Turun yliopisto
University of Turku

Tilburg University

IAE AIX
Aix-Marseille Graduate
School of Management
Aix*Marseille Université

DATA QUALITY CHALLENGES IN NETWORK AUTOMATION SYSTEMS

Case Study of a Multinational Financial Services Corporation

Master's Thesis in
International Master of
Management of Information
IMMIT

Authors/Laatija(t):
Jan van Roozendaal
ANR 631888
j.l.c.f.j.vanroozendaal@uvt.nl

Supervisor/Ohjaajat:
Dr. Hans Weigand
Dr. Sc. Hannu Salmela
Dr. Sc. Patrick Rousseau

01.06.2016
Brussels

Turun kauppakorkeakoulu • Turku School of Economics

PREFACE

With the writing of this thesis comes an end of a very pleasant and rich personal learning experience. For the last two years I have placed myself into new situations which have helped me to grow as a person. Getting out of my comfort zone and studying abroad has been a choice I will never regret. Being granted the opportunity to study at three European universities to pursue multiple Master's degrees is something some people within my personal network would not have expected beforehand. I have perceived the IMMIT program as the best way to conclude my educational track, and looking back at it I believe I have reached the maximum achievable.

First and foremost I would like to thank my parents for their continuous support and love throughout the two years of the IMMIT program. All the Skype sessions we had while abroad have helped me to feel at ease when things were getting a bit too much. I also would like to thank my sister and her fiancé for their support and their continuous teasing; graduating from a triple-degree program does not guarantee winning in Rummikub, Trivial Pursuit or TV-quiz shows. There is also a family somewhere near the city of brotherly love I would like to thank for their support as well during the IMMIT experience, especially *A* who has always been an excellent motivating person.

Of course, I would like to thank my supervisors both on-site and at the University for their efforts and guiding me during the writing process of the thesis. I also would like to thank the company where I could do my internship for having their faith in me and allowing me to gain my first couple of months of professional experience, along with my kind colleagues.

I hope that the staff of the elementary school from eighteen years ago will somehow read this message. My parents have told me multiple times a story from when I was still a toddler. The headmaster at the time said the following to my parents: either I was very intelligent, or I was a complete bonehead. Imagine how difficult it would be as a parent to respond to that statement. I do not like to call myself intelligent, but I am glad to conclude with confidence that the latter has been disproven. With perseverance, not being afraid to take on a new adventure, and knowing that there are always people to support me, sometimes in unexpected ways, you can achieve more than you think, because I have done just that.

The lions were coming. And again George Hadley was filled with admiration for the mechanical genius who had conceived this room. A miracle of efficiency selling for an absurdly low price. Every home should have one. Oh, occasionally they frightened you with their clinical accuracy, they startled you, gave you a twinge, but most of the time what fun for everyone, not only your own son and daughter, but for yourself when you felt like a quick jaunt to a foreign land, a quick change of scenery. Well, here it was!

…

"You've let this room and this house replace you and your wife in your children's affections. This room is their mother and father, far more important in their lives than their real parents. And now you come along and want to shut it off. No wonder there's hatred here. You can feel it coming out of the sky. Feel that sun. George, you'll have to change your life. Like too many others, you've built it around creature comforts. Why, you'd starve tomorrow if something went wrong in your kitchen. You wouldn't know how to tap an egg. Nevertheless, turn everything off. Start new. It'll take time. But we'll make good children out of bad in a year, wait and see."

"But won't the shock be too much for the children, shutting the room up abruptly, for good?"

"I don't want them going any deeper into this, that's all."

The lions were finished with their red feast.

Extract from "The Veldt" (1950) by Ray Bradbury (1920-2012)

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

The following is a list of frequently used abbreviations and the section number for when they are first introduced.

[1.2] **BYOD**	Bring Your Own Device
[1.2] **IoT**	Internet of Things
[1.2] **IPv4/6**	Internet Protocol Version 4/6
[1.2] **DDI**	Combination of **D**NS, **D**HCP, and **I**PAM
[2.2.1] **CWA**	Closed World Assumption
[2.2.1] **OWA**	Open World Assumption
[2.2.4] **CFD**	Conditional Functional Dependency
[2.3] **KPI**	Key Performance Indicator
[2.3] **KGI**	Key Goal Indicator
[2.3] **CobiT**	Control Objectives for Information and Related Technology
[2.4.1] **DNS**	Domain Name Server
[2.4.2] **DHCP**	Dynamic Host Configuration Protocol
[2.4.2] **TCP/IP**	Transmission Control Protocol / Internet Protocol
[2.4.2] **LAN**	Local Area Network
[2.4.2] **MAC**	Media Access Control
[2.4.3] **IPAM**	IP Address Management
[2.4.4] **(G)UI**	(Graphical) User Interface
[2.6] **NDS**	Network Discovery System
[2.6] **ICMP**	Internet Control Message Protocol
[2.6] **SNMP**	Simple Network Management Protocol
[2.7.2] **PNA**	Passive Network Appliance
[2.7.2] **BPMM**	Business Process Maturity Model

1 INTRODUCTION

The first chapter of this thesis will introduce the research topic, its problem indication and its context to ultimately provide motivation as of why this topic has been chosen for further investigation. It has to be identified which stakeholders could potentially hold this problem and, by answering a defined research question, how it can be solved or how improvements can be made to respectively resolve or mitigate the problem. After defining the research question, further theoretical background is required to gain a better understanding of the topic at hand to provide a complete answer to the main research question. This is provided in the second chapter via an academic literature review.

1.1 Introduction & Problem Indication

The Oxford Dictionary defines *control* as the ability to manage a machine or object, the power to influence behavior of people or entities during given events, or the ability to limit or restrict a certain activity. People desire to be in control, sometimes proactively when placed into a new environment such as a new job (Ashford & Black, 1996). It is not strange to think that organizations also desire to be in control for all activities they are involved in to maintain their market position in their respective industries. But rather than attempting to be in control in new, unexplored territory, people and organizations have to *maintain* control over their activities in their current, known environment as well. A method that organizations use to reassure that they are in control of their activities is *auditing*. However, similar to how a student may be nervous for an exam result, organizations may be nervous about possible compliance issues as it could put pressure to both management and reporting teams, and could even lead to re-organization (Ghose & Koliadis, 2007). *Monitoring* current activities should give organizations assurance of their current status concerning performance, security, or financial reporting, to name a few. Consulting the Oxford Dictionary once again, it can be described as keeping a continuous record or log about something, or observing whether an activity or process is behaving correctly or reasonably. Similar to how a watchman on a tower would observe whether intruders would arrive or keep track of who is arriving and departing, certain employees could be assigned to assess whether in the Internet network infrastructure any suspicious behavior is occurring.

There may however be a problem with such monitoring practices, and it is the limitation of making judgment solemnly on what we can actually ‘see’. For the watchman on the tower, it would be the limitation of visibility over a given geographical space; for a network administrator, it would the amount of data it can fetch given a particular network device, entity, or system. Now comes the judgment of what they perceive: should it raise any alarms or is it of no concern? For a watchman, it could be that the colors or sigil on a banner indicates whether an outsider of the perimeter is friendly or an enemy. Another indication could be the number of persons to assess the level of threat or whether or not artillery weapons are present. Therefore, the *size* of the group and their *colors* or *sigils* could be seen as possible *dimensions* a watchman can use to judge over the given activity, defined as *features* of a given situation by the Oxford Dictionary.

A network operator has to monitor activities within a given infrastructure; however, instead of a large open field where artillery weapons could be driving about, the dimensions applicable to monitoring are not tangible, but abstract. Network monitoring, and the assessment of whether an alarm should be raised or not, is dependent on the data streams coming in. But how can network operators deduce from the data when an unwanted event is partaking? This starts with the actual *quality of data*, and thankfully, many researchers have discussed this topic (e.g. Fan & Geerts, 2012; Wang & Strong, 1996; Fisher & Kingma, 2001; Pipino et al., 2002; Liebchen & Shepperd, 2008; Wand & Wang, 1996). The researchers are in agreement that data quality can also be broken down in abstract dimensions, but some differences exist in which dimensions should be taken into consideration.

To give examples of such dimensions, the values that can be found in a datasheet need to *accurately represent* the entities they are describing, and that all attributes of these entities are considered to be taken into consideration in the datasheet, often referred to as *completeness*. Another example could be that the data needs to be relevant to the current status of entities, meaning that data cannot be *outdated* and is relevant to current business practices. The combination of these data dimensions, and possibly of additional data dimensions, constitute the overall data quality the network operator can work with.

But for now, we have only described the comparison of a watchman versus a network monitoring tool for when there is only a single one present. It would be logical to have multiple watchmen at different towers in order to cover a larger geographical area. But then there is the challenge of information asymmetry: not every watchman perceives the same thing. Whenever one watchman has to raise an alarm, a communication system should be in place to notify other watchmen and residents of an incoming attack, such as a bell. This way, all people who could be in possible danger are given the same information, it being that they are under attack. However, not all involved people know exactly what the perceived dimensions of the danger was; only the first watchman could still be the only one knowing the finer details at the time.

When multiple network monitoring tools are installed throughout the network interface, the challenge remains the same: whenever one monitoring tool raises an alarm, other monitoring tools should be made aware that suspicious activity is going on. Though, the initial question should be if monitoring tools know from each other what they are monitoring. In other terms, this means whether monitoring tools can agree on the data that is collected via some form of integration. In the case of a disagreement, it could be that multiple monitoring tools are assigned to collect data from the same entity, possibly with different datasheets as a result. There is therefore ‘no single truth’ on the status of the network infrastructure in question. Another possibility could be that one monitoring tool is dependent on the produced datasheets from another tool. If certain data quality dimensions within that particular datasheet are underperforming, the quality of the overall monitoring is already negatively affected.

So, a question organizations may have is how data quality could be improved so that produced network monitoring results are more reliable. The problem does not stop here however as the psychical network infrastructure, such as physical pieces of hardware, cabling, and end-user devices such as PCs, is now accompanied with virtual or cloud network infrastructures which, as the term *virtual* would suggest, are intangible. The watchman in the tower would suddenly be out of its elements: how would it be possible for him to watch over intangible entities? Organizations may struggle to have visibility over their entire network as it could be partially physical and partially virtual. So, instead of stating that the watchman has to look out for intangible entities, we could replace it by visualizing how the visibility of the watchman could be hindered due to a large hill blocking the view. One could suggest to build another tower at the top of the hill for improved visibility to look over the horizon, such as how network monitoring tools are continuously developed to improve end-to-end visibility throughout the network infrastructure.

The purpose of the thesis is to investigate on the challenges of monitoring paired with data quality and the different types of network monitoring practices currently in place. An action plan has to be developed for a particular organization in order to communicate clearly to other groups of stakeholders not only *how* to possibly improve on data quality for network monitoring practices, but also *why* it is necessary to be aware of current data quality performance during data collection phases. Only then will organizations have a better sense of control over their current activities occurring within their environment, and are they possibly more likely to be in compliance with mandatory, established frameworks.

The following sections will describe the research context and the problem statement in more technical detail, and to whom the research of the thesis could be useful to identify groups of problem owners for this particular topic. Once this has been described, a research question needs to be defined which will dictate the flow of the thesis: which topics are going to be discussed in the theoretical background section, which elements are required to provide a complete answer to the research question as possible, how can this insight be applied for the designated case study, and how should conclusions be presented.

1.2 Problem Statement / Research Context

The purpose of the problem statement is to present the possible knowledge gaps within the topic. This will allow for a better understanding of the reasoning regarding the selection of the research question by inspecting existing theories to justify why the research question is of relevance and importance. Maintaining a consistent high level of data quality is becoming an increasing challenge as new developments such as BYOD (Bring Your Own Device), IoT (Internet of Things) and the rollout and increased use of IPv6 addresses allow for larger data pools to be created (Kind et al., 2008). Furthermore, network environments have been expanded from traditional, physical hardware tools providing the network to virtualized infrastructures and (hybrid-) cloud practices. Because of this increase of network environment complexity, achieving end-to-end network visibility is also becoming an increasing challenge. Being heavily dependent on network monitoring practices to be run via human input is more likely than ever to cause errors on the long-term without substituting human input with automated processes. Also, the introduction of a more visualized tool for monitoring should cut down time on network measurement data interpretation and allow for more conscious requests for network configuration changes.

Thankfully, such visualization tools are available with the introduction of Infoblox DDI and NetMRI. Lee et al. (2014) speak of five monitoring operations layers that create the workflow from network measurement data collection to analysis and data presentation. While this may seem that the introduction of such tools are the solution for increased network visibility, reduced time needed for data interpretation, and increased understandability of the network model, improving data quality is not as straightforward as it may seem.

One of the challenges of the thesis is to understand the importance of data quality on network monitoring practices and how this may possibly be improved. However, the question is whether improving data quality also translates into improving the accuracy of representing the so-called network model. For this, it is important to decide the dimensions of data quality to be considered in the thesis. Dimensions have been selected by analyzing the primary work of Fan & Geerts (2012), Wang & Strong (1996), and Wand & Wang (1996), to name a few. Dimensions derived from the literature review are data consistency, completeness, and accuracy, amongst others. So, a question could be whether altering network monitoring practices by focusing on improving on the data accuracy dimension lead to improved accuracy of the network model. It is important to understand which data dimensions are considered essential for the practices of network monitoring in order to provide a consistent definition of data quality throughout the thesis.

Understanding the possible trade-off between accuracy and flexibility regarding network monitoring can be explained with the following example: focusing on increasing the accuracy of network measurement data can be achieved by dedicating more machines to fetch such data; however, the compromise may be that it may take longer for such types of analysis to become available as each measurement may be more time-intensive than before. Another possibility could be that these machines support a smaller variety of network analysis types. As a researcher, it is important to consider the trade-offs between accuracy, efficiency and flexibility when presenting suggestions for improved data quality on network monitoring practices.

Finally, let us think that with the use of a network monitoring tool, a consistent view of network data should be achieved among multiple work groups. Combining this with the suggestions from researchers that network monitoring practices should be more automated, this becomes an interesting scenario in terms of data and security governance. Both Lee et al. (2014) and Reddy et al. (2014) acknowledge how improved techniques on discovering devices on a particular network and improved data structures may strengthen the position of the organization in terms of IT security compliance. However, how does the introduction of a network monitoring tool impact data governance practices? Who can access the network model representation? Who can zoom in on segments of a network for closer inspection? Who can access the network measurement data and request new configuration changes? While governance is not the main focus of the thesis, the topic of data governance is briefly described in the appendix using Wende’s (2007) model (see Appendix A).

1.3 Relevance of research to Problem Owners

As a researcher, it is important to relate to the stakeholders who can be perceived as the problem owners of this particular case. How can one communicate in an understandable way the impact of data quality and possible suggestions for improvement keeping in mind that not every stakeholder may share the same expertise? In order to achieve this, and being relevant to the Master’s, when discussing data quality and automated network monitoring practices, it should be done in a more holistic view for better mutual understanding between different groups of stakeholders. Stakeholders need to understand why they should be concerned about network data quality and how they can benefit from increased data quality. Increased operational efficiency and lowered operational expenditures are two examples of possible effects of improved data quality. However, from an enterprise governance prospective, stakeholders can enjoy the benefits of increased visualization of their current IT security and data governance status for upcoming audit checks. This allows for reduced time spent on checking current security practices and whether or not violations are made with compliance standards (e.g. SOX, HIPAA, PCI).

Several researchers (e.g. Batini et al., 2007; Eppler & Helfert, 2004) have discussed the possible types of costs associated with low data quality, and how trying to achieve perfect data quality may not only be realistically impossible, but cost-wise a very poor strategy. Eppler & Helfert (2004) discuss how total costs for data quality could be broken down in a simplified manner where preventive, corrective, and detecting cost of data errors paired with repairing costs can be showcased in cost curves. An optimum point in the total cost curve can be identified using predictive cost calculations, and the goal should be to enjoy both a adequately high level of data quality while ensuring that the lowest total cost point is achieved. Of course, this optimum point can change when preventive, corrective or detective practices on data quality are improved upon or automated, for instance by introducing a network monitoring tool. The concern of the researchers however is that cost analysis of data quality methodologies for improvement is often neglected, and that the proposed solution may be beneficial in terms of data quality, but less so in financial terms. Stakeholders can enjoy from the insight these authors have given; for instance, when developing a solution for gradually improving data quality within a given sector or department, the pitch of proposing such solution can be strengthened by considering and clearly identifying the financial impact of when data quality would remain poor, and the level of investment needed for data quality to gradually improve.

Finally, this thesis may provide value for those types of stakeholders who are not technically knowledgeable but would like to feel more comfortable when being involved in network operations discussions regarding data quality and network monitoring practices. As this thesis is aimed for a broader audience, not all aspects of network technology are discussed in much detail. Instead, a more comprehensive, narrative form is used. It is a challenge to make sure all stakeholders involved are on the same page and that each of their concerns, whether in the field of security, finances, or automation, are all given enough attention. The thesis aims to address each of these fields in order to give a global, holistic view of network monitoring and data quality improvement practices.

However, when discussing the benefits of the research towards stakeholders, the existing challenges amongst stakeholders should also be acknowledged. Not all stakeholders may hold the same sense of awareness regarding the importance and potential benefits of improving data quality on network monitoring practices. The challenge may already start at the term ‘data quality’: do all stakeholders have the same understanding of the definition, or is it interpreted differently by stakeholders? Are stakeholders using similar terms interchangeably without realizing the difference between the two? Stakeholders may have a narrow mindset regarding the idea that may be pitched to them, as they would like for their department to gain the most benefits or for their preferred approach to be taken for certain practices.

Another challenge to be addressed is the internal credibility issue: are all relevant stakeholders aware of the proposed changes, or are they excluded to contribute to improve on these changes? Are there unanimous results all stakeholders would like to achieve, or do they differ as well? Asking this question is needed to gain an idea on when stakeholders define a project as a success given their own interests; if stakeholders can agree on a mutual key performance indicator, it simplifies the assessment of considering the project to be a success or failure. Finally, stakeholders may show a lack of commitment along the timeline of the project, which can cause future financial issues regarding funding and budgeting. Lack of commitment from a stakeholder would not only indicate loss of interest in the project to financially invest in, but it also hints that the project will be a failure in general, possibly due to the given idea that not all wishes of the stakeholders are being accounted for.

1.4 Proposed Research Question

In order to define a proper main research question, it would be advisable to use the six properties of Foss & Waters (2007) as guidelines for defining a complete yet well-defined research question for the thesis. Before each property is discussed, it would be best to present a first, preliminary research question in order to showcase its development towards a more fitting one. The initial research question is the following: *How can data quality be improved within centralized network management systems to achieve more automation?*

The first property of Foss & Waters (2007) is *theoretical construct*: does the question accurately show in words what the researcher wants to learn about? Other people may define this property as *relevance,* as a question that is irrelevant to the described theoretical background and further research has no value for the research. In this case, we could say that the topics of data quality improvement and centralized network management systems are relevant for the research to be conducted for the thesis. However, it should be known whether the relation between the two topics is depicted correctly. This means, are we interested in data quality improvement within the scope of network management systems, or do we mean how network management systems have the potential to improve data quality? To clarify, the latter question would be infeasible to answer since network management systems depend on the data they are retrieving. Thus, the focus of the problem should be on improving data quality in general, and then relate this practice to possible challenges or benefits for network management systems.

The second property is *recognisability*: does the research question indicate which literature needs to be reviewed in a logical structure? The structure that can be followed for the theoretical background chapter of the thesis could be that it should be known how academic literature defines data quality, why it is considered important since the question indicates that improvement of data quality is to the organization’s interest, how it can be improved, and finally how centralized network management systems work. The latter may require a further breakdown into the practice of network management and which data is related to network management systems. It would seem that the initial research question does showcase a logical structure for literature review; however, it is not as specific. The initial research question indicates that centralized network management systems allow for those relevant problem-owners to identify not exactly *what* data needs to be improved, or *where* data quality may be improved upon, but simply *how*. This means that by increasing our understanding of both topics, as a researcher the task is to answer which functionalities or benefits centralized network management systems can be used to improve on data quality. Recognisability is also related to whether terminology is used similar to what experts in the field would prefer to use. Does the use of such terminology paired with the theoretical construct presented describe the research area of interest accurately?

The research question should also *transcend the data* to be used for the purposes of conducting the research. The research question should be answered, and the researcher should be able to deduce which method would be suitable to answer the research question. It would seem obvious that in this scenario having experience with using centralized network management systems would be useful to give a constructive answer. Lacking such experience would not suggest however that the question cannot be answered. Using a retrospective case study approach related to these topics could help the researcher to formulate a possible answer to the research question. In this case, a prospective case study approach is used in combination with the insight gained from performing a thorough literature review in order to answer the question based on newly gained experience. The research question should therefore allow the researcher to predict which data might be useful to answer the question or solve the problem at hand.

This brings us to the fourth property regarding a well-defined research question, *significance*. It is important to consider whether the research question allows for new insight within the field of research, for instance from another perspective. Given the initial research question, it could be that many authors have already discussed on methodologies to improve on data quality. Though, how many authors have narrowed down its scope to network management practices? Has the combination of data quality with network management systems ever been used before? Now, while this may hint that the research question allows for *originality*, it should not be considered to be the same as *significant*. There may be a reason why a combination of topics could not be considered significant, perhaps due to their initially perceived irrelevance or because there are no known current problem owners for the presented research question. A question to which no one is interested in the answer should not be considered as significant, thus this property could be related to the earlier explanation as of why the chosen research topic could be relevant to problem owners.

Continuing with the fifth research question property, *the capacity to surprise,* is related to the predictability of the answer of the research question. If the researcher, problem owners and other readers could already correctly guess the answer, the effort spent into elaborate research would probably not be recognized as valuable. For instance, asking the question how it comes that apples fall down from trees, the literature review on explaining how gravity causes it to happen will not receive much appraisal. Researchers need to ask questions because it is their job to explore topics and phenomena. If the researcher would conduct a very predictive research, the lessons learned from the research will be very scarce. Asking a question to which a researcher does not directly know the answer of allows for a more learning experience, from both the researcher's and the reader's perspective, and could be applicable to this proposed research question.

Finally, the sixth property of Foss & Waters (2007) is related to the *complexity* of a question. If the research question could be answered with a simple *yes* or *no,* the research question lacks a robust structure in order to provide a challenge for the researcher to develop a fleshed out, detailed answer to a more complex subject. For instance, asking the question whether apples fall from trees means that no explanation should be given on how the law of gravity is responsible for this phenomenon, though it would be very advisable to include it in the answer. This is why the initial research question has been purposefully defined as a *how*-question to eliminate the *yes*/*no*-type of answer.

It would therefore seem that the initial research question passes the criteria for a well-defined research question as by Foss & Waters (2007). The research question is not predictable, does not lack complexity, can be associated to existing problem owners, gives an indication which topics need to be reviewed and what data could be collected to answer the question, and, most importantly, accurately describes the topics that are desired to be discussed by the researcher in this thesis. The only point of criticism could be that the initial research question lacks clarification on what fields of data are relevant to the scope of the research. One could claim that, given the way the research question is defined, customer order data could be considered as relevant to centralized network management systems. Thus, the research question should be altered to the following:

How can network data quality be improved within centralized network management systems to achieve more automation?

This way, the scope for the topic of data quality is limited as such that only network data is considered. Of course, network data would have to be defined, but the revised research question still passes the six properties of Foss & Waters (2007) to conclude that a well-defined research question has been formulated.

However, the research question should be accompanied with several sub-research questions in order to provide a complete answer. The research question suggests that there is a logical connection between network management systems and data quality, but it may not be as clear as it may seem. Sub-research questions help to complete the picture around the main research question, meaning that the answers to these questions help to support the importance of the research question itself. The answers to the sub-research questions also allow to fill any knowledge gaps that may exist upon initially presenting the research question. Furthermore, using sub-research questions is a suitable technique to better relate the combination of network management systems with data quality practices, and more specifically how the general benefits of data quality improvement can be applied in the context of the problem statement:

a) How can data quality assessment be paired with network management practices?

b) What are the benefits of improving (network) data quality?

c) How can the introduction of a network management system help achieving the goals of improving network data quality and automation?

