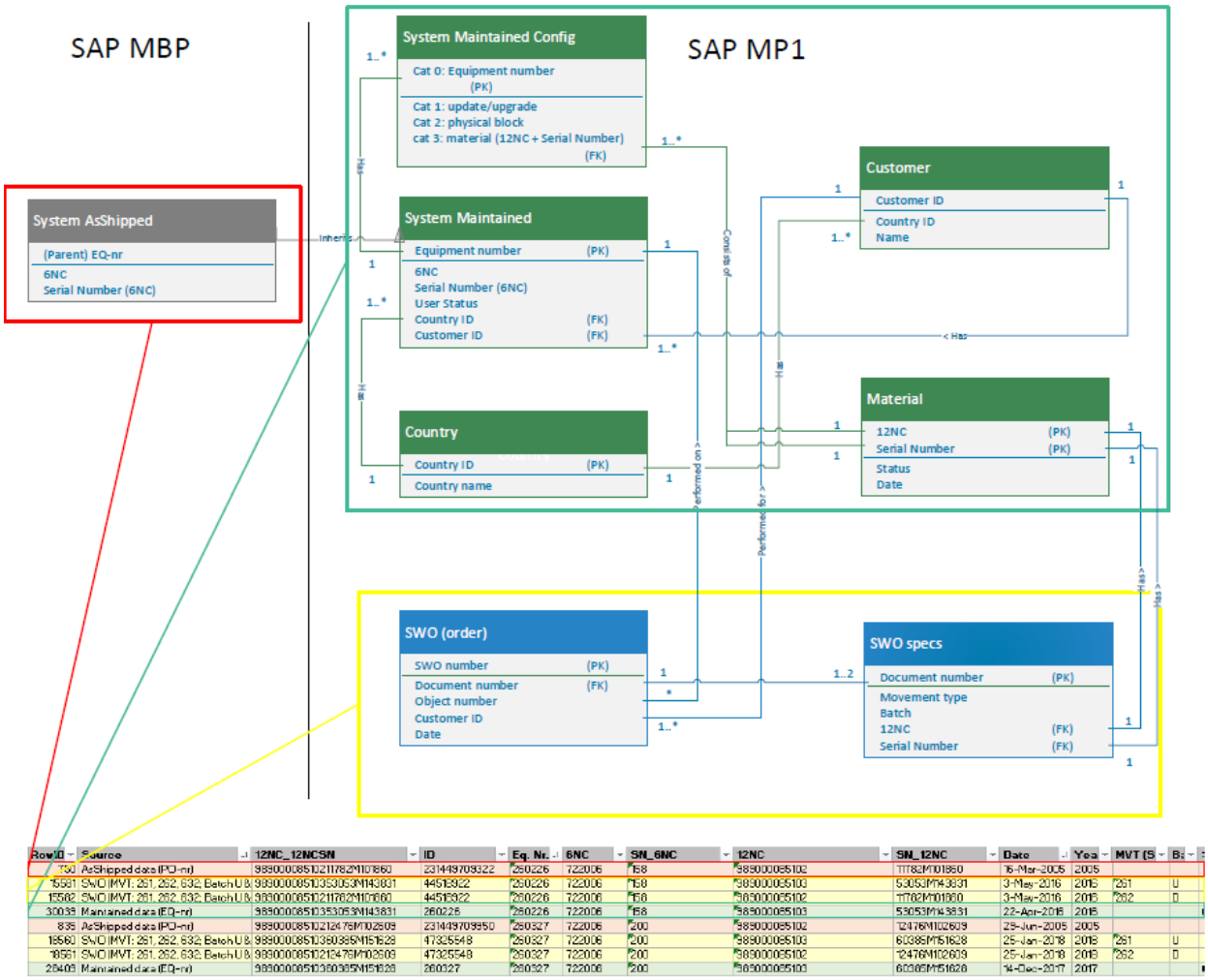
All required entities are collected for all sources. For analysis however, it is required and convenient to put them all together in one sheet. Here it is important to first add all entities for one source and then for the next. When finished, a new attribute “Row ID” is created, starting from 1 at the top and numbering all the way down to the last row of data. Reason for this is to always be able to return to the original setup by sorting on Row ID. For readability purposes the cells with AsShipped data are colored red, SWO data yellow and Maintained data green. Additionally an attribute “ID” is created which is unique for the source the data comes from, to be able to trace back the data in the original datasets. For the AsShipped data this is the Purchase Order Number, for SWO data the SWO number and for the Maintained data it is the (parent) equipment number.

Since a good overview per system is desirable for initial analysis and understanding what happens with the materials in a system, the data in excel is sorted in a specific, hierarchical way. First sorting is done on equipment number, which makes sure all the data of a system is grouped together. Next is sorting on source (or background color since those correspond), with AsShipped (red) sorting on top as this is the origin of the material’s data. Next sorting is done also done on source (color), but now for maintained and restraining it to the bottom. This results in the SWO data falling in between the AsShipped and Maintained data for every system. Ideally the data represent a timeline of the materials per system, with the AsShipped data first, then the first SWO followed by the next, until there are no more SWO data and the Maintained data shows as final row per system. This is achieved by adding a final sorting on the data value, low high. Eventually, the dataset will look like demonstrated in the example of Figure 9, where the material is clearly visualized per equipment number.