The remainder of the thesis is organized as follows. Section 2 will discuss the relevant topics via an extensive literature review as deduced from the research question. In Section 3, the research methodology applied in this thesis is explained, followed by the discussion of a case study in the following section. Finally, in Section 5, there is room for discussion, conclusions, and further remarks regarding the described work.

2 THEORETICAL BACKGROUND

The following chapter will discuss the theoretical background of relevant topics to the problem statement of the thesis. The theoretical background can be seen as the first step in the research design, where critical analysis is offered on the topics of data quality, data governance, the practice of network monitoring and its latest developments regarding monitoring systems. After the theoretical background and literature review have been concluded, the next step is to define the research methodology to be applied when considering the case study research approach. This will be discussed in the following chapter.

2.1 Introduction of this Chapter

This section of the thesis will elaborate in further detail some of the core elements and theoretical concepts previously mentioned in the problem indication and problem statement in Chapter 1. In order to understand how the proposed research question may potentially contribute toward new insights within the academic research field of improving data quality in general, and even more specifically improving the data retrieved from high-end network monitoring systems, it is necessary to perform a thorough literature review to gain in-depth knowledge of the theoretical background of the topic in question. Any reader of the thesis should know not only the definitions used of relevant concepts such as *data quality* and *data governance,* but more importantly the scope of which these concepts are used. To give an example, which shall be elaborated further in the literature review, it would be perceived as infeasible to describe in elaborate detail all dimensions of data quality as given by multiple researchers. Furthermore, to improve the quality of the thesis, more attention will be given to carefully chosen dimensions of data quality due to their strong relevance toward the problem statement. This will eliminate for the proposed solution, based on the literature review and further research, to be elaborate and broad but most importantly, improbable. Providing an answer to the research question toward the reader should require knowledge beforehand of the discussed topics; thus, the topics of (a) data quality, (b) DDI and NetMRI systems solutions, and (c) network monitoring practices will be discussed. After discussing each topic, a brief summary should be presented to showcase to what level these concepts are interrelated to each other. This will showcase the value of the thesis as the combination of discussing data quality within the realm of network operations is still exciting due to, for instance, the rollout of IPv6 addresses allowing for complex and dynamic networks and the *Internet of Things* (IoT). Important is to keep in mind that network operations also includes security aspects; therefore, there should be an understanding how data quality can contribute to maintain or improve toward the desired level of security for company internal networks.

2.2 The Importance of Data Quality

Before discussing the dimensions of data quality and its definition chosen for the purpose of this thesis, it should be known beforehand the impact of poor data quality on business performance and its strategy. Liebchen & Shepperd (2008) discuss data quality in the setting of software engineering, and mention from the beginning how data quality should be interpreted differently depending on the context. The authors mention how within software engineering, data quality is perceived as the level to which dependent variables can be predicted through the data in question. The scope of their research was limited to understanding the aspect of *accuracy* of the data and how within academic articles of software engineering data accuracy is being considered as a concern for engineering or implementing applications. For clarification, the authors use the term '*noise*' as a substitute term for data accuracy.

From their first query consisting of 'data quality' and 'software' 552 articles were found for which only twenty-three articles addressed data quality. It is mentioned by the authors that there seems to be a slight increase of attention for data quality over the time period of 1998-2008. The value of their literature review of the twenty-three articles comes from their findings that not every article discussing data quality should be perceived as a threat for empirical data analysis. Wesslén (2000) discusses how data errors during the data collection process can be categorized as random, intentional and omitted. He perceives random data error not as a threat for data quality as the huge amount of data should be ignored. Statistically, there should be an assumption that as many data as possible falls under the 'true' or correct value, and thus the level of noise should average out on the long run. However, Liebchen & Sheppard (2008) counter Wesslén's statement by mentioning how, statistically speaking, while the level of noise may be indifferent when using large data pools, the level of variance of the data can be influenced by the noise.

This is where the discussion of 'unbiased' versus 'biased' data comes into play. Both Wesslén (2000) and Liebchen & Shepperd (2008) agree that intentional, biased data errors pose a larger threat for empirical data analysis, though they disagree on the unbiased random data error as previously discussed. Over half of the twenty-three selected articles of the literature review discussed manual quality checking as a possible method of tackling the data quality problem. However, an example is given from the case study of Johnson & Disney (1999) showcasing that software engineers analyzing data collected from their own work (89 projects in total) were responsible for almost half of the data errors due to their biased, but incorrect calculations. Thus, there seems to be a mutual agreement between all previously mentioned authors that data quality issues within data collection practices should be improved not by using manual quality checking practices, but to allow for tool support as an external measure. This is so that the practice of triangulation can be applied in data quality practices, meaning that multiple methods are used in order to validate results. This is described by Liebchen & Shepperd (2008) as a preventive technique for data quality loss. In this case, it would be to verify the true value of the data if known, or a possible true value in the case it is not predetermined. However, two articles from the literature review suggest that the usage of quality metadata (data about data) should be used instead to assess the quality in a more visualized way. One of these is from Mendes & Lokan (2008) in which they concluded in their research comparing single-company versus cross-company data models that data accuracy of single-company data models was superior when examining factors such as size and the effort required for the project along with project delivery rate. These factors acted as dependent variables for choosing whether a single- or cross-company data model should be applied. Thus, data about the project was used by running it through two separate data models, creating metadata, and applying this metadata as a practice for assessing data quality.

This is where the main challenge of data quality assessment lies. The true value of one particular piece of data is difficult to know. Reflecting this to the problem indication and problem statement described in the first chapter, presenting a joint action plan to improve data quality within a company for multiple groups of stakeholders (network operators, network engineers, and security engineers, to name a few) should start with an understanding of how each group perceives the threat of data quality. It could be that one or multiple stakeholders agree with Wesslén's (2000) opinion that in the case of large data pools, random (or unbiased) errors in the data should be ignored. Other stakeholders could argue that manual data quality checking should be incorporated and assigned to those with the most expertise. Nevertheless, the findings show that data quality concerns are growing and that there is no unified solution to this problem. While different methodologies of tackling data quality concerns are going to be discussed later on in this chapter, the next step is to understand how poor data quality can impact business strategy.

The paper of Redman (1998) can be used as a basis for understanding the impact of poor data quality on business strategy. The typical issues described are that it should be expected that an error rate between 1-5% is present for the collected data, that multiple databases increase the risk of data inconsistency, and that often due to redundancy of data decision-makers find themselves to be in a situation where necessary data becomes unavailable. Redman (1998) continues to categorize the impacts of poor data quality on the operational, tactical, and strategic business level. Expected consequences could be decreased customer satisfaction for potentially handling their data improperly, along with increased operational expenditures (operational impacts), increased risk to poor decision making and increased difficulty of reengineering (tactical impacts), and increased difficulty to either set or continue with a business strategy (strategic impact).

However, the value of the paper comes from the author's attention given to what he describes as 'softer impacts'. The data-driven culture of a business can also be seen by zooming in on the needs of the departments. Another tactical impact given caused by poor data quality is the increased organizational mistrust. Imagine if a department X is dependent on the data owned by department Y. When the data quality of department Y is perceived as poor, department X may choose to set up its own database, creating said mistrust. Such course of action taken by department X leads to the creation of business silos; corporate culture is lost and efficiency of information sharing has been mitigated. Furthermore, it creates corporate data governance challenges and is described by Redman (1998) as a strategic business impact leading to data ownership issues. Another soft impact is based on the perceived lower customer satisfaction rate: front-end employees who may not be held responsible for poor data quality practices have to deal with customer complaints and could lead to a lower morale during work, categorized as an operational impact.

Several explicit examples are given in the paper of De Veaux & Hand (2005) of how poor data quality can lead to unknowingly wrong decisions or conclusions. An important thing to note is that they describe how in the age of data mining practices, data owners are actually increasingly dealing with secondary data analysis. It is described as analyzing data initially collected for another purpose. An example given by the authors is how the data for money transactions is collected so that during secondary data analysis, transaction patterns may be detected. The authors also use a simplified categorization of how data errors can occur: (a) missing data values, and (b) distorted data values. An example of how missing data values can cause negative impact on the business can be found in banking. In the work of Hand (1998) the main research focus is to understand the consequences of rejecting applicants for a bank loan: was it a bad decision to reject this applicant? This question is asked because accepted applications are given a data value, typically in the *yes/no* variant, when the true outcome of the loan is known: has the applicant been able to repay the loan? Of course, while assuming that a rejected applicant would not have been able to repay the loan, the true outcome shall never be known. This can be seen as a form of missing data to which Hand (1998) suggests a predictive model when household income is considered a dependent variable for either accepting or rejecting the applicant for a loan. Applying a model replaces a missing value and the risk of the applicant not being able to pay the loan back has been converted from being assumed to being measurable. Further work on this scenario has been performed by Sohn & Shin (2006) where a confidence interval of *survival analysis* is applied. To put it simply, their proposed solution to determine the true class (yes/no regarding repayment) of a rejected applicant is by determining the time to which the first delayed repayment occurs.

A distorted data example could be the conversion of metric systems of measurement during international data exchange. Launched on December 11, 1998, the Mars Climate Orbiter was a space probe intended to collect weather data and to land on the south pole of the planet Mars. The project was estimated to cost around $125 million (Pollack, 1999) and was considered a failure after it went sixty miles off-course and was lost in space. The reason for what could have been presumed a miscalculation was the discrepancy of using the English metric unit of *pound-seconds* while the scientists at NASA presumed it to be written as *newton-seconds*. Further research on this particular case has been performed by Sauser et al. (2009) discussing a project management approach to potentially allow for upfront analysis of the characteristics of the project, and thus to the elimination of inconsistent methods of entering data and striving toward a unified, agreed upon structure to enter data into programmable applications.

De Veaux & Hand (2005) also mention the need for triangulation during data collection practices and refer it as a *duplicate performance method* where comparisons of data between double entry systems are being made. Another suggested method mentioned in their article is similar to the approach of Mendes & Lokan (2008) where metadata information should be used to not only check for data consistency, but whether variables are related in order to create a fitting predictive model for data. However, this may become an increasing challenge with larger data pools.

The question now is how this acquired knowledge can be applied to the problem statement of the thesis. What organizational impacts can be caused by poor data quality in network systems by either missing or distorted data? Important to keep in mind is that network systems play a critical role in the field of security. Missing or incomplete data about devices connected to the internal network of the company without fully understanding its device name or verification of its IP address can cause for operational impacts in the form of increased operational cost and effort to fetch additional data (or place it manually) and rename the device for naming convention. More importantly is the impact in the field of security. If an internal system contains a lot of ‘clutter’ in its data, a business impact on the tactical level would be for the security engineer team members to make poor decisions based on the potentially false results of the security and risk assessment of the internal network. A strategic impact would therefore be to discontinue the pursuit of a business strategy due to insufficient confidence that safety is ensured by relying on the data collected from network systems. Similar impacts can also be said for distorted data values: an IPv4 address not showcased in a 32-bit format will raise some security questions, and the presented data value is of no use to verify the IP address. It is important to keep in mind that it is very unlikely to think that a system can achieve data collection without any data errors included. There should be an understanding of the fault rate acceptable even if triangulation cannot be achieved to verify data values. As seen in the literature review thus far, it is the type of unbiased data errors to which a compromise has to be agreed upon. Intentional or biased data errors made by humans may lead to further discussion and the delay or the cause of poor decision-making. Examples have shown us how minor data errors can have grave consequences in virtually any project or process depending on data for its end result. It has also been discussed how one may miss out on new opportunities by not analyzing the data further or, in the case of the rejected applications for a loan, dismiss on developing a prediction model and perform a “what-if” analysis to predict data values which would otherwise have stayed being missing values.

The following section focuses on the perception of the decision-maker: how do they interpret the data they are being presented with? Are there any factors that could improve the awareness and interest of decision-makers in regard to data quality? And, more importantly, is it possible to apply this knowledge for the joint action plan to the particular stakeholders in question? Chengalur-Smith et al. (1999) discuss the possibility of data quality tagging for decision-makers. Their argument for conducting research on achieving a better understanding of the relationship between decision-making strategies and data quality is that due to data warehousing data is used by multiple parties and often for multiple, unintended purposes. Tagging in this definition is described as allowing a data value to be added to the original data describing its quality. The quality is pulled from metadata describing the quality of the original data. The researchers have chosen for two variants of visualizing this value of quality to potential decision-makers: (a) an interval scale between 0 and 1 describing the quality evaluation of the related data in relative terms, meaning that a score of 0.6 does not mean that 60% of the data in question is considered to be correct but is a mere indication that it contains higher data quality than a data set scoring 0.5; and (b) an n-point ordinal scale using words such as ‘excellent’, ‘good’, ‘average’, ‘poor’, etc. Two types of decision making paradigms were also discussed: *conjunctive* and *weighted additive* decision making. In the conjunctive format, decisions are made by setting a minimum acceptable level on multiple criteria and judging the different options on whether they meet all of the criteria. Only those options whose criteria have been evaluated higher than the set minimum are considered. Weighted additive decision making is, what the name suggests, a method in which each criteria is given a certain weight of importance. The end result is the evaluation of the criteria multiplied by its weight, and the option with the highest sum is suggested to be the overall best decision. An example would be the Analytic Hierarchy Process (AHP) of Saaty (2004) as it considers placing weights on *multi-criteria* decision-making practices.

Their conclusion was that including too much detail about data quality could actually work counter-productive as the risk of information overload to the decision maker is present for complex decision scenarios. It would be that the interval format describing data quality would be less preferred than the ordinal format in such particular case. However, the inclusion of data quality measurement in the interval format for each criteria considered in the decision-making process did influence the final decision. Conjunctive decision making also seemed to work more effectively compared to the weighted additive variant, possibly due to information overload and the attitude of the participants to focus only on the weights with the lowest and largest impact. However, combining conjunctive decision-making and the inclusion of data quality tags did create a challenge of understanding the impact of data quality if the weights are assigned equally throughout all criteria. Adding more complexity by adding exact figures of data quality assessment in an already complex decision case would therefore not guarantee an improvement of more conscious decision making. The results of Chengalur-Smith et al.'s (1999) experiment also showcased that the simple decision-making scenario paired with data quality assessment in the interval format allowed for the smallest level of to which the information of data quality was ignored. Thus, while not explicitly mentioned by the authors, it would seem that exact details of data quality information in simple decision-making scenario could have a stronger positive impact than providing descriptive, yet vague tags of data quality.

Taking their insights one step further, Fisher (1999) discusses whether time constraints and experience levels of the decision-makers themselves also have an influence to how they perceive data quality information using both interval and ordinal formats. The researcher asked the question not only if people would use data quality information, but if experts would use such information more than novices. Important to understand is that *experience* is divided into (a) general work experience measured by the number of years a person has worked, and (b) domain-specific experience. Based on the results of his experiments, experts do place a higher value on data quality information than novices, but the amount of 'general' experience does not seem to be a predictor regarding the use of data quality information. It is that those possessing domain-specific experience are more alert to data quality information, but only when it is included. When data quality information is not included, there does not seem to be a difference between those possessing domain-specific experience and those lacking it. With novices, results were found similar to that of Chengalur-Smith et al. (1999) where novices would be assured of their own decision and were neglecting data quality information tags.

As for the time constraint, it would seem that using the ordinal format for data quality information allows for quicker decision-making for novices in the case of short time constraints. However, the interval format did not seem to have much influence because of its complexity. The length of time constraints did not seem to have a significant effect on the decision-making process regarding experts. However, the inclusion of data quality information did seem to have an effect on the experts as they felt some sense of time pressure compared to the experts who were not provided with data quality information who did not feel time pressure irrespectively of the length of the time constraint. Though, the most important observation is that regardless of whether data quality information is available, the level of time constraint did not seem to influence decision-making.

Combining the efforts of understanding both the importance of data quality from the perspective of impact on business performance and the perspective of the decision-makers, it has now been made clear not only why data quality is to be considered important but also in what presentable format it can be perceived as important to those responsible for making decisions based on the data. Applying this to the problem statement and research question presented in the thesis, the challenge of making different groups of stakeholders aware of the importance of data quality might be tackled in the joint action plan. For instance, if it would ever come to the conclusion of including data quality tags from here on out, this section of the literature review has discussed the possible risk of information overload. Another advantage of having performed a literature review on this topic is to gain a better understanding of the possible impact of poor data quality on network systems and to clearly present it to stakeholders. However, the definition of data quality used in this thesis has not yet been discussed, and the interrelationship between data quality and network systems, in particular DDI-systems, still has to be discussed in further detail in order to possibly set up a structure for answering the main research question. Thus, the following sections shall discuss the dimensions of data quality and established methodologies of improving data quality.

2.2.1 Basic overview by Fan & Geerts (2012)

The first chapter of Fan & Geerts' book (2012) discussing foundations of data quality management gives an overview of how the quality of the data should be assessed. Their speculations made for the increasing importance of data quality management next to overall data management has been due to the trend of big data management and the increased efforts of using enterprise resource planning (ERP) software packages. The authors break up the issues of data quality into five categories: (a) consistency, (b) deduplication, (c) information completeness, (d) currency, and (e) accuracy. Each of these issues will be addressed briefly, with more attention given to the data accuracy issue.

Data consistency deals with the validity (well-founded conclusion) and integrity (truthfulness) of those data values that are representing materials or persons in the real world. Usually, data consistency is linked with relational databases and data dependencies due to the relations data values may have. An example is how the country calling code for the Netherlands is '+31'; data stating a Belgian home phone number starting with '+31' would therefore not be consistent with the established rule. *Data deduplication* practices are put into place to check whether multiple entries of data are referring to the same real-world entity (person, building, product, etc.) in question. An example is given where in payment fraud, it should be checked whether the card holder is also the user and therefore responsible for any suspicious purchases. With poor data quality, attempting to match data tuples while values may be inconsistent or missing makes it time-intensive and costly to achieve deduplication.

Data information completeness asks the question whether a query can be answered in a given dataset or whether external data is needed. Decisions that need to be made based on the results of such queries can be prone to being inaccurate and bias in the case of information incompleteness. Both Fan & Geerts (2012) and Abiteboul et al. (2006) describe two types of assumptions on how one should deal with incomplete information. The first one is the *Closed World Assumption* (CWA) where it should be assumed that all attributes relevant to the database concerning an entity have been collected, and that no other conclusions can be made except for those supported by the database. The only cause for information incompleteness would be missing data values of attributes within the database. The second assumption is named *Open World Assumption* (OWA) where the database in question may also have information incompleteness due to the possibility of it not having collected all tuples of the real-world entity. Fan & Geerts (2012) mention that although the CWA-type of database management is applied, it is inappropriate to assume a database is either fully closed or fully open. *Partially closed* databases are preferred by the authors to achieve complete information.

The fourth central issue, *data currency*, is concerned with establishing some form of identification of whether the values stored in a database are considered to be currently accurate. The inclusion of a timestamp would solve this issue, but research by Dong et al. (2009) shows that data integration without a uniform method of information extraction, such as how to include timestamps, leaves room for uncertainty of the integrated data. Thus, data quality issues due to currency are related to the difficulty of identifying the most recent data values of entities.

Finally, *data accuracy* relates to the closeness of a value to what is presumed to be the true value of the entity. However, as it may be often the case, the true value is not known and thus *relative accuracy* is used instead. For this, studying the meaning of the data should be performed in order to deduce the accuracy. A registered pupil at elementary school with an associated age of ‘64’ is bound to be inaccurate, especially if it is listed with other tuples containing pupils of which the associate age is within the same appropriate age range. More attention will be given to data accuracy problems later in the thesis.

Although it has now been possible to divide data quality issues into five categories as done by Fan & Geerts (2012), it is far more difficult to place a problem into a single issue category due to their interactions. To give an example, data deduplication can only be performed whenever a true value is known, but the true value has to be deduced from relative data accuracy, which in its turn can only be done when there are few to none information completeness problems. It may therefore be that focusing on one aspect of data quality could indirectly lead to improvements of other data quality aspects. For this thesis, it has been chosen to focus more on data accuracy for the following reason: making assumptions within the field of security can be a risky approach to assess the level of threat within an internal network. Focusing on how data accuracy can be increased without being heavily dependent on relative accuracy should give employees a stronger sense of confidence in the data and, consequently, a stronger sense of security.

2.2.2 Comparison with other Authors

In order to define a proper definition for data quality throughout the thesis, it is suggested to reflect on what other academics have written within the research field of data management. Wang & Strong (1996) define it as suitable data for a potential customer of said data to be used, while Fisher & Kingma (2001) suggest rather than using a single definition, a set of variables should be selected which combined creates a broad concept of data quality. Similar to Fan & Geerts (2012), accuracy, consistency, timeliness, and completeness are used as relevant dimensions; however, relevancy and fitness for use are mentioned by Fisher & Kingma (2001) as most frequently used dimensions but are not mentioned in section 2.2.1 discussing the first chapter of Fan & Geerts' (2012) book. To clarify, the relevance of data focuses on how an end-user can use the data to solve a relevant business issue; data used for other purposes would lover data relevancy. Fitness for use is relevant to the discussed work of Chengalur-Smith et al. (1999) and Fisher (1999) since it is related to the format of how the information derived from the data is presented to the responsible decision-maker.

Pipino et al. (2002) also acknowledge that data quality should be perceived as a concept consisting of multiple dimensions, and derived from their literature review sixteen dimensions have been identified as relevant for the data quality issue. Relevant dimensions that have not been mentioned earlier are *security* and *ease of manipulation,* though Redman (1998) does mention security and privacy within a separate category of data quality issues. However, Pipino et al. (2002) go one step further by categorizing the valuation, or judgment, of data quality into a two-by-two matrix using the level (low or high) of objective and subjective assessment. The work of the authors is mainly focused on developing data quality metrics and an approach of assessing data quality in order to create an effective joint action plan for data quality improvement; therefore, it will be used in the thesis later on.

Liebchen & Shepperd (2008) prefer to focus only on the earlier mentioned dimension of *fitness for use* of Fisher & Kingma (2001) and describe it as to what extent the data can be used for a solemn purpose such as predicting variables. The authors also describe how the dimensions of accuracy, completeness and timeliness are subject to change in terms of their importance dependent on the problem in question. While it might be debatable how the fitness for use of data may change depending on the type of end-user, Liebchen & Sheppard (2008) focus on this dimension within the field of software engineering only and in terms of predicting depending variables for a software project. Finally, Wand & Wang (1996) give critique on how several popular dimensions such as *correctness* and *accuracy* are not consistent between authors and comply with Liebchen & Shepperd (2008) on focusing only on the use of data. However, Wand & Wang (1996) discuss how, within their context of systems design, there has been a lack of describing data quality from a ‘data-centric’ point-of-view: purely looking at values and structure of data rather than focusing on misinterpretation of information due to noise within communication or the value of information based on established criteria. Though, the authors see shortcomings of having such perspective on examining data quality as the specifications of data, including its structure, should be defined beforehand in order to assess its quality.

Based on this briefly described review of gaining a better understanding of how data quality is described within its research field, the definition assigned to data quality in this thesis will be in the form of multi-dimensional aspects with a stronger focus on data accuracy and fitness for use with regard to decision-makers as the designated end-user of the data. In particular, the five dimensions described by Fan & Geerts (2012) shall be used as the basis for defining data quality throughout the thesis with possible additions of the security dimension as given by Pipino et al. (2002) and the *fitness for use* dimension as mentioned by Wand & Wang (1996), Fisher & Kingma (2001), and Liebchen & Sheppard (2008) with regard to data-driven decision-making policies as discussed earlier (Chengalur-Smith et al., 1999; Fisher, 1999).

The data quality model of Wand & Wang (1996) is targeted towards the system design audience in which data quality should be seen from two viewpoints: internal and external. The viewpoints are described with the aid of using two types of systems to explain the data representation process: an information system should be seen as a black box which is able to represent the given state of entities of the ‘real-world system’. However, in order for this to occur, the design of systems operations to create usable data for decision-makers, such as data collection, entry, and delivery, should be done in such a manner that a perfect implementation is achieved, meaning, all data from the real-world system do not contain any deficiencies. An ontological approach has been used in order to define data quality: the real-world system should be represented by an information system while respecting the states and laws of the real-world entities. The word ‘state’ is referring back to the dimension of *currency* where data can only be captured in one point of time and should be captured correctly for that particular moment. Wand & Wang’s (1996) work on defining the intrinsic dimensions of data quality is not only helpful to formulate a generally accepted definition for the thesis, but also relates problems of these dimensions to flaws in the design or operation process of systems, and more particularly how these data flaws between real-world systems and information systems can be described. Completeness, unambiguousness, meaningfulness and correctness are the chosen intrinsic dimensions of the authors. To compare with Fan & Geerts (2012), all of the five dimensions given in the first chapter are related to the internal view of data quality focusing on design and operation according to Wand & Wang (1996), but the dimension of *timeliness* is considered to be relevant to both the internal and external dimension, the latter being more focused on the usability and value of data for end-users. Other data-related dimensions to the external view could be relevance, scope and freedom from bias.

For the end of this sub-section, the paper of Wang & Strong (1996) tries to understand data quality from the perspective of data consumers. They have established four aspects of four criteria based on the literature review and their understanding of *fitness for use*: data is *accessible*, *interpretable*, *relevant*, and *accurate*. Based on their first survey, a hundred-and-eighteen attributes of data quality were identified of those using data for decision-making for their business. Their method was to ask each respondent their first thoughts of what aspects are relevant to data quality, followed by allowing the respondent to select a list of thirty-two other aspects found during Wang & Strong's (1996) literature review. Their second survey is intended to measure the importance for each of the mentioned aspects, and the following section will delve deeper on the importance of accuracy.

However, Wang & Strong (1996) do make a formidable effort to organize the data quality aspects into twenty dimensions, and four main categories have been derived from selectively grouping these dimensions together: *intrinsic*, *contextual*, *representational*, and *accessibility*. The authors claim that the four dimensions are consistent with their defined criteria for data quality, as intrinsic data quality represents the accuracy of the data in its own right. As for contextual data quality, it can be interpreted as the relevance of the data as *relevancy* is recognized as one of the related categories, together with timeliness and the value-adding purpose of the data for decision-making. Representation and accessibility of data quality are, in short, consistent with the established criteria of, respectively, interpretability through systems collecting or using data in a non-foreign language and the criteria that data should be accessible, meaning users know how to retrieve it.

Examining the definitions given of data quality for the thesis, it would seem that the dimensions given by Fan & Geerts (2012) are distributed over three data quality dimensions of Wang & Strong (1996), with deduplication, currency and accuracy falling into the dimension of intrinsic data quality, consistency towards representational data quality, and information completeness towards contextual data quality. Furthermore, considering also the criteria of fitness for use and security, respectively they could be placed best under the dimensions of contextual data quality and accessibility of data quality. Interesting from this brief analysis is that the definition and criteria chosen to be considered to assess data quality are distributed in all four data quality dimensions defined by Wang & Strong (1996), allowing for a greater level of confidence that multiple perspectives on the term 'data quality' are taken into consideration in the thesis.

2.2.3 A stronger Focus on the Accuracy Dimension

While Herrera & Kapur (2007) also give their own categories of dimensions related to data quality problems, such as validity so that theories can be tested based on the data and coverage so that the data may or may not be present to answer the research question, the main focus of this thesis is to understand how they perceive accuracy of data quality. It is described by the authors as the level of error avoidance during the processes of data collection and presenting the data. The two types of errors discussed are referring back to Wesslén's (2000) typology of biased versus unbiased types of data; either end-users are examining the data from a subjective viewpoint, or data errors have been fetched due to a wrongly implemented methodology. Herrera & Kapur (2007) explicitly mention that the biased data quality errors should not be blamed on experts on the workspace, as has been discussed earlier having explained the work of Chengalur-Smith et al. (1999) and Fisher (1999). However, the challenge described is how biased decisions can be understood better in the hopes of improving the presentation of quantitative datasets for better usage in the future. During their research of comparing data quality methodologies, Batini et al. (2009) describe two types of accuracy: syntactic and semantic. Syntactic accuracy would be described as the closeness of a data value given the domain of the data; this does not necessarily mean that the data value is correct, but whether is it accepted as a value within a certain data domain. The authors mention that syntactic accuracy is considered only in their data quality methodologies chosen for comparison. Semantic accuracy is best described as the level of correctly representing the real-world state of an entity (Zaveri et al., 2015). For completeness, data semantics by Madnick & Zhu (2006) are explained in the appendix (see Appendix B).

Referring back to Wang & Strong (1996), they have placed the category 'accuracy' into the dimension of intrinsic data quality. To explain further, this dimension discusses how, given the example of product quality, the quality demands of the consumer go further than that of the product manufacturer. Whereas the manufacturer would like for the data to be accurate and interpreted as objectively as possible by the decision-makers for green-lighting the manufacturing of the product, consumers care about the reputation of the manufacturer and thus, indirectly, they care about how the manufacturer assesses the level of believability to the data, and for IT professionals the reputation of the source from which the data was fetched from. Accuracy of the data should therefore, according to Wang & Strong (1996), not be understood as the believability of data or as its level of objectivity, but should be considered a separate category of data quality criteria that focuses on the quality that data can carry on its own before any biased or unbiased examination has been performed. Wand & Wang (1996) explain the difficulty of attempting to define 'accuracy' as a data quality dimension as it is derived from their literature review that, while accuracy has been mentioned most frequently as an important dimension for data quality, descriptions of the dimension may differ and may be overlapping with other established dimensions, notably 'correctness'. The authors suggest that defining 'accuracy' may be meaningless. Instead, since they focus on separating information systems state with real-world state, as explained earlier, *inaccuracy* can be defined as the result of the information system incorrectly defining the state of the real-world entity, on the condition that a valid state was entered rather than a meaningless data entry. Consistent with Fan & Geerts (2012), inaccuracy is presumed to be strongly correlated to incompleteness and uncertainty of the data, and should be considered an internal dimension of data related to design and operation as Wand & Wang (1996) perceive the derived value and use of data as an external dimension (or view) of the data.

2.2.4 Methodologies of Improving Data Quality

In order to understand different possible methods of how to improve data quality within the situation given at the problem statement, it is best to understand how other researchers approach established data quality methodologies and possible methods for data quality improvement. As has been discussed earlier, there is a debate amongst researchers about which dimensions should be included within the topic of data quality. The paper of Batini et al. (2009) gives a very detailed description of the comparing method used for thirteen carefully selected data quality methodologies. Though it is not the purpose of the thesis to explain which methodology would be most complete and fit for the problem statement at hand, the paper gives a clear insight as to how data quality methodologies differ and which data quality improvement techniques are more favored than others. The paper starts off with a brief description of how data quality methodologies are typically composed: in sequence, the three steps included are (a) *state reconstruction* focusing on, for instance, data collection procedures to understand quality issues; (b) *assessment*, where the source of poor data quality is determined, and (c) *improvement*, describing the techniques to be applied to improve data quality to a new certain minimum acceptable level. However, the authors give a further breakdown of the assessment and improvement phases of the typical data quality methodology. Important for the thesis is to understand the different strategies applied in the improvement phase. Two types of strategies are identified by the authors: *data-driven* and *process-driven*.

Data-driven strategies are focused on data values and methods for modifying them to be compliant with the data quality dimensions of accuracy, completeness, timeliness and consistency. Besides acquiring new, perceived higher-quality data to replace biased data sets and the standardization of values to create consistency, other data-driven strategies suggested are to assess the level of trust of data sources, cost optimization by minimizing the cost of effort to achieve data quality, and the localization of errors by establishing data quality rules. Such rules will be explained later in the thesis, notably conditional data dependencies. However, a more high-level technique of data-driven strategy is called *data and schema integration* by Batini et al. (2009), and is focused on data access through multiple disparate sources on one unified view or platform. Again, a breakdown of three different types of such disparities is given: (a) technological, (b) schema, and (c) instance-level. Technological differences can be explained by having multiple products of different vendors within the IT infrastructure. In the case of our problem statement, DDI-technology can be bought as a package from one vendor with data collection and storage provided by another piece of technology. Data integration should be applied in such scenarios for it to remain compatible in such complex infrastructure. Using different data models or having different data values that actually represent the same ‘real-world’ entity are examples of schema heterogeneities which hinder data consistency. Finally, instance-level disparity is described as the event when multiple sources provide conflicting data values for the same ‘real-world’ entity. Causes of such events could be separate established processes that allow for such data sources to be used for data collection. Again, data integration as a data-driven strategy for data quality improvement should prevent such situations.

Two process-driven strategies for data quality improvement are given by Batini et al. (2009) in the field of (a) control, and (b) redesign of processes (often called business process reengineering or BPR) to prevent poor data quality control. The authors explain that the choice for either data-driven or process-driven strategies for improvement is partially determined on the time perspective: data-driven techniques seem to be in favor when focusing on one-time solutions or cost optimization on the short term. However, in the long term, process-driven strategies allow for further detail to understand the cause of data quality problems, but could be expensive on short-term timeframes in the case of process redesign.

Interesting in the approach of Batini et al. (2009) is their monitoring of the types of information systems considered in the different data quality methodologies. While the majority of the methodologies considered focus on monolithic systems, other information systems are considered too: data warehouses, distributed and cooperative information systems, web information systems and P2P (peer-to-peer) information systems. Reasoning for the attention given on monolithic information systems is that it simplifies the scenario considering data storage and queries: data is not shared between multiple applications, leading to an increased risk of data duplication. As explained by Fan & Geerts (2012), data dimensions may be interrelated and thus, redundancy of data could affect or cause multiple issues related to other data quality dimensions. Important for the thesis is to understand to which category of information systems DDI-technology can be placed in to achieve a better understanding of fitting data quality improvement methodologies. A first examination would leave to suggest that DDI-technology could be placed under the category of *distributed information systems*, meaning that a division of tiers is applied within applications to divide presentation and data management to improve workflow and interoperability. While some methodologies are focused on distributed information systems, others implicitly consider them as their main focus lies on monolithic systems

(Batini et al., 2009). An explanation given for the evolution of focus by data quality methodologies toward new types of information systems is the growing number of data sources and the introduction of new data quality dimensions such as completeness, currency, and consistency. Trust of data sources have been questioned since the introduction of the Internet and the derived Web-based and P2P information systems. Growing need for cooperation between multiple departments would mean that accessibility of data should now also be considered an important data quality dimension.

A summary of the comparison of the different data quality methodologies by the authors is done in a classification of *complete*, *audit*, *operational*, and *economic* methodologies. As the name would suggest, economic methodologies focus on cost evaluation, while operational methodologies only consider technical issues during the assessment and improvement phase of the methodology structure. Audit methodologies are described as those methodologies spending much focus on assessing the data quality issue but not providing sufficient support for improvement practices. Complete methodologies are the preferred type as not only the assessment and improvement phases of the methodology are fleshed out, but also pay attention to the combination of economic and technical challenges. Only two have been considered to fall into this category: *Total Information Quality Management*, or TIQM (English, 1999), and *Comprehensive methodology for Data Quality management*, or CDQ (Batini & Scannapieco, 2006). Complete methodologies are considered to be most fit for developing frameworks in data quality assessment and improvement programs for those organizations with critical data being one of their top priorities, such as the company in the particular case explained in the problem statement of the thesis. However, Batini et al. (2009) do conclude by stating that even complete methodologies may be considered biased. This is due to their concern that both TIQM and CDQ are primarily focused on structured data while semi-structured or even unstructured data may be present at a company as well. Batini & Scannapieco (2006) suggest in their methodology that knowledge management techniques should be applied in such scenario. As for the interrelationship between data quality and process quality, related to data-driven and process-driven approaches for improvement, the authors acknowledge that improvement on data quality is not enough for better decision-making, and that attention should be given of understanding data from the perspective of the decision-maker, as has been discussed in section 2.2 with the work of Chengalur-Smith et al. (1999) and Fisher (1999). Finally, there is still room for further discussions whether or not methodologies should be expanded even further to consider data quality dimensions such as security, as addressed by Redman (1998) and Pipino et al. (2002), and dependencies between data dimensions should be examined further.

While the typical structure of data quality methodologies have been discussed along with an explanation of both data-driven and process-driven approaches to improve data quality, further research on understanding dependencies of data dimensions could be done by examining the work of Fan (2008) and Fan & Geerts (2009) discussing (conditional) dependencies of data values. Fan (2008) revisits established data dependency types in the hopes of finding new methods for data quality improvement by altering or expanding data dependency rules. Not going into full detail of the semantics of the data dependency, both conditional and matching dependencies are briefly discussed to convey the message better. Conditional dependencies are used to detect inconsistent data by establishing some rules (hence the word *conditional*) to see to what extent tuples of a data record follow what is assumed to be a correct relation. However, conditional dependencies can only be established once a proper functional dependency has been found. For instance, the country code, area code and a phone number should specify a geographical location in the form of street, city, and zip code. A more formal notation could be the following: [country code, area code, phone number] → [street, city, zip], e.g. [country code = 31, area code = 013, phone number = 1234] → [street = Drive Av, city = Tilburg, zip = 9999AA]

So, for a given combination of data values of country code, area code and phone number, there should be only one correct combination of values of the street name, city and zip code, creating a conditional functional dependency (CFD). However, the danger of creating a set of CFDs could be that conflicts due to inconsistency on the conditions may arise (Fan, 2008). Advanced examples of CFDs are given in the appendix (see Appendix B) where inclusion dependencies and matching rules are discussed. Testing dependencies against data when the dependency itself is faulty is a waste of effort. However, two methods of improving data quality with dependencies given in the paper are the following: (a) data repairing, covering the process of cleaning and imputing new data to improve on the original database; and (b) consistent query answering, where it is measured to what extent the output of the query changes due to the data repairs, having in mind that the goal of data repairing is not to change answers on queries per se but to minimalize any difference of query answering compared to the original database. Though, despite these suggested methods, conditional data dependencies have to be tested on their performance possibly through algorithms for testing. Furthermore, the process of data repairing is strongly related to that of object identification (or data deduplication), and a combination of both should be developed, forcing the combination of conditional and matching dependencies.

Further improvement methodologies not considered for the research purposes of the thesis but discussed can be found in the appendix (see Appendix B). It discusses the 'corporate householding problems' by Madnick & Zhu (2006) describing how the receiver's expectations of the data may differ from what was initially thought of by the provider (human or machine), ultimately suggesting more development into contextual and semantic data. Furthermore, a follow-up study by Fan et al. (2009) regarding CFDs and relational candidate keys (RCKs) is also discussed (see Appendix B).

2.3 Data Quality and Key Performance Indicators

To think back at the research question, in order to catch any improvement of a business process organizations may choose to use certain *key performance indicators* (KPIs) to measure business process performance. The purpose of using KPIs is to have a quantitative measurable approach to performance or other intangible concepts, such as employee satisfaction. Examples of such measurement approaches are the balanced scorecard of Kaplan & Norton (1996) and the performance measurement matrix of Daniel & Keegan (1989). KPIs can have multiple measurement units, such as a percentage, an average score based on an interval from 1 to 10, or a financial ratio. Besides the difficulty of selecting the right KPIs for each project, Masayna et al. (2007) discuss the difficulty of establishing a possible link between KPIs and data quality. Their argument is that since organizations are now dependent on the produced results from KPIs to make project decisions, the data required to produce the KPI results should be of proper data quality. The question remains however which data quality dimensions should be considered for each KPI established, and Hey (2006) stresses that organizations struggle to identify suitable data quality dimensions holding either high risk or value. This is why Masayna et al. (2007) have presented a framework in order to achieve both the accuracy of KPI results and increase the awareness of stakeholders regarding the importance of data quality.

The authors then continue to give a breakdown of the types of KPIs applied by organizations. They could be either (a) process-based, (b) activity-based, or (c) outcome-based indicators. Process-based KPIs are related to the compliance of business processes given certain guidelines that the organization has to follow. Examples of activity-based indicators can be the headcount of employees within a given project or the total dollar amount spent thus far on the project timeline. Finally, outcome-based indicators can be measurements of the rate of success of a project or process, or a quantitative measurement of the effects of the project (e.g. goals achieved). We shall see that most of the KPIs established by the organization in the case study are outcome-based indicators.

Masayna et al. (2007) stated that very few previous studies have touched upon the topic of data quality and KPIs, and reasoning for exploring the topic is derived from their shared belief that multiple enterprise systems, such as an ERP and CRM systems, and additional data files are distributed over the organization, creating a larger risk in data incompleteness, data accuracy, and data timeliness. MacMillan (2007) adds that consolidating data into one centralized system is also seen as a challenge since organizations would have to put their trust into a single system. Furthermore, their belief is that organizations neglect the activities regarding data collection and data quality when developing their KPIs, meaning that the form of the data and its quality are not considered. Based from their literature research, organizations needs to define which data sets are to be used for their KPIs, along with documenting the data source, the owners and/or stakeholders of the data, which data should be prioritized, and which data quality dimensions should be taken into consideration for the selection of data to measure the KPI. The framework of Masayna et al. (2007) is now presented in the following figure.

It should be clarified that the goal of the researchers is to improve data quality for the sake of making KPIs more accurate, and not to define KPIs simply based out of data quality dimensions. This is because, as it has been discussed in the literature review, multiple researchers apply different definitions of the same named data quality dimension or use different sets of data quality dimensions. While the activities for developing KPIs and data quality policies are self-explanatory in the network, both organizational and external influences need to be explained. The most prominent external influences organization are faced with are enterprise governance regulation standards, such as the Sarbanes-Oxley Act of 2002. Of course, both KPIs and data quality policies are dependent on the actual strategic goals of the organization, as the KPIs should quantitatively express the performance of the organization working toward their strategic goals, and in its turn these KPIs are computed from the data which should be reviewed in its relevant dimensions.

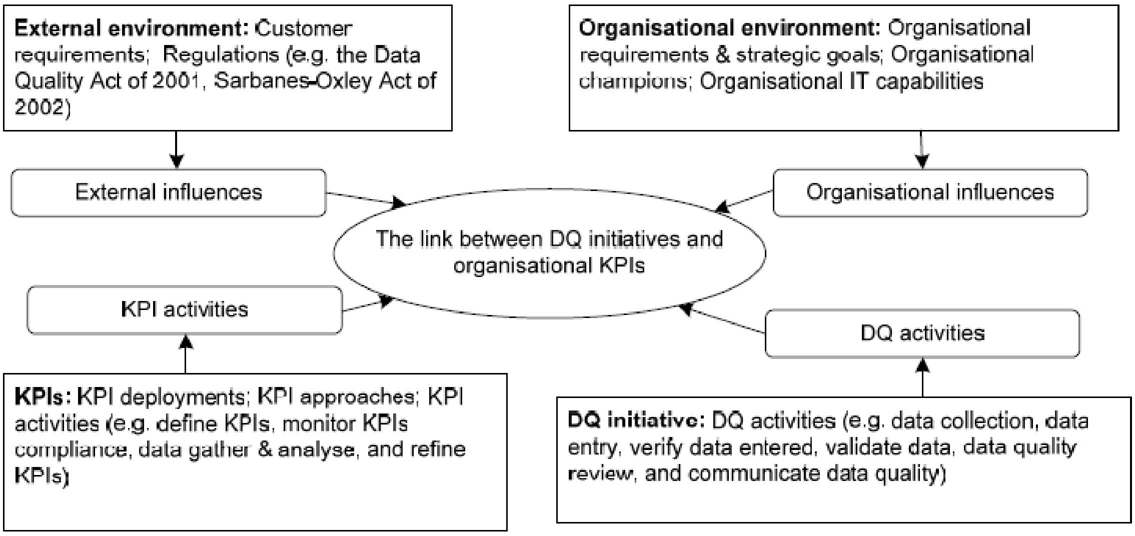

External environment: Customer requirements; Regulations (e.g. the Data Quality Act of 2001, Sarbanes-Oxley Act of 2002)
Organisational environment: Organisational requirements & strategic goals; Organisational champions; Organisational IT capabilities
External influences
Organisational influences
The link between DQ initiatives and organisational KPIs
KPI activities
DQ activities
KPIs: KPI deployments; KPI approaches; KPI activities (e.g. define KPIs, monitor KPIs compliance, data gather & analyse, and refine KPIs)
DQ initiative: DQ activities (e.g. data collection, data entry, verify data entered, validate data, data quality review, and communicate data quality)

Figure 1 Theoretical Framework of Linking Data Quality and KPIs (Masayna et al., 2007)

Sufi Abdi (2013) discusses Masayna et al.’s (2007) framework and adds the practice of *performance measurement* in order to strengthen the link between data quality and KPIs. Using Lichiello & Turnock’s (1999) work, it is described as the practice of regular data collection and data reporting in order to present achieved results which should be incorporated into KPIs. It requires not only input from stakeholders to be aware of possible data quality problems, but also well-defined goals (both short-term and long-term) and technical assistance in case of failures. Because of the inclusion of performance measurement, Sufi Abdi (2013) states that blind data collection and developing KPIs is not enough for the established link to be strong. Indeed, Tarokh & Nazemi (2006) state that the performance data should be relieved as well besides the data quality dimensions of the raw data in order for organizations to make better decisions.

Finally, for a clear view of the interrelations between data quality, its dimensions, and the established KPIs and business goals, we refer to the doctoral dissertation of Masayna (2006) before the theoretical framework was developed. A breakdown was used between KPIs and so-called KGIs, which stood for key goal indicators. The reasoning for separating performance metrics with goal metrics was based from the CobiT management guidelines (2004) and is to eliminate the blur between performance from business processes and goals derived from the organization's strategy, as strategy has to be supported via its business processes. It can therefore be derived that data quality dimensions are not only influencing the presentation of the KPIs, but thus also the belief of the organization having achieved its strategic goals by evaluating the results of their defined KPIs.

Coincidentally fitting with the research topic, Rodriguez et al. (2009) discuss data quality problems related to KPIs in network information centers including DNS services and logging of associated events. Two business processes that have been described in their work is the administering of an domain within DNS and the modification of information related to these domains. Examples of KPIs that were introduced are the latency between request and reply for a payment and the time between adding a domain name and its first successful support request by a user. The latter is a proxy for determining the quality of support services and documentation available for the user. The authors stress that these business processes have been chosen since it allows for the creation of events and thus data, which can then be reviewed periodically for its quality.

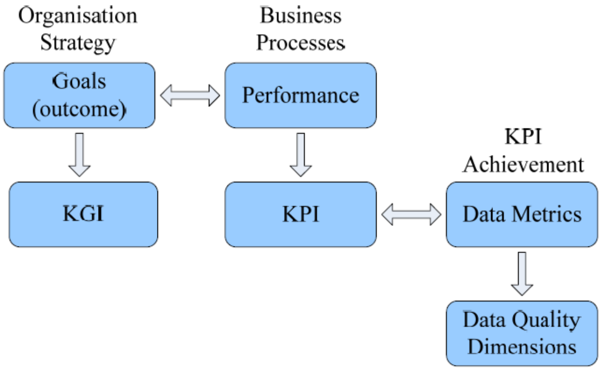


Figure 2 Interrelation between KGI, KPI, and Data Quality (Masayna, 2006)

Rodriguez et al. (2009) then continue describing three types of problems which are derived from the earlier reviewed work of Wand & Wang (1996) regarding data quality in event logs. The first type is called the 'meaningless state' and suggests that while an event has been included in the data log, there is doubt whether the event has actually occurred. The second type, labelled ‘uncertain data’, also assumes that while an event is present in the data log, certain values are doubted to be correct and thus whether it accurately refers to a real-world event. Finally, the third type is described as ‘incomplete representation’, and is the opposite of the first problem type. While one may think that a real-world event has occurred, there is no trace of it when going through the data logs. Another possibility could be that an event such as *ProcessEnd* has been recorded, hence that at some point a *ProcessStart* event must have happened but cannot be traced directly.