RowID	Source	12NC_12NCSN	ID	Eq. Nr.	6NC	SN_6NC	12NC	SN_12NC	Date	Yea	MVT (S	Ba
750	AsShipped data (PO-nr)	98900008510211782M101860	231449709322	260226	722006	158	989000085102	11782M101860	16-Mar-2005	2005		
15581	SWO (MVT: 261, 262, 632; Batch U &	98900008510353053M143831	44518922	260226	722006	158	989000085103	53053M143831	3-May-2016	2016	261	U
15582	SWO (MVT: 261, 262, 632; Batch U &	98900008510211782M101860	44518922	260226	722006	158	989000085102	11782M101860	3-May-2016	2016	262	D
30039	Maintained data (EQ-nr)	98900008510353053M143831	260226	260226	722006	158	989000085103	53053M143831	22-Apr-2016	2016		
835	AsShipped data (PO-nr)	98900008510212476M102609	231449709950	260327	722006	200	989000085102	12476M102609	29-Jun-2005	2005		
18560	SWO (MVT: 261, 262, 632; Batch U &	98900008510360385M151628	47325548	260327	722006	200	989000085103	60385M151628	25-Jan-2018	2018	261	U
18561	SWO (MVT: 261, 262, 632; Batch U &	98900008510212476M102609	47325548	260327	722006	200	989000085102	12476M102609	25-Jan-2018	2018	262	D
28409	Maintained data (EQ-nr)	98900008510360385M151628	260327	260327	722006	200	989000085103	60385M151628	14-Dec-2017	2017		

Figure 9 Part of data representation.

Essentially, the data and entities as described in our UML class model are now all merged together in one large dataset. Figure 10 further illustrates how the class model relates to the dataset.



Row ID	Source	12NC_12NCSN	ID	Eq. Nr.	6NC	SN_6NC	12NC	SN_12NC	Date	Year	MVT (S	Ba
750	AsShipped data (PO-nr)	989000085102117821M101860	2314497093322	260226	722006	158	989000085102	11782M101860	16-Mar-2005	2005		
15581	SWO IMVT: 261, 262, 632; Batch U &	989000085103530S3M143831	44513922	260226	722006	158	989000085103	53053M143831	3-May-2016	2016	261	U
15582	SWO IMVT: 261, 262, 632; Batch U &	989000085102117821M101860	44513922	260226	722006	158	989000085102	11782M101860	3-May-2016	2016	262	D
30038	Maintained data (EQ-nr)	989000085103530S3M143831	260226	260226	722006	158	989000085103	53053M143831	22-Apr-2016	2016		
836	AsShipped data (PO-nr)	989000085102124761M102609	2314497093950	260327	722006	200	989000085102	12476M102609	25-Jun-2005	2005		
18560	SWO IMVT: 261, 262, 632; Batch U &	989000085103603S5M151628	47325548	260327	722006	200	989000085103	60385M151628	25-Jan-2018	2018	261	U
18561	SWO IMVT: 261, 262, 632; Batch U &	989000085102124761M102609	47325548	260327	722006	200	989000085102	12476M102609	25-Jan-2018	2018	262	D
28409	Maintained data (EQ-nr)	989000085103603S5M151628	260327	260327	722006	200	989000085103	60385M151628	14-Dec-2017	2017		

Figure 10 UML Class Diagram linked to data representation

Appendix C: Data gathering queries in SAP MBP and MP1

The AsShipped data is relatively easy to gather, running a query in SAP MBP, table IH10 for the required 12NC's in the material box and selecting the right entities for those 12NC's returns all the data required. For the maintained data the procedure is the same, except this time the query is ran in SAP MP1. Within IH10 in SAP MP1Use Excel's VLOOKUP function to fill the parent equipment number column by looking up the values of the lower equipment number and the data is ready for use.

The SWO data queries are all ran in SAP MP1. Here the 12NC's are entered in the material number box, movement type 261, 262 and 632 are entered in the movement box. As final action : put the data together in one sheet by using Excel's VLOOKUP function.

Appendix D: Formulas used for traceability data analysis

Table 18 Formula's used in Excel for Data Analysis.