Data quality is therefore computed by using the metrics of probability, reputation, and confidence intervals in order to determine whether values of data attributes are realistic for it to describe a real-world event, if the data source can be trusted, and if the exact value of a data attribute can be determined to some extent. For instance, reputation is used by going through earlier cases where meaningless states or incomplete representations have occurred within the data given a certain type of event. A data source having trouble with sequential recording of event logs where processes stop and start gains a bad reputation to seasoned users for whenever such event logs need to be reviewed; however, this is considered to be an objective assessment of reputation. A subjective method could be explained by taking the example of different business departments or outsourcing units. The greater the distance between the two business parties, the less trust there is between the two regarding data quality. Rodriguez et al. (2009) give as argument that while business departments may share corporative data processes, these are considered to be separated and thus ‘hidden’ from each other between a business department and an outsourcing unit, creating visibility and reputation issues. Confidence intervals could be introduced for whenever a range of possible values is available, assigning each value with its probability of occurrence, and is easily applicable for when first names are presumed to be misspelled.

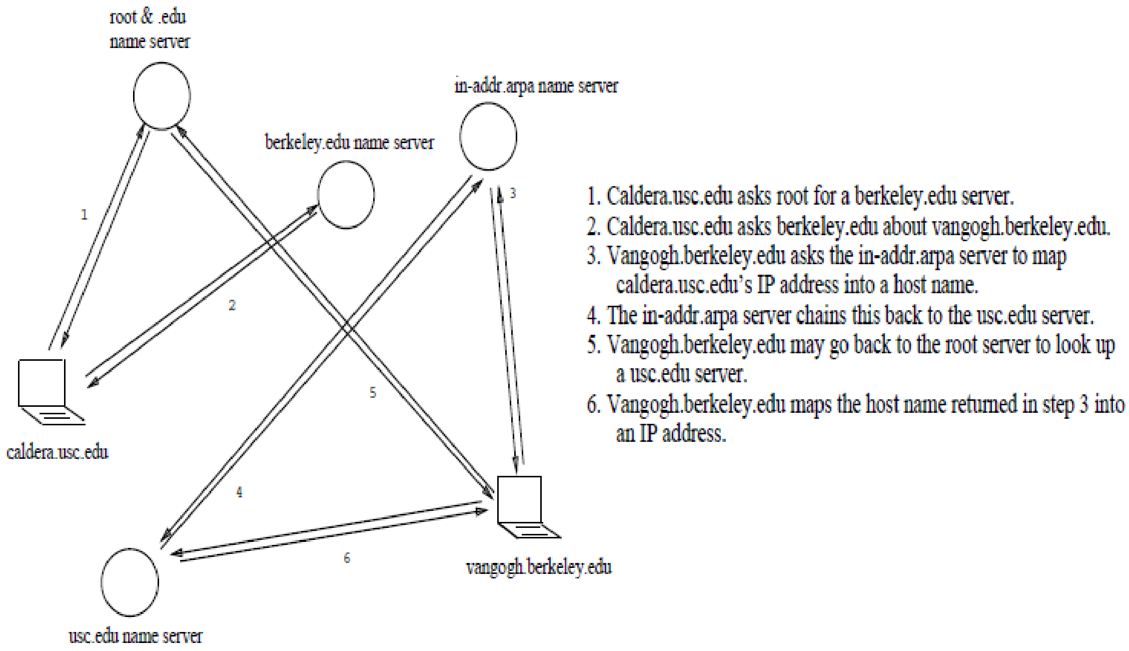
The final point addressed in their discussion is that users should be made aware of the level of data quality through the graphical presentation layer of their data collection and analysis tool via an indicator of some sort. Ideally, when data quality is indicated to be poor, users should have the availability to drill-down to the possible root of the problem and determine the type of data quality problem that has occurred. This also allows users to filter through data for which the indicator suggests that none up to a few (minor) data conflicts have been detected. While accuracy, completeness, and timeliness are data quality dimensions frequently mentioned throughout academic literature, Rodriguez et al.’s (2009) contribution is the addition of pairing KPIs with data quality by taking into account reputation of multiple business data processes, and the level of trust toward data sources paired with the number of previously occurred data quality problems. With this, an evaluation of data uncertainty can be computed for the data designated for the calculation of KPIs.

2.4 DDI Systems

To relate these insights to the problem statement of the thesis, it is important to understand via a brief overview what DDI technology consists of and how its functionalities can relate to data quality. While the term DDI is not always preferred by some, e.g. the white paper of Rooney (2013) explaining the overall best practices of DDI, the term is an acronym for three other acronyms: DNS, DHCP, and IPAM. Of course, each of these will be explained briefly in this chapter in their respective sub-sections. After each element of DDI has been discussed, the white paper shall be used for further discussion as of how these technologies are strongly related to each other.

2.4.1 DNS

The paper of Bellovin (1995) provides a brief overview of the Domain Name System (DNS) used within the DARPA Internet, which later supported the development of the ARPANET, which used packet switching as a method of digital communication. He describes it as a ‘distributed database’ which is used to pair IP addresses with host names both ways for identification purposes. While his description is fairly technical, the explanation given to how a server may delegate authority to other subdomains on other servers is relevant. A server hosting a so-called top-level domain may itself have information on lower-level domains, but may be dependent on the responsibility of a larger domain. Danzig et al. (1992) give a simplified explanation and describe DNS as a naming service for host names to IP addresses and vice versa. Furthermore, it is responsible for seeking background processes regarding e-mail transfer. A six-step overview is given by the authors to illustrate how DNS traffic works when one user attempts to visit a website *caldera.usc.edu*.



1. Caldera.usc.edu asks root for a berkeley.edu server.
2. Caldera.usc.edu asks berkeley.edu about vangogh.berkeley.edu.
3. Vangogh.berkeley.edu asks the in-addr.arpa server to map caldera.usc.edu's IP address into a host name.
4. The in-addr.arpa server chains this back to the usc.edu server.
5. Vangogh.berkeley.edu may go back to the root server to look up a usc.edu server.
6. Vangogh.berkeley.edu maps the host name returned in step 3 into an IP address.

Figure 3 DNS Traffic for a single login (Danzig et al., 1992)

Rooney (2013) stresses the importance of DNS services since they provide functionality for translating text-based addresses, such as an URL used in the example given above, into their respective IP addresses allowing for communication between computers so that the end-user may enjoy web browsing and e-mailing. The challenge here is that devices having IP addresses can either be configured stably or dynamically, which is related to DHCP and will shortly after be explained. In essence, DNS needs to have a strong sense of legitimacy and security; since it is responsible for securing a connection with a website, websites may more often than not perform what Rooney (2013) describes as a *reverse DNS lookup*: the website seeks for an IP address, as can be related to the third step of DNS traffic of the figure above. While multiple types of servers and routers are suggested to have static IP addresses, more accessible devices to end-users such as printers, smartphones and computers are likely to use dynamic-configured IP addresses. This creates complexity for correctly pairing IP addresses with host names and can create challenges for IP address managers.

2.4.2 DHCP

Dynamic Host Configuration Protocol or DHCP in short, is usually coupled to the Transmission Control Protocol/Internet Protocol (TCP/IP) for dynamic networks with wireless computers such as laptops. The paper of Perkins & Jagannadh (1995) discusses how, back in the day, wireless communication was considered to be a standard for future machines. The main functionality of DHCP would be for devices to be connected to the Internet without the need of a static IP address given in advance. The authors discuss how with the use of DCHP computers would be able to connect to Internet addresses related to the infrastructure; a computer at a company should not be granted access to content that deviates from the work at hand. At the time, IP addresses were coupled to a specific domain name (DNS) for those hosts which were holding an Internet addresses. Wireless computers connecting slightly differently each time while its identity remains the same can cause a conflict, according to Perkins & Jagannadh (1995). DNS-names would have to be 'dynamic' in order to computers to be associated with multiple, different IP addresses.

DHCP would allow for non-regular types of users to connect to the Internet without knowing the full details of how the local area network (LAN) of a particular organization is configured (Perkins & Luo, 1995). An Internet address is given from a service machine present at the organization ensuring that these are appropriate for the network. The aim of DHCP is to enable 'global connectivity' while also lowering efforts from the administrative side related to network support groups. The dynamic aspect is explained from the idea that laptops could be moved around while still being connected to the Internet by releasing previous Internet addresses and acquiring new ones while still being built around the client/server model principle, perhaps using time intervals. Furthermore, flexibility in configuration options allows for devices to be connected almost instantaneously, creating the 'plug-and-play' effect since device configuration would not be necessary. For this to take effect though, the client must be able to connect to a related server. The contrary of dynamic leasing would be IP reservation, and allows for administrators to manually configure settings for stationary devices, such as printers, routers, or firewalls. Though, dynamic IP leasing eliminates such manual configuration. The following figure demonstrates the IP address request process via DHCP.

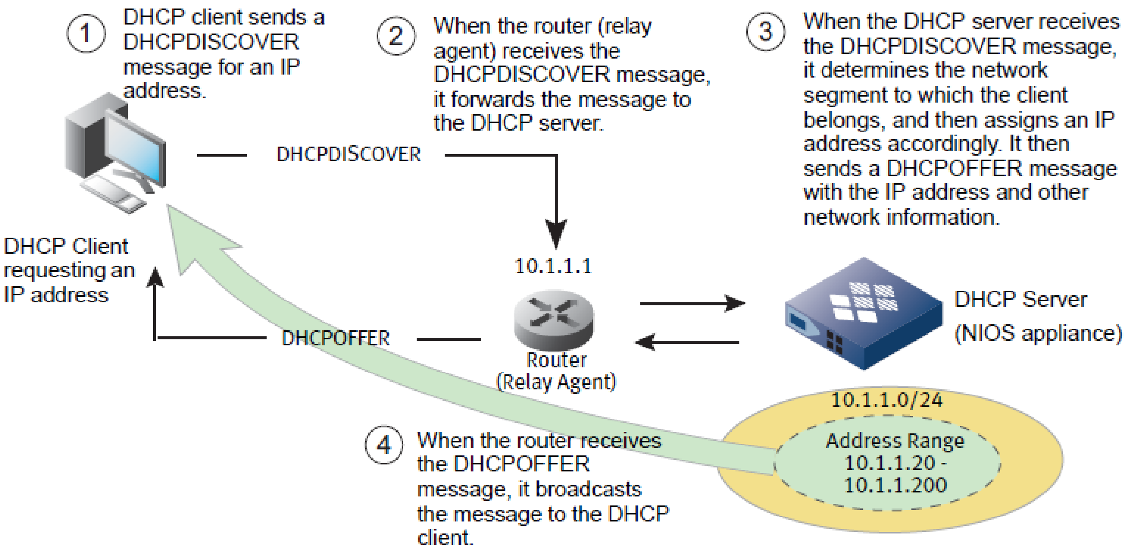

1
DHCP client sends a DHCPDISCOVER message for an IP address.
2
When the router (relay agent) receives the DHCPDISCOVER message, it forwards the message to the DHCP server.
3
When the DHCP server receives the DHCPDISCOVER message, it determines the network segment to which the client belongs, and then assigns an IP address accordingly. It then sends a DHCPOFFER message with the IP address and other network information.
DHCPDISCOVER
DHCP Client requesting an IP address
10.1.1.1
DHCP Server (NIOS appliance)
DHCPOFFER
Router (Relay Agent)
10.1.1.0/24
Address Range 10.1.1.20 - 10.1.1.200
4
When the router receives the DHCPOFFER message, it broadcasts the message to the DHCP client.

Figure 4 DHCP Process Steps for IP Address Request (Infoblox, 2012)

Perkins & Leo (1995) also discuss security issues with DHCP and their criticism on the protocol design neglecting potential malicious attacks. At the time, since wireless networks were still a new concept, the dependency of organizations to logically structure the connectivity of all devices to internal network was still at a high rate; encrypting data links between wireless devices and the network was still overlooked at. Intruders infiltrating the internal network pretending to be an internal client of the network by establishing a link of some form would be able to gain access through DHCP. Data traffic between DCHP servers and clients could be disrupted in the form of sending incorrect Internet addresses potentially causing damage to the organization. The authors mention that such scenarios could be avoided with a configuration in the DHCP server to only accept recognized and accepted media access control (MAC) addresses which act as an identifier for network communication. However, this would be in contradiction with the idea of using DHCP; the goal of minimizing administrative work to verify each connection with the network would be debunked using such approach for security purposes. Finally, Perkins & Leo (1995) address another security related to DHCP and DNS combined: if an intruder were to gain access to names that have been recognized by the organization's DNS along with the address given by a DHCP server, that intruder would gain information on how domain names are mapped and the security breach will reach beyond that of the local network.

2.4.3 IPAM

The paper of Roberts & Challinor (2000) describes how IP address management (IPAM) for IPv4 addresses should be held responsible to regional organizations and internet service providers for both organizations and individual customers. Their concern for inefficient IPAM was that the number of available IP addresses would run out sooner than when it would be carefully organized. Of course, with the roll-out of IPv6 allowing for more possibilities and combinations to create an IP address, the concern for having not enough IP addresses in the future to go round should be neglected for now. Nevertheless, IPAM is useful for organizations to get a better understanding of their network structure and IP routing practices. Referring back to the white paper of Rooney (2013), who uses the definition '*IP address inventory management'* rather than IPAM, a distinction is made between address *planning* and address *allocation*.

Address planning involves organizations to predict the capacity, or number of IP addresses, needed in order for their network to be fully operational. Internet service providers could allow for a block of IPv4 or IPv6 addresses to be obtained in order to keep a clear and consistent structure of IP addresses throughout the organizational network. The

IP address planning is dependent on the number of end-users on the network, both stationary users (e.g. employees) and mobile users (e.g. visitors). Again, Rooney (2013) expresses his concern that IPv4 addresses eventually run out soon and thus, a call for proper allocation of IP addresses is given so that administrators can easily perform network management tasks which require the need for IP addresses, such as effectively placing the network of one department into quarantine in case of a security breach. Address allocation on the other hand focuses more on routing; the author mentions the term *Open Shortest Path First* (OSPF) which refers to a routing protocol: traffic from one address to the other should happen as efficiently as possible by proper IPAM. An example is given by Rooney (2013) to showcase how to use a root address in order for consistency to be achieved across multiple routers and continents in which an organization could be operational. Every IP address in this example starts with 68.177, the root address for a single large network, while the remainder leaves address space for the host such as for every machine to be connected to the network (Roberts & Challinor, 2000). The root address is divided into three smaller blocks for the continents of Europe (68.177.0), North America (68.177.128) and Asia (68.177.192). This way, each continent is provided with a block of possible IP addresses for their departments. Such consistent structure of using the same root address allows for easier network configuration by administrators. Finally, Rooney (2013) mentions that the consequences of improper IPAM could be the assignment of duplicate IP addresses across the network causing communication and identification errors, and the possibility of sub-networks to be inaccessible. This could occur when neglecting a hierarchical network structure. When an IP address from one address block accidentally gets assigned to another network region than intended, this can create problems in terms of accessibility to that particular address block. The previous illustration shows how address blocks are allocated with consistency and accessibility in mind.

2.4.4 Infoblox DDI

The question is to understand why the information gathered from DDI tools and technology is considered to be crucial. For this, it is important to know the functionalities of DDI and its components, which have been discussed in sections 2.4.1 – 2.4.3, briefly discussing DNS, DHCP and IPAM respectively. More specifically, this particular case uses Infoblox’s DDI infrastructure technology. The need for DDI can be explained by some emerging trends, virtualization being one of them. It can be explained as creating a virtual machine or server to an already existing piece of hardware; the purpose for doing so could be to emulate given situations for testing purposes. Deploying such virtual environment may become easier; however, the configuration of its related network infrastructure is still to be perceived as a manual task. Examples of such tasks could be the to assign IP addresses (as has been discussed within the IPAM section), to keep track of DNS records and simultaneously keep track of all virtual machines deployed. This is where a paradox is created – the number of virtual servers deployed can grow faster and thus there is an increasing demand of management to assign IP addresses, update DNS records and understand which virtual servers are being deployed where for what purpose. So, the question that is proposed is: *How can DDI technology create a business benefit in the form of lower management/maintenance for workers?*

The criticality of information in the DDI system may be explained by the following: not only is it to achieve particular business benefits and thus the investment and implementation of DDI technology may be justified in the long run, but also because it allows for automatic tracking of each (virtual) server with proper IPAM and DNS tracking to ensure the safety of the network – complete information from a DDI system allows organizations to understand the network infrastructure in full detail.

An important detail to keep in mind is that virtual environments can be dynamic: they can be deployed from multiple (physical) locations, they can be shut down more often than traditional servers, and are able to 'move' (perhaps in the sense of being deployed from another physical machine). Manual work related to IPAM and tracking DNS records in this sense would be quite hectic: an automated process is needed for IP tracking, IPAM, and DNS. The criticality of information in the DDI system is therefore due to the heavy dependence of the organization on the system: to move from manual to automatic processes, the organization should have strong confidence that the information fetched from the system (which will then also be used by the system for said automatic processes) should be high in terms of data completeness, integrity and accuracy, or CIA for short. The white paper of Infoblox (2015) suggests that a typical division of work regarding network management should be the following: the virtualization team handles IPAM and DNS for virtual 'resources', while the network team does the same but for physical networks only. Acknowledging that such division is not always the case in an organization, network engineers are demanded more and more to manage DDI between 'various resources', most likely meaning both physical and virtual servers. Because of this, every network engineer that can gain access to a DDI environment has suddenly entered the virtualization environment, breaking the traditional division between the two teams, and allowing for a blend in which everyone can be called simply a 'network engineer' without referring to physical or virtual assets. The criticality of information in the DDI system is therefore due to the idea that more employees, regardless of their exact role as a network engineer (physical or virtual), can gain access to a DDI environment/platform. The information should therefore meet the 'CIA' criteria for every type of end-user; incomplete or incorrect information might impact multiple groups rather than the data/information problem is stuck within one user group.

Another challenge with virtualization is lower network visibility: referring back to the IPAM section, remember how an ISP can get hold of a ‘block’ of IP addresses to distribute among its clients. A block of such IP addresses can also be reserved for a virtualization environment delegated by a possible virtualization team. Having knowledge of each IP address along with subnets and ports is needed to track every single device in the network. Of course, this allows for network troubleshooting or possible security incidents to be tracked and resolved faster. In a virtual environment, imagine how network engineers may see the virtualized environment as a ‘black box’: it may not be known how IPAM occurs within said environment. A proper IPAM tool within a DDI can therefore be beneficial to the organization as it has the potential of placing both physical and virtual networks together in a single tool; information for IPAM (think of location, MAC addresses) for virtual networks and its environment can be fetched as well. Such visibility is improved by delivering *Authoritative IPAM* – authority is spread but different teams may still collaborate with each other and see the progress of deploying and configuring a new network. The criticality of information in the DDI system is therefore due to the ‘black box’ principle: if the process regarding DNS/IPAM is unknown for within virtual networks, it is necessary for a DDI technology tool to create network visibility of such processes so that network engineers can demand a higher level of control of virtual network environments. Knowing what information comes in and out while avoiding the visibility issue of the ‘black-box’-principle can be perceived as a benefit. Other trends given besides virtualization for the changing network environment are security, the cloud, the rollout of IPv6 addresses and, regarding connectivity of devices, the Internet of Things (IoT). Because of this, we introduce a new type of environment besides physical and virtual: the cloud environment. Infoblox’s DDI system should allow for network engineers to perform DNS, DHCP and/or IPAM-related tasks while being in compliance with security policies and service availability (the term *five-9s availability* is often used and refers to the 99.999% figure).

Such integrated DDI solution, if we keep in mind the fact that automation of tasks is to be desired by the organization, should allow for network engineers to have more time on strategic tasks rather than repetitive maintenance tasks. This creates operational efficiency; fetching IP addresses and ports to be presented on a single user interface, knowing that part of their maintenance tasks could now be automated, allow for an improved workflow. The criticality of information in the DDI system is, again, related to the automation process. Covering some of the tasks of network engineers, the DDI solution could allow for lower operating expenditures (OpEx), creating a financial benefit for the organization and a part of the workflow can be done automatically. Furthermore, since data will be collected in a centralized manner for such streamlined workflow, OpEx can be deduced even further by delegating access rights for each user in different levels; each user can be part of the same DDI tool but may be placed into different groups regarding access rights and administration.

As mentioned before, the introduction of the cloud environment also means that the increased visibility feature for both physical and virtual environments is extended. Organizations now gain the freedom of creating differently configured environments while allowing for central management of DNS and DHCP. Infoblox presents its DDI- solution as a '*zero-admin database*' with great emphasis on achieving "*consistency between service and management views of IP network data"* paired with IP and MAC address storage features. Its combination of allowing both central administration and granular access rights for appropriate workflow allows for multiple tasks to be done by different types of users: the DDI-technology solution allows for auditing and reporting in order to be in compliance with IT-related standards, for instance in the field of security, and for finding solutions to trouble-shooting related issues. For instance, an example given in the white paper (Infoblox, 2015) is how direct port control in a particular network can be done through the user interface. Automating tasks allows for more time for workers to perform strategic-related tasks. Having the GUI as a powerful tool to potentially see network performance and allowing for laymen to effectively showcase the workflow of DNS, DHCP and IPAM practices in a closed-loop methodology can allow for reduced operational expenditures.

Being flexible with the scalability of devices being connected to the organization's networks, caused by for instance the emergence of the Bring Your Own Device (BYOD) principle and the increasing adoption of IPv6 addresses is what can create network reliability; to what extent do you feel that proper, easy-to-identify secure connections can be traced down while not hindering the capacity of the workload stressed to the servers of that network? In essence, with the pointers discussed, it is about making network management less complex, despite the complexity of adding virtualization and (hybrid) cloud computing practices. DNS, DHCP and IPAM should require end-to-end visibility while still knowing the end-hosts (think of BYOD, IPv6, etc.).

To delve a bit further into how exactly the improved workflow will be achieved, it is best to look at what Infoblox describes as authoritative IPAM. The data that needs to be fetched from such systems are protocol and IP address data, along with the data of the network infrastructure. Both types of data need to be integrated in one form or another, which allows for the elimination of business silos – a situation in which one pool of data is only collected and known by one part of an organization (e.g. a department), or where one department decides not to use corporate data and uses its own methods of fetching it, possibly causing inconsistencies. The visibility of the network is explained by how the different types of network data are correlated. Accuracy of data is improved by understanding how IP addresses and port usage is allocated; when a port or IP address is coupled with another entity than previously determined, this is considered to be a conflict. IP and port reservation techniques can be applied when such conflicts have appeared by the DDI system. Historical information of port capacity can be fetched from DDI technology, and historical data is likely to be stored on a network database.

To summarize, information in DDI technology is crucial to achieve automating processes. In its turn, automating process will relieve a part of the workload of network engineers so they can focus on more strategic tasks. This will lead to a reduction in operational expenditures and increase of operational efficiency. The increasing complexity of combining physical, virtual, and (hybrid)-cloud network environments puts greater emphasis on data quality as there needs to be consistency achieved by completeness, integrity, and accuracy. This will also strengthen the position of the organization in terms of compliance with IT security standards; auditing with incorrect data while unaware leads to grave consequences. DDI is used to eliminate the ‘black box’ of network operations in virtualized environments in particular as DDI technology allows for end-to-end visibility by applying a single graphical user interface (GUI). Simplified network management allows for better preparedness for scalability due to trends like BYOD and IPv6 rollout. DDI-technology with a centralized data collection approach allows for lower-level support teams to also be involved in network engineering tasks on the same user interface. This means that a delegation can be achieved for auditing or reporting purposes regarding security compliance, trouble-shooting support, and strategic plans for the organization’s networks, to name a few.

2.5 Infoblox NetMRI

Another technology component that should be looked into is Infoblox NetMRI. Again, what are its functionalities and how does it relate to DDI-technology? In other terms, how can NetMRI contribute to the achievement of consistency between service and management views of network data? Of course, as has been discussed earlier, data quality can include many dimensions and is left to the interpretation of the researcher – this thesis has applied to use the dimensions of Fan & Geerts (2012) as a baseline and consider whether dimensions can be granted a unified definition or if other dimensions should be taken into consideration which are not consistent throughout the academic work examined in the thesis.

2.5.1 Functionality of NetMRI

As can be derived from the name, NetMRI is used as a tool for ‘scanning’ networks of an organization in order to reduce security risks. However, NetMRI can also be used as a sort of prediction tool to check on network stability when changes are applied. The datasheet (Infoblox, 2014) delves deeper into the topic of virtual networks and discusses layer-2 and layer-3 virtual constructs. For simplicity purposes, these will not be addressed as the thesis is merely focused on the importance of proper network data rather than giving a fully technical description of the construction of a virtual network.

NetMRI attempts to be the solution for maintaining proper management in the areas of security and compliance in a ‘dynamic and complex’ environment which may include both virtual and (hybrid) cloud network configurations. This relates back to the explanation given to the need of DDI-technology. Prepared for IPv6 rollout, NetMRI is intended to protect organizations from malfunctioning networks when minor configurations or new network units are implemented. Another important feature of NetMRI is compliance management. While it has been discussed that within DDI technology authoritative techniques can be used within the practices of IPAM, meaning that authority rights can be altered for different types of work groups, NetMRI is more focused on the security and legislation part of the topic regarding compliance. Nevertheless, both DDI and NetMRI solutions can be seen as a form of management support, meaning that the workload for management teams can be reduced and enables the delegation of tasks to lower-level support teams using such authority system.

NetMRI keeps a historical log of device configurations so that network performance can be compared side-by-side. This allows for management to understand the impact of any change in the configuration, and a best-of-breed approach is therefore enabled. However, ‘network performance’ should be interpreted in the form of health (connectivity and capacity), security, and the level of conformance with compliance. Not only is this to reduce manual network management, but it also allows for assumptions to be eliminated; one cannot assume that a proposed solution will have complete documentation on its effects. NetMRI allows to show in a simplified user interface the effects of configurations and allows for a scorecard to be generated.

The level of human error has to be reduced by deducting the amount of time spent on command line interfaces (CLIs) or writing scripts in order to develop a solution. Usually, these types of activities are to be done by experts in the domain and can be considered time-intensive. However, given the problem statement of the thesis, the field of analytics that NetMRI allows is to be considered more relevant. Diagnostics can be performed depending on the type of task.

Infoblox describes how organizations may struggle with two types of compliance requirements: one regarding the policy of internal security, and the other about external authorization compliance. A company may choose to focus on either one or the other, but ideally both. Though, the consultation of such compliance documents is said to be done most likely only when an IT audit is scheduled. Of course, discovering vulnerabilities in terms of security shortly before a scheduled audit is not only a major inconvenience, but it showcases that networks, servers or systems were operational with that security flaw still being present during a given period of time. Undetected security issues that are in conflict with any given compliance leave the organization vulnerable. With NetMRI, it is to be desired that security and authorization conflicts are to be showcased through the user interface. Allowing policies to be known by NetMRI given a standard such as Sarbanes-Oxley and HIPAA allows for reports to be made more easily.

The analytic tools provided by NetMRI allow for a more holistic view of the network rather than narrowing the scope to a single connected device on a particular network. NetMRI is there for prevention purposes, as can be deduced from the paragraph discussing how it is strongly related to security compliance and governance practices regarding authorization. Prevention is also stimulated by the high visibility features that have been discussed in DDI as well: creating new servers on a network or managing capacity can be tracked easily by knowing the connected end devices from which location by which user.

While this allows for device identification, it should still be noted that naming convention of devices is critical to detect rogue ones; an inconsistent policy of naming devices makes it a harder challenge to separate company devices to external or rogue devices, especially in the dynamic environment described earlier. Concluding from this it can be said that NetMRI can be paired with DDI as a packaged solution.

2.5.2 Business Benefits of using NetMRI

Again, most of the business benefits can be derived from those by using a DDI solution, being visibility, security, and reducing repetitive workload. However, NetMRI brings more focus on security compliance and governance. NetMRI also allows for detection of conflicts, similar to how DDI wants to achieve consistency of network IP data for both support and higher management groups. Probably the most critical feature of NetMRI is the ability to keep a historical track of configuration settings and comparing performances based on different configuration settings. This allows for a smoother auditing process and allows the organization to gain confidence in that it can detect flaws related to its compliance rather than it to be undetected. Again, in a dynamic setting where traditional, virtual and cloud environments are applied, employee efforts of tracing down the source of the conflict and re-establishing a healthy network can be more time-intensive and will hinder operational efficiency.

To summarize, NetMRI is a network scanning tool that not only allows for up-to-date network data based on discovering network devices and monitoring performance based on certain configuration settings, but also acts as a prevention tool for decreasing network performance. This is because sets of configurations are saved on a historical log and thus a best-of-breed approach can be achieved for optimum performance. The up-to-date aspect of it allow for network security flaws to be exposed which would otherwise might have been undetected. This allows for the organization to be more secure within the field of IT compliance and audit reports.

2.6 Understanding Network Discovery Systems (NDS)

Reddy et al. (2014) discuss how there exists a framework for network management and configuration, called the Network Discovery System (NDS). As has been discussed with DDI, it allows for a more centralized approach where scanned devices (through DDI or possibly from NetMRI) can be managed by a NDS and can be stored in what is described as a Configuration Database (CDB). It is to be used both by network operators and network engineers due to its centralized approach. While the authors have developed their own software module to be applied within the NDS framework, it is strongly based on NetMRI's device discovery techniques.

Several protocols that are considered to be essential for proper network management are: (a) ICMP, (b) SNMP, (c) NetBIOS, and (d) TCP. A combination of all four techniques is also included and is simply referred to as (e) Full. The study describes how each of these protocols are related to the practice of device discovery, and they will now be shortly discussed.

ICMP stands for Internet Control Message Protocol and its functionality is to see which hosts are active on a specific given network. An echo is initiated by ICMP, containing some data. An echo can be understood as a test procedure: an identical response (also known as the echo response) has to be returned in order for the echo to be complete. In this scenario, ICMP waits for an echo response from a host in order to discover IP and MAC addresses. This is because a host should be paired with an active IP address on the network. Simple Network Management Protocol, or SNMP in short, is more of a monitoring protocol to the devices that have been detected.

The third protocol in the set, NetBIOS (Network Basic Input/Output System) is also a detection-oriented protocol which allows for different computers (or devices) using the same application to communicate with each other using a local area network (LAN). Similar to ICMP, it sends requests and expects replies from active hosts in the network using IP addresses as a query. A NetBIOS reply means that a host is detected for which the following information is fetched: its IP, MAC address and possibly its operating system. Furthermore, the NetBIOS name is an address given for identification purposes of those NetBIOS resources on a particular network.

Transmission Control Protocol (TCP) on the other hand is communication-oriented and when referred to it is often paired with Internet Protocol as it is considered to be one of the critical protocols, creating the TCP/IP term. It is responsible for the transfer of packages that, when all of them have been retrieved on the receiver's end, create a message. These packages are managed by the Internet Protocol.

The following table then presents a brief overview of the functionality of each protocol in terms of discovery of devices within a given network. The 'Full' discovery type is, as announced earlier, a combination of the previously explained protocols. One may ask the question why the 'Full' discovery type is considered when NetBIOS is also able to retrieve the same types of data. This could be done as a form of safety measure: if the ICMP echo request fails to receive a response, then the slower types of discovery type processes should be neglected already. For instance, based from the overview given, TCP discovery can also fetch data about the OS besides the IP and MAC addresses; however, one should ask the question whether it is worth the extra time to gain extra data (OS) besides the other data types that could be gained by ICMP.

Table 1 Different Discovery Protocols (Reddy et al., 2014)

Discovery Type	Returned Data	Guideline	Mechanism
ICMP	IP address, MAC address	Use ICMP for a rough and fast discovery	ICMP echo request and reply
NetBIOS	IP address, MAC address, OS, NetBIOS name	Use NetBIOS for discovering networks running some NetBIOS services (likely to be Microsoft)	NetBIOS query and reply
TCP	IP address, MAC address, OS	Use TCP for an accurate but slow discovery	TCP/IP packets
Full	IP address, MAC address, OS, NetBIOS name	Use Full for a general and comprehensive discovery	All of the above in sequential order

Now that the different discovery types have been discussed along with their distinctive mechanisms, Reddy et al. (2014) provide the steps needed for IP discovery process practices. A network engineer can decide to place a discovery request to what is called the *Grid Master*, which in its turn will forward the request to one of its *grid members*. The network environment is then scanned and the results (detected active holds) are then placed as *discovered data*. The *Grid Master* in its turn then updates the database with this discovered data, eliminating the chance of outdated databases, spreadsheets, and the like.

Such terminology is also used when discussing the Infoblox Grid, in which the *Grid Master* (and a reserve candidate in case of disaster recovery management) sends requests to its grid members. These grid members could be involved within a given virtual environment or perform externally from said environment. It is no surprise that the similarities are striking as Reddy et al. (2014) developed their device discovery module around NetMRI. Their module would wait for the NetMRI execution to be finished and for the results to be placed on the NetMRI database. Then, the data gets transformed into message-structured formats that can be seen as information to the user.

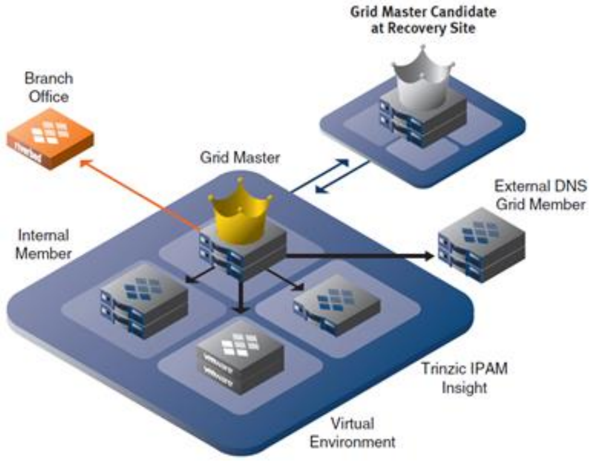


Figure 5 Infoblox Grid (Infoblox, 2015)

In order for this data transformation to happen though, there needs to be a determined data model in place which allows for the output to be presentable and easily interpretable. Going back to the IP address discovery procedure, one may put in a single IP address or a subnet. The output in this particular case could be a given status, such as *completed*, *stopped*, *running*, etc. In the worst case scenario, it may display an *error* status. Such statuses give a clear overview for the user which operations that have been requested have been performed successfully.

The IP discovery process is then described in further detail based on Reddy et al.'s (2014) developed module, which showcases a more realistic process: IP blocks consisting of 64 consecutive IP addresses are being placed as input for the network scan. Any information that can be fetched from the discovery of active hosts on the network based on the IP addresses is then placed into the database, and the Grid Master is aware of this updated database. The database illustrates the centralized data approach as multiple groups of users depend on the updated database based on constant monitoring: network operators need to detect errors on real-time and analytics for auditing reports need up-to-date information in order to assess the current state of the organization.

Reddy et al. (2014) conclude by stating that the procedure of network scanning can be an ethical issue: to what extent is it allowed to fetch information about users on a network to ensure the safety of an organization? Would users accept this kind of policy if they were informed of it? Who is responsible for attempting to fetch network-related data and who is in charge for the maintenance of the database? These are questions which are referring back to the topic of governance, and more likely than not an administrator distributes the types of rights to different work groups or individual users of such NDS.

To summarize, Reddy et al. (2014) have built around NetMRI (a *wrapper*) to create a discovery module by using the discovery engine of NetNRI and other elements based on the Network Discovery System (NDS) framework. Different discovery types exist based on what protocol is used, though ethical concerns are present when discussing the topic of network scanning.

2.7 Network Monitoring

The following section shall now describe in further detail the definition of network monitoring and the current trends and challenges that researchers have acknowledged. First off will be the paper of Lee et al. (2014) which discusses the current trends of automating processes within the field of network monitoring. It is then followed by the paper of Pras et al. (2008) bringing some more economic focus on new network monitoring practices, and share several reasons for being skeptical on network monitoring visualization tools. Finally, Schultz et al. (2011) discuss their own proposed solution for automated real-time passive network monitoring for easier investigation on network behavior during incidents.

2.7.1 Trends of Network Monitoring

Lee et al. (2014) discuss how in current times network monitoring is still largely dependent on human input regarding network operations, and the authors stress the need for a more automated technique to improve the workflow within what they describe as the *three-stage cycle of network operations*. Furthermore, their paper also attempts to describe categories of issues for different groups involved in network management. Their primary focus is on network monitoring, which shall be discussed shortly. To clarify, the paper was written for a wide audience and is therefore relatively simplified, meaning that both experts in the field as well as students can understand the core concepts of network management. The aim for the thesis is to understand the concerns and challenges of network operations from a management perspective rather than from a technical perspective. Finally, the authors present a more universal view of network management and present the relationship between monitoring, design, and deployment.

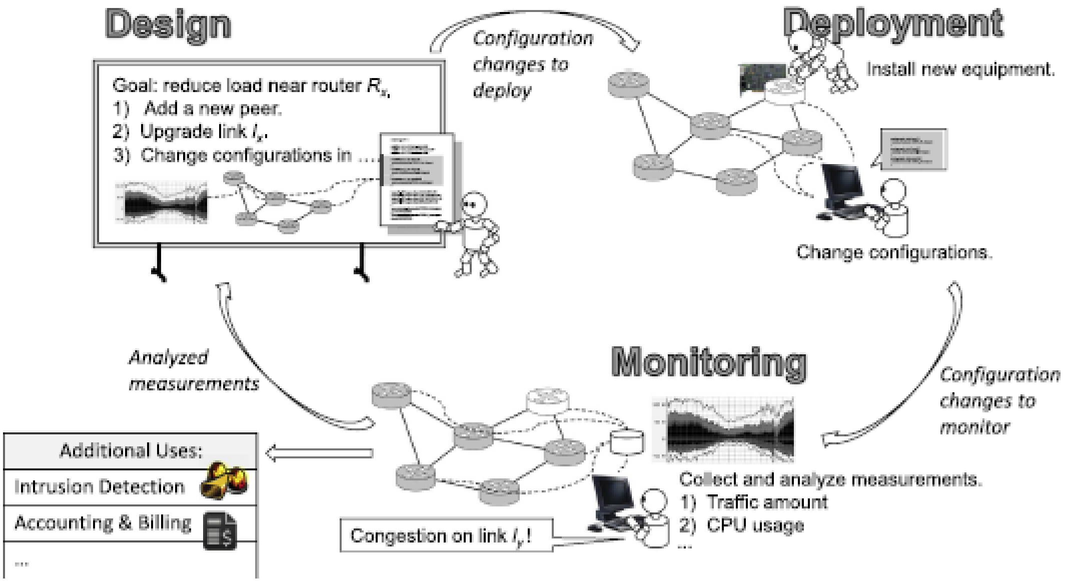


Figure 6 Network Management Operations: Monitoring, Design, Deployment (Lee et al., 2014)

Lee et al. (2014) define the cycle in three stages: (a) design, (b) deployment and (c) monitoring. As has been discussed before with DDI and NetMRI, users can choose to either measure several statistics of the network performance or can deploy configurations to strive for a higher level of said performance. Configuration is considered to be placed in its own category (design) which includes altering operations in terms of protocols and connectivity between the related devices of a network. Design changes are made based on the results from monitoring: suggestions are made for configuration changes based on the measured performance of the network under current configuration settings. These changes are then to be deployed which could involve installing new equipment. This new piece of equipment then has to be recognized by the monitoring tool in order to assess whether the design change has been leading to better performance. The cycle is illustrated in the previous figure. The authors then continue to use a term called the *network model*: it is described as how the network is currently behaving based on what can be read from the monitoring operations. Of course, this model has to be created based on several types of analysis and measurement reports, and a baseline for the performance behavior has to be created. This means that the difference between the actual performance level and the desired performance level has to be measureable in order to identify whether any configuration changes are needed. Three causes for difference in performance levels (or mismatch) can be given. The first one is related to *human error*: configurations that are deployed could be faulty designed. Compatibility issues between software and hardware equipment tagged as network devices due to future updates could be another cause. Third, due to the possible deployment of new pieces of equipment, several configuration settings that are currently in place may not apply anymore.

Table 2 Network Management Operations (Lee et al., 2014)

Operations Group	Network management functions
Monitoring	*Monitor and troubleshoot a network* Measure network behavior, identify problems Verify whether configuration changes are applied correctly and lead to improvement
Design	*Design configuration changes according to requirements* Translate needs from the monitoring reports into configuration changes Understand existing configurations
Deployment	*Deploy configuration changes to the network* Deliver configuration changes When expected performance is not met (based on monitoring), roll-back to previous configuration settings Ensure compatibility of software of network devices through forcing updates and patches

The description of the cycle has not considered any automation yet between the three types of operations. As is mentioned with DDI-technology, the automation of several network operations allows for reduced workload toward employees allowing more time to focus on higher-level strategic tasks. Lee et al. (2014) define how, similar to programming languages, an *if-then-else* clause structure can be created. The challenge here is to specify precisely the state of the network and translate them into conditions. Design options should then be defined according to the condition given in the clause. This way, a design change can be triggered automatically and the authors refer to this technique as *Policy-Based Management* (PBM). Of course, the challenge with PBM is that network operators now need to be familiar with the language including the ‘if-then-else’-clauses for when new clauses need to be defined. Second, given the increasing complexity of network environments, such automatic solutions need to be heavily customized in order for them to be functional throughout multiple aspects of network operations.

Regarding possible automation for design as a network operation, evaluation of configuration is one of three suggestions of the authors. This can be related to one of the functionalities of NetMRI: an attempt is made to check on network stability in case of configuration changes. As can be seen from the table, it might be necessary to roll-back to a previous configuration setting in case the conclusion of the evaluation comes out negative. The second suggestion can be related to the earlier concern of using ‘if-then-else’-clauses: simplifying how configurations are documented could be beneficial to network operators who are less familiar with programming languages. The third suggestion is also included within DDI and NetMRI solutions: improved visualization tools for easier contextual understanding of monitoring diagnostics.

Since the focus of the paper is primarily on the network monitoring operation, the authors continue with describing how the network model, meaning the current operational status of the network, can be assessed through network monitoring by dividing it into five layers. These five layers are the following: (a) *collection layer*, where raw data is fetched from the network; (b) *representation layer*, where the data is transferred into a particular format for storage purposes; (c) *report layer*, where the collected data gets transferred to a particular management station (this could be the *Grid Master* in Infoblox solutions); (d) *analysis layer*, where interpretation of the data can be performed by analyzing and performing measurements on, for instance, detecting faults/errors and amount of traffic on servers and routers; and (e) *presentation layer*, which could be interpreted as the (G)UI, where visual and textual representations of network data measurements are given.

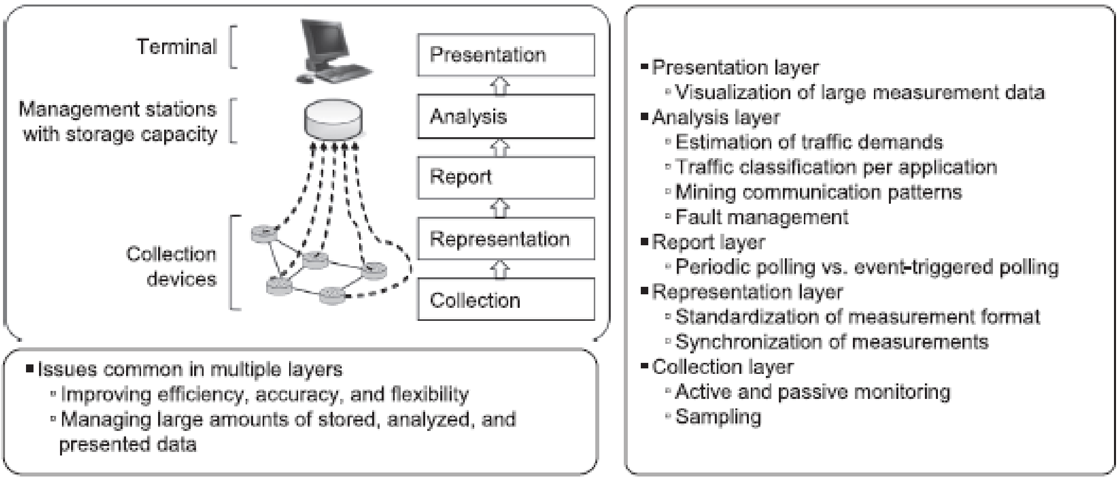

Terminal
Management stations with storage capacity
Collection devices
Presentation
Analysis
Report
Representation
Collection
▪ Issues common in multiple layers
▫ Improving efficiency, accuracy, and flexibility
▫ Managing large amounts of stored, analyzed, and presented data
▪ Presentation layer
▫ Visualization of large measurement data
▪ Analysis layer
▫ Estimation of traffic demands
▫ Traffic classification per application
▫ Mining communication patterns
▫ Fault management
▪ Report layer
▫ Periodic polling vs. event-triggered polling
▪ Representation layer
▫ Standardization of measurement format
▫ Synchronization of measurements
▪ Collection layer
▫ Active and passive monitoring
▫ Sampling

Figure 7 Monitoring Operations Layers (Lee et al., 2014)

Kind et al. (2008) describe how monitoring becomes an increasing challenge due to increasing link rates, the rollout and gradual adopting of IPv6 practices and increasing use of Internet due to more portable devices (e.g. laptops, tablets, smartphones). However, Lee et al. (2014) also discuss how, based on their literature review, significant attention is given to the analysis layer only where it has to be made clear when network operators should be alarmed of any malfunctioning within the network system. They move on with possible issues and critical choices regarding data collection for measurement purposes for each of the five monitoring layers.

Much of the focus for this thesis will be given on the collection and analysis layer since they are closely relevant to the problem statement. Within the collection layer, there is the decision whether to prefer active or passive monitoring. The latter is used when dedicated network devices have been installed in order to monitor network behavior with actual traffic to pull data from. It is considered passive since there is less interference with network behavior. Furthermore, with passive monitoring operators will have to wait for an event, e.g. an error, security breach, or odd data figures, to occur. Active monitoring on the other hand allows operators to observe what they wish for, but a drawback can be that the level of interference with current network operations can be considerably higher than in passive monitoring. Another difference with passive monitoring is that traffic simulations are used instead of real traffic; this is to allow the operators to measure network behavior on different scenarios, such as stress testing the network to search for any bottlenecks. On the contrary, usage of CPU and memory for devices is preferred to be done through passive monitoring.

However, this is not the only critical decision that has to be made considering the collection layer. Sampling of packet flows is a method that can be used for the collection process to not be overloaded. This must be done carefully in order to maintain a high level of accuracy during measurement practices. Lee et al. (2014) propose from their literature review that using a hash function is an accepted method: a hash range is selected for measurement, and packets for which the hash value falls within the range are selected. Of course, this particular range has to be determined by network operators and this allows them to control the size of packet sampling for multiple measurement tasks.

Regarding the representation layer, the format of data representation needs to be agreed upon and be converted into a standard so that multiple analysis tools can perform under the same type of measurement data. Another important factor of collecting data for measurement purposes is to snuff out redundancy: the bandwidth that is needed to collect the data and transfer it from the network device collecting the data to an accessible management station should not be reserved if data redundancy in the collection phase takes place. Another question would be the frequency to which measurements are performed.

Again, a decision can be made between two different types of frequencies and this concern relates to the report layer. Relatively similar to the decision of active versus passive monitoring, measurements can be triggered through a defined time-lapse (e.g., every five minutes) or when a certain event has occurred. This decision is called periodic versus event-triggered polling. Again, a trade-off between efficiency and accuracy is being proposed. Periodic polling allows for more accuracy as new events can get discovered, but event-triggered polling is considered to be more efficient. The catch is that such events need to be carefully specified. Any event not included to trigger a measurement will go undetected using this methodology. The report layer can also be related to data governance. The transfer of measurement data has to be performed in such a way that only those with the proper authority rights will gain access to the data in question. Authentication and encryption practices should not be neglected for network monitoring practices.

From them all, the analysis layer is the most extensively discussed in the paper. Lee et al. (2014) identify six analysis functions relevant to the layer, but only three will be discussed briefly since they are strongly relevant to the topic of the thesis. The first one is labeled *general-purpose traffic analysis*, and the suggestion is made to combine measurements into one presentable package. Looking back at the different discovery protocols described by Reddy et al. (2014), measurements can be run on the basis of an IP address, of a given port number, to name a few. The discovery process has to be consistent in the case of network-wide analysis. The second function is *automatic updating of documentation*. Discussed in the explanation of DDI and NetMRI, documentation of any measurement analysis has to be up-to-date for all user groups; manual updates will cause delays in the updating process. Non-updated documentation sheets will lead to configuration changes that are not relevant to the current network behavior, leaving current issues possibly open and potentially causing unnecessary ones in the process.