Custom Attribute	Formula
Countif maintained per EQ-nr (rule 1&3)	*=COUNTIFS(E:E;E2;B:B;"main*")
Countif Shipped per EQ (Rule2)	*=COUNTIFS(E:E;E2;B:B;"AsShipped*")
Countif SWO per EQ (Rule3)	*=COUNTIFS(E:E;E2;B:B;"SWO*")
IF EQ has Maintained and Shipped(Rule3) 1=yes	*=IF(COUNTIFS(1:1;B1;2:2;">0";3:3;">0");1;0)
IF EQ has only Maintained(1) OR AsShipped(2), none or both (0)	*=IF(COUNTIFS(1:1;B1;2:2;">0";3:3;"0");1;IF(COUNTIFS(1:1;B1;3:3;">0";2:2;"0");2;0))
6NC validation	*=IF(SUMPRODUCT(--(N2=Rule5!$J$1:$J$48))>0;"valid 6NC";"invalid 6NC")
Tubes maintained in system	*=COUNTIFS(E:E;E2;B:B;"main*")
Latest date (top-down)	*=IF(E2=E1;IF(NOT(OR(LEFT(B2;4)="Main";RIGHT(R2;1)="2"));MAX(O2;X1);X1);O2)
Latest date bottom-up	*=IF(E2=E3;Y3;X2)
Flag most Recent(261)	*=IF(AND(Y2=O2;NOT(RIGHT(R2;1)="2");NOT(LEFT(B2;4)="Main"));"most Recent";"")
Maintained data correctly updated (1=yes)	*=IF(LEFT(B2;4)="main";COUNTIFS(E:E;E2;Z:Z;"most recent";C:C;C2);"")
Maintained lines per EQ	*=COUNTIFS(E:E;E2;AA:AA;"0")+COUNTIFS(E:E;E2;AA:AA;"1")
Count 261 per EQ (Rule 9)	*=COUNTIFS(12:12;B12;30:30;"261")
Count 262 per EQ (Rule9)	*=COUNTIFS(12:12;B12;30:30;"262")
261:262 violation (Rule 9)	*=IF(B13>B14;B13-B14&" inbound > outbound";IF(B14>B13;B14-B13&" outbound > inbound";""))
last inbound	*=IF(E2=E1;IF(OR(LEFT(B2;3)="AsS";RIGHT(Y2;1)="1");MAX(V2;AJ1);AJ1);V2)
last inbound	*=IF(E2=E3;AK3;AJ2)
Count same inbound 12NC_SN per EQ-nr	*=IF(Y2="262";COUNTIFS(E$1:E1;E2;C$1:C1;C2;Y$1:Y1;"<>*2");"") & IF(LEFT(U2;5)="valid";"";" X")
Should not be in Maintained	*=IF(LEFT(B2;4)="Main";COUNTIFS(E$1:E1;E2;C$1:C1;C2;Y$1:Y1;"*2");"") & IF(LEFT(U2;5)="valid";"";" X")
SN_12NC check	*=IFERROR(IF(AND(LEFT(T2;5)=TEXT(VALUE(LEFT(T2;5));"0");RIGHT(T2;6)=TEXT(VALUE(RIGHT(T2;6));"0");OR(MID(T2;6;1)="M";MID(T2;6;1)="-"));"OK";"NOK");"NOK")
SN_6NC check	*=IFERROR(AND(P2=TEXT(VALUE(P2);"0");P2<>E2;LEN(P2)<5);FALSE)
EQ-nr Dupe Check	*=IF(COUNTIF(E:E;E2)=(I2+J2+K2);"OK";"NOK")
6NC + SN check	*=IF(AND(O2="valid 6NC";Q2="valid");"OK";IF(AND(O2="valid 6NC";Q2=

	"Invalid");"Invalid SN_6NC";IF(AND(O2="invalid 6NC";Q2= "valid");"invalid 6NC";"Both invalid")))
12NC corresponding	*=IF(LEFT(B2;4)<>"Main";IF(LEFT(B2;3)="AsS";IF(OR(R2="989000086501";R2="989000086502");"correct tube";IF(AND(X2<=2014;OR(R2="989000085101";R2="989000085102";R2="989000085103"));"correct tube";"incorrect tube"));IF(Y2="261";IF(OR(R2="989000085101";R2="989000085102";R2="989000085103");"correct tube";"incorrect tube");"correct tube"));"")
1	*=IF(LEFT(B2;4)="Main";IF(N2=722044;IF(I2<>2;1;"");IF(I2<>1;1;""));"")
3	*=IF(LEFT(B2;4)="Main";IF(M2=1;1;"");"")
5	*=IF(O2="invalid 6NC";1;"")
7	*=IF(AH2<>"";IF(AH2<>1;1;"");"")
8	*=IF(H2<>"";1;"")
9	*=IF(Y2="262";IF(AL2<>"1";1;"");"")
10	*=IF(LEFT(B2;4)="Main";IF(AM2="0";"";1);"")
11	*=IF(S2=AA2;"";1)
12	*=IF(AN2="NOK";1;"")
13	*=IF(AQ2<>"OK";1;"")
14	*=IF(AR2="incorrect tube";1;"")
Violations/line	*=SUM(AS2:BB2)
Violations/system	*=SUMIF(E:E;E2;BD:BD)
IB affected directly	*=IF(LEFT(B2;4)="Main";SUM(AS2;AU2:AV2;AY2:BB2);"")
IF Maintained affected	*=IF(LEFT(B2;4)="Main";IF(BF2<>0;"Systems with affected maintained data";"Systems with clean maintained data");"")
Maintained clean/affected	*=IF(E23579=E23580;BL23580;BH23579)