The final analysis function discussed for this layer is *fault management*, and is subdivided into fault *identification* and fault *localization*. Simply put, fault identification allows for acknowledgement of a fault within a network configuration, and fault localization showcases the source of how and where the fault occurs. The reasoning for automating both processes is cut down time on troubleshooting. Lee et al. (2014) discuss how multiple work groups may be involved in this task. If the operations staff is not able to fix the problem based on customer complaints and trouble to localize the root of the problem, the network engineers are called in in order to check on the configuration of the network machines in question. Network designers must handle with more complex issues that may involve multiple devices. Automating such processes is again a trade-off between reduced manual work and accuracy of fault localization. Given the level of granularity to which the network structure is defined, a finer level of granularity increases the accuracy as to identify the source of failure; however, to build the network graph will take more time and can be a complex task on its own. Finally, the presentation layer allows for more user-friendly visualization, possibly in the form of a dashboard to speed up the process of interpreting the data and allowing employees to create more time for more strategic tasks. Important to note is that a change in one layer could affect changes in functionality in other layers. An example given is how a new practice for analysis is likely to require a new method in terms of collecting and transferring measurement data.

Throughout the explanation of the five network monitoring layers we have seen that more often than not, decisions for monitoring practices are based on trade-offs in three areas: accuracy, efficiency, and flexibility. Sometimes, finding a balance between the three is difficult and when discussing the topic of data quality, one may think that accuracy is the prior focus. However, striving for higher accuracy could mean that there should be a sacrifice in flexibility; collecting measurement data using multiple machines means that there are less options in terms of types of analysis that can be performed. Another perspective could be that increased flexibility allows for more types of data to be collected, but the accuracy for each type of data could deteriorate slightly. However, when the objectives of network monitoring can be programmed in some language, this allows for increased flexibility as more types of analysis can now be programmed in advance. Thus, such flexible monitoring techniques allow for better efficiency and accuracy: if the programmable monitoring objective is deployed, this creates focus in terms of knowing the type of measurement performed allowing for less maintenance (creating efficiency) and a better understanding of the network model and its behavior (creating accuracy).

The question remains though whether more collected data will actually lead to improved accuracy for data quality purposes. Large pools of data only increase the difficulty for analysis and errors in the data are bound to occur more often as the data sample grows. *Consolidation* and *aggregation* are two suggested techniques by Lee et al. (2014) in order to reduce the amount of data while not sacrificing on accuracy of monitoring the behavior of the current network model. Consolidation is where the stored data of previous measurements can be used for different types of functions, such as seeking out the average value within a given timeframe. This timeframe has to be agreed upon based on a threshold; for instance, an organization could agree upon the policy to delete measurement data older than three months. Aggregation can be seen as bundling multiple flows together where data could be extracted from. This way, different data types such as IP, port number, and timestamps can be collected simultaneously from the aggregated flow. Of course, users will have to define which types of data are required for each measurement practice.

The dilemma of maintaining high data quality and extracting larger pools of data is only gaining more attention due to the increased complexity of network environments, the increasing number of devices that can be connected, and the rollout of IPv6 allowing for more network capacity. Lessons learned from this paper is that there is no one best solution when it comes to network monitoring: there are several critical choices that can be made which all influence the trade-off between accuracy, efficiency and flexibility. If we pair this insight with our problem statement of improving data quality in network monitoring systems, one could say that focusing on the accuracy of the network model would be a logical strategy. Though, it could be that flexibility in network monitoring can also stimulate data quality in network monitoring in general; think about the different types of data that need to be fetched rather than focusing on fetching error-free data from only one dimension of the entire network.

We have learned that network monitoring is only one part of the cycle: network configurations and deployment of new network equipment is delegated to other workgroups and should not be the main focus of the thesis. However, the entire network monitoring process is responsible for future configuration settings and deployment strategies; when the interpretation of measurements are based on low-quality analyzed data, this creates an invalid representation of the network model.

2.7.2 Further Challenges with Network Monitoring

The paper of Pras et al. (2007) is used to identify further challenges within the field of network monitoring as the authors believe that cooperative management in the form of automated processes combined with the human role as network operator still needs more investigation in order to identify how its benefits can be achieved. The first main issue has to do with *distributed monitoring* and the unfeasibility to monitor enterprise-wide networks with a high level of accuracy. This is because multiple monitoring *protocols* have to be established for different research purposes, with each protocol having its own requirements of the estimates (or dimensions) at place, such as the maximum allowed delay within the network, data accuracy, to name a few. Pras et al. (2007) suggest that a management application should enable such protocols can be changed at any given time in order to get the monitoring data in question to be matching the established estimates within the protocol.

However, the authors give more attention to how challenges can exist with using a more visualized approach for data analysis, and come up with four reasons why such visualized interface may not be gratifying. The first one is related to the scalability concern of network topologies and the corresponding challenge of creating a proper network topology view, as coined by Kind et al. (2008). The second concern has to do with the possible lack of a feature to allow zooming in on the network topology or data sets to explore the root case of a particular problem; again, both the work of Lee et al. (2014) and Kind et al. (2008) have described the importance of such feature.

The third issue is related toward security auditing, as the authors are concerned that the visualization tool will only highlight high-volume activity in terms of network traffic, and that unusual, small-volume traffic will be ignored. While the authors understand that having a simplified view on high-volume traffic is useful for network planning purposes, it does not guarantee that network operators are seeing all unusual activity or problems occurring in the network. Finally, the final reason given for visualization tools to be possibly unsatisfying is the limitability of the visualization tool to only allow offline analysis of data. The need for reaching almost real-time analysis of network measurement data puts more stress on the accuracy dimension of data, since the closer to real-time the data gets, the less time becomes available to assess on data accuracy. Combining this with the increasing size of data pools, capturing real-time measurement data becomes even more challenging. It is the question to what extent Infoblox tackles these challenges proposed by Pras et al. (2007).

They also make clear how with the use of semantic models, meaning that the models explicitly show the types of relationship that entities may have with each other, can allow for interoperability between different applications and allows network operators to understand data exchange between these applications better. Furthermore, the authors also explain the challenge of network monitoring from an economic point of view. Investing into increased availability of a network for instance can be seen as a form of risk management, since the outcome is still unknown since payment has to be done upfront before placing the network into operation again. Other objectives can be added to a particular list which creates the total picture for risk management within network operations; other types of risk could be packet-loss rate and overload in some of the links within the network. A choice can be made to tackle these possible risks: or there will be some economic investment in order to create a technical solution, or they are neglecting leaving the network infrastructure untouched.

Schultz et al. (2011) go further into the topic of real-time network possibilities and have developed their own *passive network appliance* (PNA) monitoring system in order to allow users to have an extensive overview of network behaviour at any given time, rather than waiting for analysis reports including details based on an established protocol. Discussed by Zseby et al. (2008), the authors give their criticism on data sampling as they believe it portrays a biased view on network behaviour. Their proposed solution tries to balance visibility, accuracy while trying to achieve real-time monitoring by creating a low-cost system tracing close to all packets in the network. This is to be achieved by developing a module that will be coupled with every packet traced by the PNA system. Both real-time monitoring and summary logs (of present or past) are to be enabled using their proposed solution. These summary logs are generated by processing packets in given network flows as soon as possible. The number of packets, the destination of the packet, active ports, and whether the packet is based from a recognized IP address are determined by the PNA system. This creates *flow data* which can be stored, and interpretation of the data creates a snapshot of the state and behaviour of the network at that given point in time. Over a long time span, multiple logs can be used to better understand typical usage of a network. Active logs can be stored on a central database for later inspection for when an inspection of a long-time period is scheduled. As for real-time monitoring functionalities, two monitoring activities are established: (a) IP-to-IP tracking between local hosts and remote hosts, and (b) focused monitoring on active local hosts.

The differences between the two methodologies are depicted in the following figure, where the monitoring process is displayed for every packet coming in and detected by the PNA system. Interesting to note is that for every IP-to-IP tracking, a *flow key* is generated from the local side, meaning that both local and remote hosts need to send a packet with the flow key included in order to verify whether it is actually a local host that is requesting to communicate with a remote host. IP-tracking not only gives information to verify local hosts and their requests for connection with external hosts, but it also allows for insight on how frequently two hosts want to communicate with each other, or the amount of data that is transferred between these two hosts. Unusual values could lead an administrator to declare suspicious network activity has been detected. In the figure, this real-time monitoring feature is described as ‘*connection monitor*’.

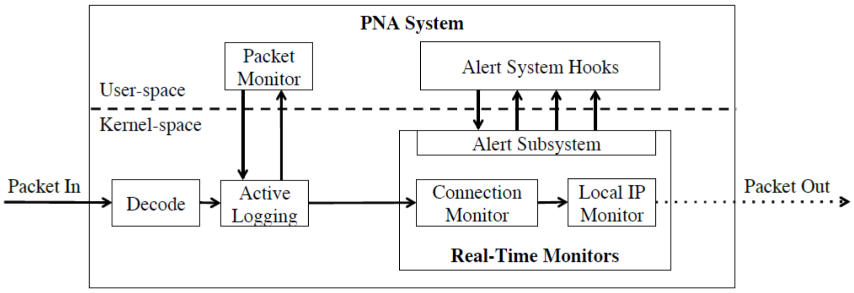


Figure 8 The Passive Network Appliance System (Schultz et al., 2011)

The *local IP monitor* on the other hand detects all activity related to a single IP address linked to the network subject for monitoring. Examples of types of data that could be extracted from IP addresses are the number of network ports available, and *send*- or *connect*-type of packets derived from interaction of an IP address with another machine or host. Both types of real-time monitors have a given threshold, meaning that at a given circumstance an alarm will trigger and the network administrator will be notified. Such circumstance could be, as mentioned earlier, an unusual high value of the amount of data transferred between two hosts, or the frequency rate of communication between the two hosts. Network administrators get information on which IP address is the 'violator'.

Schultz et al. (2011) conclude that creating a holistic view of the network should be done using the technique of taking snapshots and combining both summary logs and real-time monitoring allows traffic monitoring to be performed at the actual rate of traffic flows, rather than applying a sampling technique. Their proposed system should allow network administrators to have more time on network behaviour reports rather than actively being involved in monitoring practices, hence the reasoning for the term '*passive network appliance monitoring system*' as used by the authors.

2.8 Monitoring Maturity Levels & Auditing

This section provides information on how a framework of maturity levels can be created based on monitoring practices currently applied at organizations. Business monitoring process improvements to increase in maturity level may, according to Alles et al. (2006), also be combined with continuous auditing to keep control of data, and is supported by earlier work of Kogan et al. (1999). Two types of continuous auditing archetypes are then discussed by Vasarhelyi et al. (2004), followed by Wetzstein et al.'s (2009) publish/subscribe business process management approach and its relevance to the possible combination of monitoring and continuous auditing processes.

As will be explained later in the case study chapter of the thesis, organizations may apply their own maturity model related to network monitoring operations solemnly based on the level and scope of monitoring. As a researcher, the question should be asked whether such maturity model could be compared to other, earlier established and recognized maturity models. For instance, Lee et al. (2007) discuss the *Business Process Maturity Model* (BPMM) which discusses the capability of both monitoring and controlling relevant processes. The reasoning for choosing BPMM is because it is based on recognized standards and the *Capability Maturity Model* (CMM), making it a good candidate to be compared with the organization's developed monitoring maturity model. BPMM consists of five levels, each level discussing measurement and analysis characteristics. While the BPMM model will be shown later on in the thesis, the remainder of this section will briefly focus on how network monitoring maturity level models can be paired to continuous auditing practices, and the insight gathered will be used later on to determine the best course of action for the organization subjected to the case study research.

Continuous business process improvement may also be paired with continuous auditing practices. Alles et al. (2006) describe how a monitoring and control layer could be placed for business process controls by providing a case study of Siemens. They borrow the definition for *continuous auditing* from the *Privacy Maturity Model* (PMM) of CICA/AICPA (1999): audit reports are issued briefly following a given methodology after a particular relevant event has happened. Similar to the trade-off of accuracy and flexibility for network monitoring, Kogan et al. (1999) describe how continuous auditing can have a trade-off between orienting on control and data. If we take the example of an environment where process controls may not be automated, this could lead to loosely coupled systems. Data processing is an important element of the system, meaning that the organization could be forced to apply data-oriented continuous auditing to compensate for the current loosely coupled data processing situation.

Another reason to pair network monitoring maturity levels with continuous auditing is because the frequencies of both practices need to match. It is likely to assume that (network) monitoring capabilities increase whenever a higher maturity level is achieved. Within the field of auditing, a benchmark needs to be determined for acceptable levels of business process control settings. One may think that the higher the frequency, the better; however, performing an continuous audit every five seconds during work hours will cause some stress on the systems, and end-users may notice the impact on their performance.

A thing to remember is that although the description makes it look as if continuous auditing is the same as network monitoring due to their functionalities, continuous auditing makes sure that no exceptions in terms of, for instance, segregation of duties or vulnerabilities or loopholes in control areas, are detected and exploited. To put it simply, it monitors whether the business process controls are operating under the given audit compliance requirements. A scenario could be that a wrong configuration of segregation of duties for network monitoring practices could be currently applied. If multiple network monitoring tools are implemented in the organization, it should be defined who manages these tools and who can gain access to them. Only then can monitoring practices grow in maturity given the protection that continuous auditing provides for the organization.

Considering the architecture for continuous auditing on business process controls, there are two different types available. The system could run independently in its own dedicated layer, called the *monitoring and control layer* (MCL). It is based by the work of Vasarhelyi et al. (2004) as they claim that since such system is to be placed as an overlay for other existing systems, a middleware layer has to be established. This is to improve the level of integration of, for instance, an ERP system with separate or loosely coupled applications not included in the ERP package. An alternative would be to create a *module* that could be attached to an existing enterprise system, named the *embedded audit module* (EAM).

Using the MCL approach allows for auditors to be separate from the enterprise system and it is assumed that the enterprise data passing through the layer is not altered by personnel. The problem with MCL however is that in order to get the data, queries need to be performed targeting the enterprise system. However, this will impact the performance of said enterprise system and thus queries may not be performed as frequently as desired. This also means that there is a risk of missing a particular event during a period where queries are not performed.

Using an EAM would mean easier access to enterprise data and thus the detection of any worrying events, eliminating the need for performing queries at a high frequency; however, the data fetched by EAM can still be manipulated by enterprise personnel, and safeguarding the results based on enterprise system data will turn out to be a greater challenge. Finally, Alles et al. (2006) give another hint of their preference as implementing a MCL does not require much interaction with enterprise personnel as it is built around the enterprise system rather than attaching a module to it. The only requirement of the middleware layer is to be granted read-only access to the enterprise system data.

The reason for including continuous auditing within the section of monitoring maturity models is because the criticality of monitoring or auditing tools to fetch correct data while not interfering with segregation of duties or performance of related enterprise systems and applications can be related to the level of maturity. A higher maturity level would indicate that more sophisticated monitoring tools are implemented in or around the enterprise system or network in question, and coordination between these tools seems to be an increasing challenge as the maturity level rises. Continuous auditing is needed to make sure that no vulnerabilities or conflicts with compliance exist when enabling data sharing of sensitive enterprise data to multiple monitoring tools. This means that both monitoring and auditing practices need to be aligned correctly with the desired maturity level the organization would like to reach.

Finally, Wetzstein et al. (2009) showcase how in business process management, using the (*Web Services*) *Business Process Execution Language* (WS-BPEL/BPEL), a given *publish/subscribe channel* approach could be used for monitoring tools. This simply refers to the suspicious events a monitoring tool would like to ‘subscribe to’ in order to selectively monitor intensively during particular circumstances, also known as an event filter. While not discussed in the thesis, the authors elaborate on possible metrics and quality-of-service monitors for later analysis in order to define possible key performance indicators of enterprise systems or networks, which could simplify the continuous auditing process.

2.9 Summary

Not only has it been discussed how poor data quality may lead to a less-than-optimal business decision, numerous data quality dimensions have been discussed, with more attention given for the *accuracy* dimension. Methodologies have been discussed to improve data quality related to these data quality dimensions, such as (condinitional) data dependency techniques as discussed by Fan (2008). Network health monitoring and management systems, Infoblox DDI and NetMRI, have been disussed, along with the core elements of the Internet being DNS, DHCP, and IPAM. A brief understanding of these elements allows to focus on the possible business benefits of using network monitoring systems to investigate where data quality could be improved. Explanation has been given on how via a network scanner as NetMRI new devices can be discovered within the internal organizational network using different discovery protocols as described by Reddy et al. (2014).

Finally, within the literature review an attempt has been made to develop a fitting monitoring maturity model for the organization examined in the case study. The reason for including it in the literature review is because it should be made clear that enabling monitoring practices within an organization may be enough; historical data could be needed for root-cause analysis, to give an example. Ideally, a passive network appliance system, such as the one discussed by Schultz et al. (2011), would relieve the workload on network administrators, potentially even more if the system would be self-healing. This allows employees and stakeholders to focus more intenstively on strategic issues.

3 METHODOLOGY

In this chapter, the overall research methodology used within the thesis is discussed. To clarify, methodology can be translated as a way of approaching a research topic for further study (Silverman, 2000). In particular, the research design, case study approach, data collection & analysis practices will be described in detail in order to let the reader of the thesis understand which steps and decisions have to be taken to create the structure and flow.

3.1 Research Design

The first step when constructing the research design is the formulation of the research question, and the properties of Foss & Waters (2007) are used to create a well-defined research question. The research question is going to be the main purpose of writing the thesis: what do you want to research and how can this research question be answered best? In order to answer the research question in this thesis, it is required to understand which subjects are considered to be relevant. This means that theoretical background on multiple, relevant subjects needs to be constructed via an extensive literature review. Of course, while this may not be applicable to all literature used in this thesis, the core articles used for understanding the relevant topics are desired to come from top publishers such as *MIS Quaterly* and *IEEE*. Chapter 2 is devoted for said theoretical background as the topics of the importance of data quality and its relevant dimensions, network monitoring systems, and its techniques and challenges, were extensively discussed. A literature review allows the researcher to develop a theory for the research question at hand. However, it is important to note whether the research approach will be deductive (quantitative) or inductive (qualitative). The difference between the two will be showcased later in the upcoming sections of this chapter. What is important to know at this point in time is whether the researcher wants to perform theory building (inductive) or theory testing (deductive). The former requires an extensive literature review and the researcher to look at the problem from another perspective in order to gain new insight, and thus stimulate the formulation of a new theory. With theory testing, an established theory is tested under a new or slightly altered setting to either strengthen the theory or to discover weaknesses of the theory. If the theory does not hold up to the phenomena explained in the new test environment setting, it is up to the researcher to provide with an explanation for the failure of the theory, possibly leading to the formulation of a new or improved theory. For this thesis, theory building is applied given the performed literature review and the qualitative nature of the thesis. Important to note is that for a deductive research approach, hypotheses are defined to test a theory rather than producing one.

Once the initial conceptual and theoretical work has been performed, the stage is set for further collection of data. The thesis shall be using a case study to strengthen the theory building process. Once again, the approach given to selecting, describing, and fetching data from the case has to be defined, and can be found in the following section. The data collection and analysis phase is described in its own section. Once the data collection and analysis phases have been completed, the researcher has to refer back to the original conceptual and theoretical work. The findings of the case have to be paired with the findings from the literature review in order to test whether the newly constructed theory achieves to deliver a complete, clear answer to the research question given defined at the beginning stages of the research plan. It may be that the researcher has to acknowledge that certain phenomena from the case study cannot be explained directly using the findings of the theoretical background or literature review, and has to either acknowledge the limitations or perform further research on finding support from literature for said phenomena. Finally, the researcher reaches a conclusion where the research question is answered through the support of the theory, accompanied by the literature review and the findings of the case study.

3.2 Case Study Approach

Before continuing with our case analysis procedure, it would be wise to clarify and justify what kind of case study will be conducted. One chapter from the book of Erikkson & Kovalainen (2008) is dedicated to case study research and explains the different methodologies applicable to case study research. A case study research is there to allow complexity into the research design, meaning that whenever the research design cannot be presented in a simplistic way, an example of a case can help understand the complexity of the issue. In this case, this would be the understanding of how to improve data quality for network monitoring or network operations in general. Using the typology of Stoecker (1991), it has to be defined whether the case study in question is *intensive* or *extensive*. An intensive case study research is more qualitative driven, as much focus is spent on one (or few) particular case(s), allowing for a more ethnographic approach. An alternative is extensive case study, where the aim is to identify common patterns within the cases to elaborate and test theory. To be explicit, Eriksson & Kovalainen (2008) mention that cases can be used as ‘instruments to explore business-related phenomena’, henceforward a possible generalization to apply to other business contexts. This would suggest that related to theory building, multi-case analysis would fit grounded theory approach of Glaser and Strauss (1967). The aim of the case study is to explain the phenomena that happen throughout an individual case with intensive analysis.

To clarify the generalization issue, case study research is not fit to produce generalization for *populations*, a term used in statistical analysis and generalization. However, to generalize towards theory is what is called analytic generalization (Yin, 2013). For this, a well-grounded theory is needed along with propositions, or research questions, to be tested upon that theory, which is more fitting with the extensive case study profile. Although Erikkson & Kovalainen (2008) do not recommend to use correlational hypotheses into a case study, it is accepted that under special circumstances, attempting to prove the existence of a phenomenon that causes a correlation between variables is accepted.

Continuing with Stoecker's (1991) explanation, intensive case study research is qualitatively focused where the understanding and interpretation of the case is heavily emphasized; the author speaks of a sense-making process where the researcher's job is to define the case as such as if (s)he is actually there, understanding the perspectives of those fulfilling their roles involved in the case. Dyer & Wilkins (1991) do stress that the focus of interest should be on the case itself and not necessarily on the theoretical propositions given. Erikkson & Kovalainen (2008) give their reasoning of preferring an intensive case study, it being that the case in question could be so exceptional or unique that pairing its findings with other cases would not be advisable. Their advice for using such approach is, besides defining research questions in advance, to remain open to see what other elements of the case could be considered for further research.

Geertz (1973) speaks of a '*thick description*' when contextualizing the case study. This means that between all the rich detail that a case can offer, the reasoning for selecting the case is made clear. When it comes to interpretation, the ones of the researcher and potentially those involved in the case should be described. The researcher is ultimately responsible for developing the construct of the case which will then be used for analysis. Stoecker (1991) mentions how intensive case studies are usually written in such a narrative form where the development of the case is shown through a timeline, which is likely to be the approach taken for the thesis.

When dealing with generalization, this is not the goal of an intensive case study. Instead, the goal is to showcase how this case study should be seen as an example of a fundamental unit of some sort, upon which analysis is performed later. The challenge is to present to any reader how performing analysis on this case can be considered critical, unique, or simply interesting. Finally, Stake (1995) describes how an altered form of generalization could be achieved using an intensive case study approach, called naturalistic generalization. It is based by placing yourself as the reader and try to understand how the presented data can create an experience with the reader so that an understanding of the case can be achieved. This helps the reader to make associations between different elements within a case study without the need of presenting some form of model with variables included. In essence, it means that the reader of an intensive case study report could be 'thinking along' with the researcher.

Another feature to address is that case study research allows for both qualitative and quantitative approaches can be included allowing for triangulation in both data collection and data analysis. Erikkson & Kovalainen (2008) confirm that qualitative data could be used at the beginning stages of case study research in order to formulate the focus points of the study. However, they also speak of a complementarity approach where both quantitative and qualitative methods are placed side by side simply to provide more detail of the description of the case.

Delving deeper into the topic of triangulation, Jack & Raturi (2006) describe the pros and cons of one triangulation category, called *methodological triangulation*. The term triangulation has been developed by Denzin (1970) and refers to how several research methodologies can be applied for the same phenomenon or topic of interest. A study should consider triangulation when the research question allows for answers from multiple perspectives; using multiple approaches to narrow the conclusion into one constructive statement would be beneficial for the quality of the research. While different types of triangulation exist, such as data triangulation discussing how data collection and analysis procedures could be done in multiple ways, our main interest is methodological triangulation where more than one method or source regarding quantitative or qualitative data is considered for a single research. One of the points Jack & Raturi (2006) make is that a strategy should be developed beforehand when considering triangulation within a given research project. For a case study approach focusing on a single organization, this organization could be perceived as the unit of analysis where through an ethnographic study the behaviour and key decision points of multiple stakeholders can be considered for the narrative of the case study. Another element to consider regarding methodological triangulation is which types of validity researchers would like to reach. The balance between internal and external validity shall be discussed shortly after, but validity could be broken down into other types. For instance, when using quantitative secondary data, such as from an earlier case, statistical conclusion validity is one type of validity that could be achieved. Again, the type of validity achieved is dependent on the allocation between quantitative and qualitative data methods and their purpose for the research. However, it should be mentioned that triangulation also has its limitations. Taking into consideration that the case study used for this thesis is specific to one organization and its related industry sector, secondary data pulled from case studies discussing organization from other industry sectors would not be useful for triangulation purposes. Usable quantitative or qualitative data paired with results would therefore be limited to a select population of firms. Though, triangulation helps to eliminate possible bias in the findings of the research.

Using Okoli & Schabram’s (2010) typology of literature reviews, this one would be considered to be the most common one, it being a literature review for the theoretical background, to build a foundation for grounded the primary research. Critical to this approach is that the review of the literature must contribute to offer critique to the theory, which brings to the possible need of a rival theory.

Another aspect of the research strategy to mention is whether the case study is either prospective or retrospective. In prospective case studies, certain criteria might be established and as new cases emerge, they may be added to the study. In retrospective case studies, historical cases are included in the study and from there out, criteria are established to determine what other historical cases may be part of the study as well. In this thesis, a single, newly emerging case will be taken into consideration, but no other cases shall be added in the process. The case study would be considered having a retrospective approach when the selection for the case would be determined on the initial search for relevant literature within the given scope. However, case selection took place before preparing the literature review, thus debunking the possibility to categorize the case study approach as retrospective. Indeed, the case study can be perceived as prospective since the purpose of the research is to investigate data quality issues within network monitoring operations, a topic that has not been researched or documented extensively within the organization for which the case study is relevant. Bitektine (2007) describes how in a prospective case study approach with a quantitative theme, hypotheses based on the theories and discussed literature review are given before any data collection or analysis of the particular case has been performed, thus protecting the outcome of which case has been selected, which theories have been chosen, and how the hypotheses were stated.

Internal and external validity also have to be discussed regarding our research methodology. Knowing that internal validity is related to the extent of how our own chosen variables to measure during the case study would be considered to be the only one that creates the result or cause-effect relationship, we can say that internal validity will most likely be achieved within the given topic. To clarify, the level of internal validity explains to what extent the research scope can explain the phenomena that have happened. For instance, can the occurrence of an event be explained only by the controlled, independent variables included in the research, or is it possible that there are external factors that influence the selected variables, making them dependent? Placing it in the context of the research topic, as a researcher the following question should be asked: is it believable that the conclusion of the organization’s position regarding data quality is fully explained by the observations made via a network monitoring system, or are there any other causes that could influence the organization’s understanding of their data quality position?

External validity on the other hand asks the questions whether the research results can be generalized and placed into other settings. For instance, can the conclusions and suggestions that will be made for the organization’s case be applied to other companies who use Infoblox DDI and NetMRI solutions? Based from the literature review, insufficient data quality can be interpreted in multiple ways depending on which data dimensions are being considered. It is also very unlikely to think that the problem discussed in the case study will be identical to current problems at other organizations, let alone whether these organizations are operational in other industries. In conclusion, this means that the aim of the thesis is to achieve a high level of internal validity and a low level of external validity.

To showcase the credibility of the research, the inclusion of a rival theory may be needed in case of inconsistent findings within the case analysis. Rival theories could be established when looking at the nature of the theory: is it focused on an organizational level or only on an individual level? Looking at the nature of theories may help why similar cases may have different outcomes; was the failure caused on a global organizational level, or was just a small group of individuals responsible?

Futhermore, the difference between acquiring primary and secondary data needs to be briefly explained. In primary research, the researcher is the original data collector for the purposes of the research. Methods of collecting such data could be by conducting interviews, publishing a open survey, or simply by observing a target group and noting any intersting points of activity. Secondary data is already available and thus, the researcher is not the original data collector. Reasons for using secondary data could be for time saving purposes for when a report has to be prepared shortly. Presenting a usable case study instead of commiting much time in developing an own study could be a suitable solution for such occasion. For instance, data collected from a prevoius case study could be used as an example for another academic study, transforming the data into secondary data. Both primary and secondary data can be used in combination in order to determine whether changes have occured in the phenomena observed. For instance, a case study from a couple of years ago could be replicated in order to possibly detect these changes.

It has to be mentioned that using secondary data does not come without risks. Of course, the data from the original researcher may be of poor quality, but even riskier, it could be that secondary data is misued, meaning that the data was orginally meant for another purpose. The researcher should perform a careful consideration whether its interpration of the secondary data fits the purposes of the research. Another benefit however is that the validity of the data has already been tested by the original researcher. The choice of selection regarding the type of data that was measured is supported by the literature review performed by the first researcher; this does not have to be re-established by the researcher who would prefer to use it as secondary data.

Having explained this, it could be conluded that with using the prospective case study approach, primary data will be collected as the case is unique. While this may indicate that that there is no risk of data being outdated or inaccurate for the purposes of this research, it does carry a risk that there is no initial support from the literature review, and thus this becomes a challenge for the thesis. Another challenge then is to establish the level of validity and reliability of the research based on the data collected from the case study and the conclusions made.

3.3 Data Collection & Analysis

Dubois & Gadde (2002) describe how research design should be flexible enough for the data collection phase in order to follow the research questions with more accuracy, and urge the need for combining available theories based from the literature review with the case in question that continues to evolve during the research period, and the analytical framework that the researcher has to apply. Their theory is that both developing and confirming theories should go hand-in-hand for a case study research, rather than the separation between developing and testing a theory with the use of a case study.

Yin (2013) speaks of an analysis strategy where the case description forms the basis for the research questions, possibly causing new research questions to be emerged during the timeline, where direct interpretation practices are applied (Stake, 1995). This means that the analysis strategy is *inductive*, meaning that, as a researcher, the patterns and activities that occur within the case are to be discovered without having a framework or proposition already developed beforehand. Stake (1995) also suggests to use an *issue question* rather than an evaluative question, in order to strengthen the inductive nature of the case study. To clarify, this does not mean that established theory should be neglected during the data analysis phase. Instead, this theory can be used as a form of reference to record empirical data for the events that may occur in the case study (Blumer, 1969; Eisenhardt, 1989).

Continuing with Yin's (2013) work, knowing that the case study approach for the thesis will be of a single case, four different analytic techniques are available. The technique that seems most suitable in this situation is called *time-series analysis*. Shortly explained, this form of analysis looks at several events over a given chronological timeline. Derived from the problem statement, the goal is to measure whether or not data quality will improve over a given set of time. 'Events', in this context, could be the points of significant configuration changes in the network or every thorough monitoring scan, and have to be defined clearly within the case description or during the data analysis section.

Regarding the data collection method, a mixed-method approach of quantitative and qualitative data sources will be used in order to achieve triangulation, due to the complex social environment that is related to the problem statement. Not only is it mandatory to answer the question how network data flow could possibly be improved by using a network maintenance tool, but also *why* it should be considered important by organizations. As will be discussed in the next chapter, the network maintenance tool available at the organization is capable of generating reports for DNS, DHCP, and IPAM. Using such reports over a given timeframe would contribute toward describing the current and possible future status of the network behavior and overall performance in a concrete, statistical manner. Furthermore, multiple meetings were scheduled face-to-face with on-site supervisors to identify what the original problem at the organization was regarding managing and monitoring network data, and what goals have been established that should be fulfilled by implementing the new tool, often identified as key performance indicators (KPIs). This allows for a better narrative description of the findings of the case study.

As for the generated reports by the network maintenance tool, it has to be discussed which of these reports will be considered for the purposes of the thesis, and which metrics are going to be applied for measurement purposes. Simply stating that the network has a daily total DNS query rate of 150,000 is not going to deliver much meaning to non-experts, nor does it give a complete picture of the health of the organization's network, only a hint of its capacity.

There is a general consensus that the purpose of composing metrics for network data to DNS, DHCP, and IPAM is to allow for data reduction. Providing a data summary value for each type of measurement would simplify the analysis process. Two relevant types of compositions are proposed by Hanemann et al. (2006): (a) aggregation in time, and (b) aggregation in space. The first metric composition is related to *trend analysis*: over what given timeframe does one wish to analyse a given phenomenon? This allows for flexibility in terms of granularity: should this timeframe be a month, a day, an hour or continuous for every five minutes? The second is aggregation in space, and can be related to the network grid members. Would it be considered useful when data summaries are given for the entire network rather than for each separate grid member? In other words, for which time period does the data have to be fetched from the generated reports and analyzed, and for which grid members? A section in the next chapter is devoted to typical network performance at the organization using collected data from the period of Monday January 4th till Friday April 29th, 2016. Historical data regarding DHCP was available from September 17th, 2015 till present time, but given the deadline of the thesis, the cut-off date was also set to April 29th, 2016. Other network data generated reports were available via the Infoblox DDI user interface and could be filtered on multiple data attributes besides start and end time, such as homing down on an individual IP address, host name of a device, or device class, to name a few. The following chapter shall now introduce the case study.

4 CASE STUDY

Having discussed the theoretical background and methodology to possibly formulate an answer to the presented research question, a case study shall now be described of an multinational financial services corporation where the researcher was entrusted to view and analyze data of all international networks of the organization. The purpose of this case study is to showcase the motivation as of why a network management system was put in place at the organization, what its direct measurable effects were by using KPIs, and how data quality pulled from the system can still be improved upon by analyzing discovered data quality issues while using it. The case findings are then discussed in the following and final chapter of the thesis, where the answer to the research question is given and the conclusion is formulated.

4.1 Case Background

The organization that shall be the topic of the study is a global financial institution, operating since 1991 after a fusion between multiple Dutch companies. The organization is primarily focused on banking services for both customers and businesses. Research was done at one of the headquarters in Brussels, Belgium. The organization states that they can be considered market leaders in multiple European countries regarding both retail and wholesale banking services. Their current focus is to invest on digital leadership and deploying a *direct-first* model by achieving operational excellence and streamlined business processes. Part of their strategy for 2017 is to improve on the cost/income ratio by cutting on operating expenses in IT, with a aimed target of 0.5-0.53 (or 50-53%). Based from the latest annual report of 2015, the current cost/income ratio is currently set at 0.559 (55,9%), and has been fluctuating between the 56-60% range in 2012, 2013, and 2014.

The organization has faced a reconstruction during the period 2008-2015 due to the financial crisis. It had received from the Dutch State financial aid in order to develop a restructuring plan. Some of the initiatives taken were to sell its personal banking branch in the United States and to deprive some its own businesses in the United States, Asia, and South America via an Initial Public Offering (IPO). Two of its businesses in the insurance sector were placed on the New York Stock Exchange (NYSE) and the Euronext Amsterdam Stock Exchange for which all its shares are expected to be sold before the end of 2016. Repayment of the State’s financial aid was done in increments between 2009 and 2014 for a total of €13,5 billion, allowing for the organization to start distributing dividends again to shareholders. The 2017 financial targets of the organization show the organization’s eagerness to strengthen its standalone position.

While it has been described in the theoretical background chapter that the introduction of a network monitoring tool would release some of the workload from network operators by introducing automated processes, such as reporting, it is simply difficult to estimate by how many percent-points the introduction of such tool the cost/income ratio will decline. It is likely to think however that widespread implementation of such monitoring tool throughout all sites worldwide concerning network operations would decrease the operating expenses regarding network operations, and thus it would contribute towards achieving the target for 2017. Appendix C discusses possible cost curves relevant to the practice of data quality improvement and the taxonomy of data quality-related costs, using the work of Wasserman & Lindland (1996), Eppler & Helfert (2004), and Batini et al. (2007), in case further information is required on how the implementation of such network monitoring and management platform, in combination with data quality improvement practices, contributes toward lowering the cost/income ratio down to 50-53%. The reason for placing it in the appendix is because the main desired outcomes of the case study is not to determine the financial impact, but the managerial impact for the organization.

The organization in question used to have different software and methodologies to perform the DDI tasks at hand. Both internal and external DNS services were provided using BIND/Solaris, while DHCP tasks were performed by servers for Windows. IPAM documentation was done manually using Excel spreadsheets, which would often be out of date, reaching error rates up to 80%. The current solution integrated in the organization's network infrastructure allows for all DDI services to be performed by Infoblox hardware and software.

The choice of the organization in 2012 to implement Infoblox in their network infrastructure was partially influenced by their current external DNS services software and the recent changes at one of the primary Dutch sites of the organization. The Dutch site already switched over to a native Infoblox solution, meaning both the software and hardware were delivered from Infoblox, costing around €800,000. Meanwhile, at the Belgian site, the software VitalQIP delivered by Alcatel-Lucent for DNS services only was running on Infoblox hardware, since VitalQIP has the capability of being a full DDI management tool. Eventually, problems at the Belgian site emerged at software licenses for their external DNS services were expired which caused local staff to not have access to their services anymore. Authority could not be switched over to staff from another site, such as the Dutch one, since the hardware was considered Belgian and thus would cause possible conflicts regarding governance. This is where the possibility was discovered that integration of both Belgian and Dutch Infoblox hardware onto one network grid would be considered cheaper than simply creating an own network grid and keeping the two separated. Integration costs were estimated to be around €115,000 and €125,000, which covers the integration of the internal DHCP and IPAM services.

However, while integrating the question has to be asked who should have control over these services; in this example, would it be the Dutch one or the Belgian one? Referring back to how the grid of Infoblox works, one grid member should be crowned as the *Master Grid*, with possibility for a potential *Candidate Master Grid*. Since the Dutch site had already moved over to an Infoblox solution with its own appointed Master Grid, the pressure was placed on the Belgian site to push for equal ownership rights. Eventually, the Belgian site had successfully been granted its own Candidate Master Grid. Now comes the question of licensing costs regarding the contract. While both sides would share the same grid environment, they would still have their separate services concerning DDI. A cost split of 70%-30% was applied for the Dutch and Belgian site respectively. Other challenges at the Belgian site was to train its staff to be familiar with Infoblox technology for DDI practices. For instance, migrating the IPAM Excel spreadsheets to Infoblox directly was not advised as it was known that a high error rate in the data was present. Instead, a team was assigned to assess, correct, and implement usable IPAM data from the Excel spreadsheets to Infoblox in order to create a first accurate image of the network. This process had taken around six months to complete.

The organization would argument that migrating the data over from Excel spreadsheets to an Infoblox solution allows for more insight of a single data entry than in Excel. To give an example, in a spreadsheet an IP address would be associated with a DNS zone or device and no other information would be available. With the presentation of Infoblox, not only would an user be able to extract this data by inspecting an IP address, but (s)he would also have the possibility to inspect whether this IP address was included in a DHCP range or to investigate on its DHCP history in the case of it being associated to a device.

Key performance indicators (KPIs) that were identified by the organization when implementing the Infoblox solution to the Belgian site were: (a) the percentage of error rate, based on the number of incidents on DNS and DHCP services and the number of failures on IPAM; (b) the availability of support in case of troubleshooting or upgrading systems, as this turned out to be a significant problem with their prior way of working; (c) performance, which could be measured in multiple ways such as operational cost (savings) or even in latency network performance; and (d) criticality of survival for business process continuity. Each of these KPIs shall now be briefly discussed within the context of the organization.

As was mentioned earlier, the Excel spreadsheets holding data records for IPAM services would have an error rate up to 80%. The number of failures of DNS and DHCP would be largely dependent on which IPAM data from the Excel spreadsheets would migrate over to the Infoblox solution. Because of the organization’s extensive effort to filter and correct the Excel spreadsheet data, it was made possible to configure DNS zones and DHCP ranges accordingly to the organization’s established standards and policies. The error rate would also be influenced by the possible lack of support in case upgrades were applied to their systems but were not compatible to the organization's configuration settings. Important for the organization was to be guaranteed of software and hardware support from Infoblox in case errors emerged or services would become unavailable. Performance could be interpreted in multiple ways: operational performance would often be translated in *efficiency* and *effectiveness*, while network performance would best be measured by using *latency* speed in milliseconds. The organization has observed that on the Belgian site, their latency speed was cut tenfold from 200ms to 20ms on the first days after having completed their implementation. Finally, criticality of survival would mean that services would have to remain available after one part of the network system would become unavailable. Business process continuity would focus primarily on keeping services available for the external client, such as home Internet banking. An example would be to have Infoblox grid members in both primary and secondary clusters. If one primary grid member would fail, another primary member would be able to pick up the work and continue the availability of business processes and services. Even in the case that all primary grid members would become unavailable, secondary grid members would be assigned with taking over the task of business process continuity. Technically, one could say that the usage of two clusters for backing up critical work could be seen as double redundancy. However, as the organization is operational in the financial sector, business process continuity is perceived to be one of their primary key performance indicators, if not the most important.

These events occurred over a timeline of three years in total, with around a year and a half spent on migrating internal services over to an Infoblox solution and assigning the balance between the Dutch and Belgian site. Implementation time was estimated to be around four to six weeks as the hardware of Infoblox was already in place at the Belgian site for multiple years.

While these KPIs would suggest that the implementation of the Infoblox solution at the Belgian site was considered to be a success, a question remains whether the data generated by Infoblox hardware and presented on its software would be of proper quality for continuous high performance of DDI services. While the organization stated that the error rate of DNS, DHCP and IPAM was visibly less than with taking the Excel spreadsheet approach, it would not take away the fact that resolving an error would require proper data quality in order to determine the source of the problem and the best suggested solution. This is also reinforced by the reviewed literature by Masayna (2006), Maysana et al. (2007), Sufi Abdi (2013), and Rodriguez et al. (2009) where the possible link between KPIs and data quality has already been discussed and established in a framework.

For network monitoring practices, Infoblox IPAM has been implemented using the Infoblox Grid model as has been explained in Section 2.6, with the Grid Master and its reserve candidate located in the Netherlands, and its external members being located in the Netherlands, Belgium, United Kingdom, and Poland. Poland is recognized by the organization as a growth market where digital leadership and innovation are the main focus due to the healthy economic prognosis. Support teams are currently located at the site in Poland for regarding IT operations, but are beyond the scope of this case study research.

Access to the Infoblox IPAM graphical user interface is done by requesting an account via a form which has to be processed at the Dutch site. Through an access control server, access rights are managed for each user. For instance, the different roles that were defined for network operations were 'network operator', 'network designer', and 'network specialist'. For the purposes of the case study, read-only access was granted as a network designer to become knowledgeable about the current network architecture.

Whenever an alert has been discovered by the system, members within the Master Control Room (MCR) of the organization are notified and an incident management tool is used to direct the alert or incident to the corresponding team, such as Network Operations. Such alerts would be generated based on the results from NetMRI, which would perform regular network scans only and transfer its discovered data over to the IPAM tool. Later in this chapter, the topic of IP address conflict resolution is discussed, but what could be mentioned already is that an IP address conflict alert would be of low priority to the MCR team. Nevertheless, the organization has not placed any other process in place to resolve IP address conflicts besides MCR being responsible to refer the IP address conflict to other responsible teams.

4.2 Maturity Levels in Monitoring at Organization

The organization in question offers its employees a training course on monitoring and their monitoring maturity model. Apparently, it consists of six different levels, but only three have been fully developed within the organization. The first level of the maturity model is concerned with availability of the organization's IT systems by narrowing down the scope at individual infrastructure components. More specifically, what could impact the availability or performance of these components, and when are events registered and alerts triggered? Employees need to know of their daily used applications on which servers they are running on, and which monitoring tools are being applied.

Interestingly, the organization has defined its policy that anyone is free to implement any tool for monitoring purposes. However, the main concern of implementing different monitoring tools is that the data used by these tools is not shared. This means that every business unit, department, or even an individual user may have its own interpretation of the current status and behavior of the entity they are monitoring, leading to information asymmetry and more importantly, there is 'no single truth' about the monitoring results within the organization. Enabling tools to share data allows for users to determine whether any type of disruption within the infrastructure is caused by another application other than the one used by his/herself.

The collection of historical data concerning performance and availability of the applications used within a team is done on the second maturity level. Users should know which components in their applications are vital, which monitoring practices are applied, and where summary logs are stored. The scope of the maturity level is fixed to application components only to emphasize the importance of understanding the health of applications. Teams can request monitoring for their relevant applications. The scope is expanded in the third maturity level to business applications. The dashboard or interface from the monitoring tools implemented should have the capability for users to drill-down to the root cause of the disturbance within the infrastructure. Ideally, when data is shared amongst different tools, different business units can pull the monitoring data and perform analysis for their relevant applications and components. Teams need to have a set-up prepared for analysis whenever a disturbance is detected in order to quickly determine the root cause, or at least determine which major business applications are affected.

The organization then states that the fourth, fifth, and sixth maturity level have been specified, but have not yet been deployed. The fourth maturity level suggests that development operations teams work together with other teams in order to develop monitoring practices over the entire IT value chain. This requires a new type of dashboard where a clear overview of availability and performance statistics over the value chain are given. Again, visibility in possible impact of a disturbance is improved by the newly introduced dashboard.

The fifth monitoring maturity level introduced by the organization is related to transactions between applications. Monitoring tools should be able to trace these transactions and have predictive techniques applied. This means that monitoring tools should be able to predict whether, for instance, an occurrence of high traffic is simply a reoccurring event due to heavy load or an anomaly. The sixth and final maturity level is continuous improvement of monitoring practices as new possibilities become available and meet the demand of business units.

As discussed in the theoretical background chapter of the thesis, the BPMM model of Lee et al. (2007) could be compared with the monitoring maturity layers established at the organization to inspect whether similarities exist between either increments of maturity levels or the nature of maturity levels between the two models. The following table will now give an estimated comparison of maturity levels defined by both the organization and within BPMM.

Table 3 Comparison of Maturity Models of Case Study Organization and BPMM (Lee et al., 2007)

Maturity Levels at Organization	**Maturity Levels at BPMM regarding Monitoring**
-	*Initial* – specific manner (ad hoc)
Level 1 – *Infrastructure Component*: monitoring tools are assigned to applications	*Managed* – Partial measurement of process performance, processes are assigned to work units regarding monitoring and controlling
Level 2 – *Application Component*: historical data used to understand application health, improvement of analytic capabilities Level 3 – *Business Application Perspective*: easier analytic capabilities for applications by having teams to create a monitoring set-up allowing for overall organization monitoring control	*Defined* – Process performance is measured and controlled for the overall organization, partial performance data used for process improvement
Level 4 – *IT Value Chain Perspective*: dashboard provides a more holistic view, performance data from multiple steps in the chain can be used for improved impact analysis and improvements	*Quantitatively Managed* – process performance is measured systematically, and performance data is used for improvement purposes
Level 5 – *Business Value Chain Perspective*: continuous monitoring using predictive techniques, understanding of business transaction lead-times Level 6 – *Automatic Chain Perspective*: continuous deployment and adapting to new monitoring demands for optimizing business processes	*Optimizing* – Proactive monitoring and controlling, process performance data systematically used to optimize processes

While other interpretations of categorizing the maturity levels of the organization to those of BPMM may be possible, the main focus is given on the first three maturity levels of the organization and their corresponding BPMM maturity level. This is due to the statement of the organization that the fourth maturity level and beyond have not yet been realized. To explain, the first maturity level of the organization can be paired with the second maturity level of BPMM, because both maturity levels hint the initial steps for setting up monitoring practices and pairing them with certain entities. It gets more challenging to place level 2 and 3 of the organization's monitoring maturity model into BPMM. The conscious choice has been made to place both under the third maturity level of BPMM regarding monitoring because of the term *overall organization*. This term can be interpreted in multiple ways and thus becomes quite vague. The interpretation used in this thesis is that by *overall organization* it is meant the applications and infrastructure of all business units or departments recognized by multiple monitoring tools. However, value chains are not considered to fall under this term as they are considered to be a holistic economic perspective of the organization rather than a technical, infrastructural one. If value chains were to be included in this term, then a possible interpretation would be to place maturity levels 2 through 6 all in one BPMM maturity level, which would defeat the purpose of thorough comparison of the two maturity models.

Because of this, level 4 of the organization's monitoring maturity model is placed one maturity level higher in BPMM. Systematic monitoring and process performance measurement could be best performed when throughout the IT value chain monitoring tools are implemented in place and a more holistic view on impact analysis can be created. This allows for more data to be used in order to develop improvements on processes. The fifth maturity level of the organization focuses on creating a more holistic view based on the transactions between applications to ultimately develop a business value chain. The reasoning for placing it under the fifth and final maturity level of BPMM is that the description implies that it could be achieved in current time, while the sixth maturity level is focused on possible changes of monitoring practices and demand shifts from business units in the future. In both maturity level 5 and 6, the optimal scenario can be achieved where every application and entity in the infrastructure is proactively monitored and performance data is stored and used systematically for continuous business process improvement.

Now that the comparison of the maturity models have been explained, possible questions could be how this insight can be related to network monitoring practices, and at what maturity level the organization is currently at. Interestingly, Lee et al. (2007) describe how monitoring and control practices are considered to be key process areas at the second maturity level of BPMM, which confirms the choice for placing the first maturity level of the organization there. Based from the description of both Infoblox products (DDI and NetMRI) used at the organization, the potential for either being at or reaching the third and, for the time being, highest monitoring maturity level seems very plausible. Though, further information is required on whether different teams have already prepared their monitoring set-ups. As for the application portfolio used by different teams as described at level 3, for network monitoring it could be interpreted as which applications and its related servers are subject to monitoring given the recognized network topology.

The remaining question is whether the Infoblox solutions of DDI and NetMRI are using either the MCL or the EAM approach when it comes to enterprise system data monitoring and continuous auditing. Although their purpose may be slightly altered since it concerns network data rather than enterprise system data, both solutions could be perceived as middleware as they do not have to be installed alongside a given enterprise system. It is an important criteria that the data is not subject to any manipulation for network monitoring purposes as the performance throughout the entire network topology needs to be assessed. This would lead to the conclusion that both solutions are steered towards the MCL approach.

4.3 Typical Network Behavior Statistics

In order to get a better understanding of the size of the network being analysed, some network usage statistics are now going to be presented. This is to emphasize that network management practices at the organization may be considered complex given the fact that multiple grid members are assigned with different tasks within the scope of DDI practices. The first statistics that are going to be presented are related to the DHCP usage trend, or the trend of how many IP addresses are assigned in either a dynamic, static, or free state within the network's configured DHCP ranges. Statistics were pulled from the period of Monday January 4th till Friday April 29th, excluding weekend days. This is because the goal is to portray typical usage statistics on regular workdays. For each day included in the period, five data points were taken from different moments in the day for each type of IP address state. This means that for each day in total, fifteen data points were measured. In addition, the total number of IP addresses measured from the three states was computed. To clarify though, this total number does not directly translate into the total number of IP address available in the network, since it is not a prerequisite that all IP addresses should fall under a DHCP range. As mentioned earlier, the organization has developed multiple DHCP templates for their separate networks, thus the numbers presented should give a close estimation to the total numbers of IP addresses included in the whole network. The five data points were set on 5 AM, 9 AM, 1 PM, 5 PM, and 9 PM, but were changed to 6 AM, 10 AM, 2 PM, 6 PM and 10 PM starting on the week of March 28th due to European Daylight Saving Time. On the first measured day (January 4th), the total number of IP addresses included in DHCP ranges was set at 563,368. On the final measured day (April 29th), the total number has grown to 577,559. A spotted trend was that the total number of IP addresses would often be the same for a few days before suddenly rising to a new figure. For instance, between January 15th and January 25th, the total number was fixed to 536,525. The new total number on January 26th was 539,244 and remained so till February 1st. The most aggressive growth occurred in the week of February 8-12, where the number had grown with 21,420. The figures show that on average, 12,99% of IP addresses in DHCP ranges are dynamic, 0,63% are static, and the remaining 86,38% are free. While it is useful to know the average percentages of each state throughout the whole day, it is more interesting to look at how the percentages change over the course of a day using the different data points. Logically, four time periods were established using the available data, being from 5 AM to 9 AM or 6 AM to 10 AM, 9 AM to 1 PM or 10 AM to 2 PM, 1 PM to 5 PM or 2 PM to 6 PM, and finally 5

PM to 9 PM or 6 PM to 10 PM. It has to be clarified that the number of static IP addresses remains relatively constant, with minimum and maximum percentages of the total IP addresses in DHCP ranges are 0,61% and 0,66% respectively. The changes between the different time periods therefore are measured using only the percentages of IP addresses in dynamic state, as a 1% increase in dynamic IP addresses would almost be equivalent to a 1% decrease in free IP addresses, and vice versa. The following table has been computed which shows the differences between time periods for each regular workday in percent points for dynamic IP addresses only, having in mind that for the free IP addresses, the values are to be multiplied with minus one.

Table 4 Average Change in Percent Points Between Time Periods For Dynamic IP Addresses Per Working Day Between January 4 - April 30

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Avg.**
5:00 - 9:00	1,87%	1,87%	1,28%	1,54%	0,97%	1,51%
9:00 - 13:00	0,25%	0,28%	0,20%	0,15%	0,05%	0,18%
13:00 - 17:00	-0,70%	-0,74%	-0,57%	-0,76%	-0,59%	-0,67%
17:00 - 21:00	-0,68%	-0,80%	-0,61%	-0,76%	-0,49%	-0,67%

The following trends can be observed. For all days in the week, there is a significant increase in the number of dynamic IP addresses in the morning and early afternoon, as the first two time periods in the table show positive change in percent-points for all working days. This should not be considered a surprising result as it could be deduced that employees come in the morning to turn on their workstations, thus explaining the daily increasing spike of total dynamic IP addresses. Similarly, as the hours go by in the afternoon till late evening, the number of employees, and thus the number of devices being under an active DHCP lease, may decline again; however, this decline does not necessarily have to be the same as the increase in the morning. Remember that typically for workstations used by employees, the DHCP lease lasts for three days. Devices being granted a lease on Monday morning and are not used on any other day in the week will not have the opportunity to be granted a renewed DHCP lease for the same IP address. This would mean that the lease is terminated on a Thursday. Of course, this cannot be said for all devices as it has been explained that some DHCP leases could be as short as five minutes to a couple of hours. From a data analysis perspective though, it should be noted that possibly some of the percentage decline in the last two time periods could be partially explained by DHCP leases having reached the 72-hour duration and are finally terminated. It is very well possible that on weekends, percentage-points differences for all time periods considered could be negative as less employees are present and more leases could be terminated over the weekend due to inactivity. This brings us to another spotted trend that could be related to the new phenomenon of working from home. On Fridays, the percentage-points differences between time periods are smaller than that of any other workday. Less IP addresses are in dynamic lease possibly due to employees either having flexible workdays or being allowed to work from home one day per working week. It is also the only day where, if one would only consider the difference in percentage-points between 5 AM and 9 PM, the difference would be -0,06% (0,97% + 0,05% - 0,59% - 0,49%). One could ask, if we neglect weekends and the time periods between 9 PM and 5 AM, why the percentage of dynamic IP addresses doesn’t increase when the overall change throughout working hours is, for most days, positive. Remember that the number of total addresses has slowly increased throughout the measured days. Percentages of IP addresses being in a dynamic state of the total available has fluctuated between 9,69% and 14,18%, showing that the organization is a comfortable position considering network capacity. The following table also covers statistics from January 4th till April 30th, this time covering the change in absolute numbers for dynamic & static IP addresses.

Table 5 Average Change in # of Dynamic + Static IP Addresses

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Avg.**
5:00 - 9:00	10491	10500	7197	8708	5489	8477
9:00 - 13:00	1375	1507	1081	891	283	1027
13:00 - 17:00	-3991	-4137	-3203	-4312	-3350	-3799
17:00 - 21:00	-3752	-4483	-3445	-4282	-2784	-3749

Other data that could be analysed covering similar topic is the *DHCP Message Rate Trend*, which is measured by five grid members in total, of which three are located at one of the sites in Brussels and two at the site in London. Data of the DHCP message rate trend has only been analysed for the number of DHCP request message types. The reason for this is because it filters out those clients that initially sent a DHCP discover message to the server but were not offered an IP address. A request type message on the other hand can only be done after the client has received a DHCP offer message from the server. For the period between January 4th and April 30th, for each day the number of DHCP request messages was recorded between 7 AM and 5 PM with time aggregation of ten minutes, creating 61 data points per monitored day. Again, only working days were considered. The average trend over four months has been computed and is presented on the following figure, where the data points show the number of DHCP request messages over the last ten minutes.

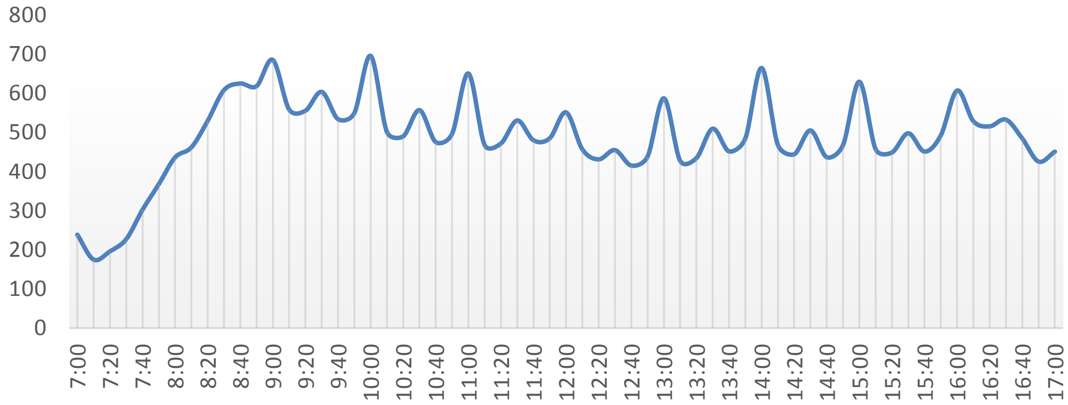


Figure 9 Average Trend in DHCP Request Message Rate Between January 4 - April 30

Again, the first noticeable trend is that the number of DHCP request messages increases in the early morning between 7 and 9 AM for when employees enter the office and start up their designated workstations, which initially triggers the DHCP discover message to be sent by these workstations before receiving a message response from the server. The message rate then fluctuates throughout the day with noticeable peaks at the exact hours of 10 AM, 11 AM, 1 PM, 2 PM, 3 PM and 4 PM. To clarify, DHCP request messages could also be sent in the scenarios of workstations and other devices renewing or rebinding their existing DHCP lease for when they were on stand-by or rebooted. Therefore, a message rate fluctuating between 400 and 600 per ten minutes does not seem to be unlikely given the scale of the organization. Another thing to take into consideration is that the DHCP request message rate does not showcase the actual trend in IP address leasing activity, as the DHCP server still has to send back an acknowledgement message containing network configurations for the client. Another method to illustrate typical network usage trends and capacity of the organization is to investigate on the total number of DNS queries per second for all grid members included. DNS queries per second show how much traffic all servers have to manage. Again, for the period between January 4th and April 30th, the average daily number of DNS queries per second for each grid member was recorded. It should be noted that two grid members were operational since March 17. It has been fluctuating between 7,000 and 9,000 queries per second for the first three months. Since April 8th however, the number of DNS queries per second has surpassed 10,000 and showcases an increasing demand of network capacity.

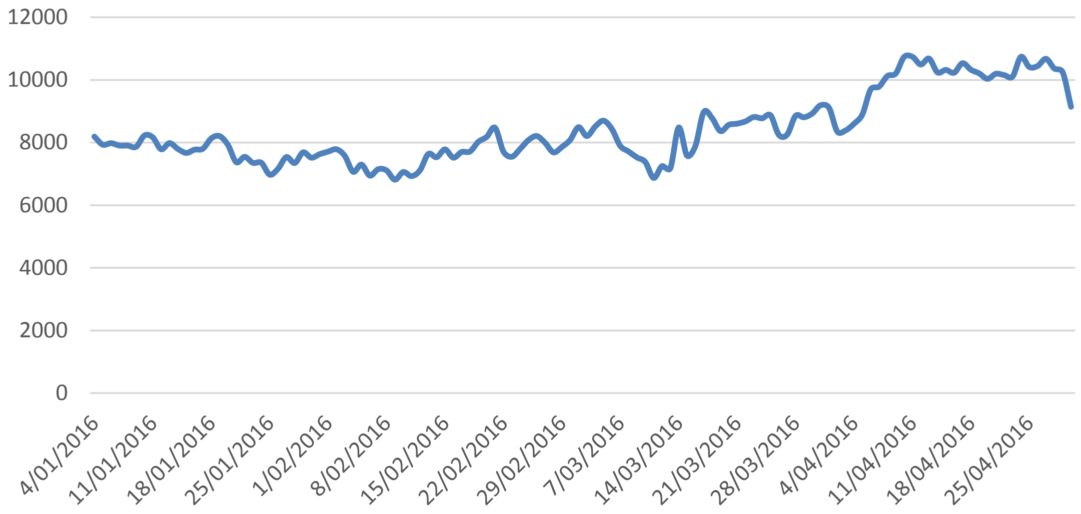


Figure 10 Average Total DNS Queries Per Second for the period January 4 - April 30

Finally, to conclude the picture of typical network behavior and performance at the organization, the traffic rate in bytes per second was recorded for the period between January 4th and April 30th for the top four grid members handling the most traffic, taking the daily average traffic rate per second including both inbound and outbound traffic. The total average number of bytes per second has then been converted to terabytes per day, with the results being shown on the following graph. Again, similar to the DNS queries per day graph, it can be seen that over the four monitored months, the average daily traffic rate increases slowly, starting from 0,25 terabytes per day to 0,36 terabytes per day. This could be due to the correlation between DNS query rate and network traffic as a larger query rate implies a larger rate of traffic to be handled by the servers. Referring back to the data analysis of the DHCP usage trend, it is worth noting that despite the possible continuing trend in network traffic rate and the increasing number of connected devices (as the number of total IP addresses in a DHCP lease grows while the percentage of static and dynamic IP addresses remains the same), the organization remains to have more than enough capacity to handle double the typical network usage trends in case of emergencies.

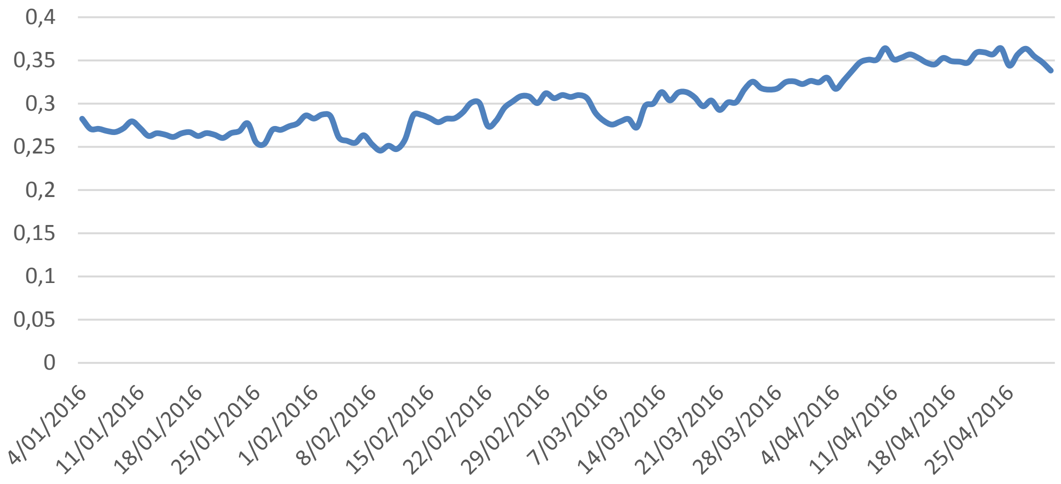


Figure 11 Average Traffic Rate in Terabytes Per Day For Top-4 Grid Members

These typical network usage and performance figures are not only discussed in order to create a better understanding of the size of the organization's network, but also to create a baseline for what could possibly happen if data quality problems were to be tackled. To put it simply, it is the question whether improvement in network data quality would influence the presented network usage and performance statistics to what effect and where. For instance, an IP address that is under conflict could still count as an IP address under either a dynamic or static state. While the organization in question typically holds on average around 86% of their IP addresses in a free state, other organizations with a lower percentage could decide to add more networks to increase the margin. If major data quality errors exist in the form of many IP addresses being in conflict as detected by a monitoring tool such as NetMRI without ever being resolved via a renewed scan or self-healing mechanisms, organizations could be presented with inaccurate data leading to unnecessary operating costs and investments. This could also mean that stakeholders within the organization need to have an understanding of the balance between the number of active IP address leases and traffic rate to determine whether typical network performance behaviour is being displayed. IP addresses whose original DHCP lease state are never updated when trapped in a conflict state could lead to data inaccuracy and problems concerning timeliness of data. Another point of concern could be that the number of devices is difficult to determine given the idea that within a state of conflict, multiple devices within a short timeframe were associated to the same IP address. Further investigation within data reports is then needed to determine whether the newly associated device that has caused the IP address conflict is still under active lease on that same IP address. Of course, unusual sudden spikes in DNS query rate or total traffic rate would leave some to suggest that there is a data quality error, but is more closely related to security monitoring and threat analysis, and such figures should not be ignored.

4.4 A First Inspection of the Presentation Layer

As has been discussed before, the DDI-monitoring tool of Infoblox (IPAM) is mainly concerned about three major network elements: DNS, DHCP, and IPAM. To clarify, DDI is often interchanged with IPAM since it encapsulates the DNS and DHCP processes. Read-only access to data management was granted for all three areas of the network monitoring tool, along with a dashboard overview of the status of the implemented network *grid* covering the abovementioned European countries. The user interface also includes a reporting tab where many types of results covering DDI-elements could be exported to CSV-files. Optionally, the user interface also makes it possible to drill-down through measurement data and apply filters, such as given start and end dates, to these reports.

The data management section has a tab dedicated to DNS. To be reminded, the basic functionality of DNS is to pair IP addresses to domain addresses in order to establish that the connection that will be made is legitimate. To take a global example of the organization in question, the main page for the intranet is divided into multiple sub-domains. The user interface provides a list of the names of these subdomains, each containing data of IP address mapping for host names and the related ports. Regarding security and trend analysis, investigating which ports are receiving the most stress and which domains are most popular within the organization might provide information about (non)-typical trends in the organization's network behaviour. While it has been explained that grid members are spread in multiple countries, the organization has chosen to merge both the Belgian and Dutch DNS domains into one DNS-zone.

The IPAM tab of the data management would showcase for each network in the organization the percentage of IP addresses used in that particular network. The monitoring tool allows for a simple overview of how the used IP addresses are distributed, and which IP address blocks are, at the moment, unused. Referring back to the hierarchical network allocation as shown by Rooney (2013), the interface allows for an overview of multiple IT blocks, with the smallest blocks containing the IP address range of x.x.0.0 to x.x.255.255, allowing for 65.536 different IP addresses, and allows for one single network to be created whenever a subnet mask 16-bit length is used. In simple terms, the network configurations deployed at the organization allows for a network to be created with room for only 65.536 IP addresses. If a longer bit-length is used for the subnet mask, for instance 24, there are only 256 IP addresses available per network. This means though that the IP-address block of 65.536 IP address can be broken down into 256 (65.536/256) networks, each holding 256 IP addresses, which is also a technique used at the organization. The user interface allows for users to zoom in on IP address blocks as such that it is possible to start from a given wide IP address range, for instance 0.0.0.0 to 0.255.255.255 (16,777,216 IP addresses!), and narrow down step-wise to an IP address block of 65,536 addresses and then further down to an IP-block of just 256 IP addresses using the different subnet masks, for instance between the IP address range of 0.0.0.0 to 0.0.0.255.

The following is an example of such IP address block of 256 IP addresses, where some of the IP addresses are currently under active lease due to them being assigned to workstations that are actively used by employees within the organization, in this example eight. Three IP addresses are labelled '*fixed address/reservation*' and are dedicated to printers, scanners, and security devices, which should not be moved frequently between networks. Twenty-one IP-addresses are labelled with '*conflict*', but the word conflict may be a bit confusing in this context. The user interface will show that the date for which the monitoring tool has last discovered a device associated for that IP address is from several months or even years ago. Some of these IP addresses could have been reserved via a DHCP-server for a workstation, such as a laptop, that is no longer in use. To be specific, the conflict can be related to either the DCHP range or the MAC address, as will be explained in section 4.5. The conflict can only be established when a discovery processes has been performed beforehand. The last IP address of a block is named the *broadcast address*. It is used for when an IP-packet has to be sent to all hosts within the subnet, meaning within the same IP address block. An example of a message to be sent to all hosts could be an announcement of configuration changes to be applied in the near future.

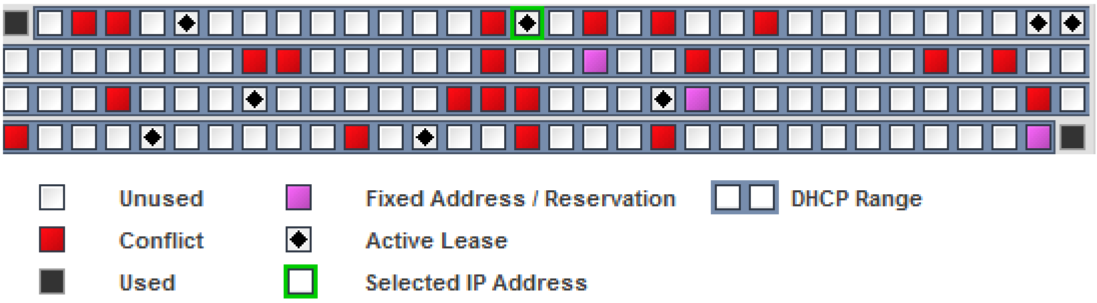


Figure 12 Example of an IP address block via Infoblox IPAM user interface

Another detail to mention in this particular IP address block is that all except the two used IP addresses fall within the range of DHCP. This means that DHCP-servers can give a lease to these IP addresses for whenever a new device is placed within the organization's environment. Of course, the ones that are currently in active lease, have been reserved, or are currently conflicting with availability, fall outside the available DHCP-range, but were originally included in the range. Introducing a new workstation would mean that the DHCP-server has the opportunity to assign an IP-address within this block by choosing any of the unused ones. This can be achieved by applying so-called DHCP range templates where the number of available IP addresses can be predetermined. This can be done by applying an *offset* value. For instance, an offset value of 10 would mean that the first 10 IP addresses in a block are excluded from the DHCP range. For the exemplary IP address block, the offset could be 1 due to the first IP address possibly being reserved for a router. Another kind of template is a fixed address or reservation template, where a given number of IP addresses are to be assigned to, for instance, printers. Both templates can be bundled into an overall network template for easier configuration over multiple networks. The organization in question uses twenty-six DHCP range templates for its networks in Belgium, The Netherlands, and the United Kingdom, covering DHCP ranges for regular users, voice-over IP, and (guest) Wi-Fi users.

Using the function of implementing such network templates could possibly be related to the data quality dimensions of consistency and accuracy to some extent. If for a given network container, each leaf network would be assigned with two printers on IP-addresses x.x.x.11 and x.x.x.12, the data values associated to these IP-addresses should refer to the real-world object of a printer and nothing else to comply with data consistency and integrity. As for relative data accuracy, determining whether or not the values in the data attributes are truthful can be supported via the applied template. If it is assured that the template is deployed over a given network, while data attributes would suggest that for IP-addresses x.x.x.11 and x.x.x.12 a different (un)known device is on active lease, network administrators could easily deduce data accuracy errors. Simplifying the scenario, if a data value ‘110’ would be given to identify the device type as a printer, the template should allow for IP-addresses x.x.x.11 and x.x.x.12 to always be associated with the data attribute of device type with value ‘110’; however, as relative accuracy is being discussed, it is not suggested to exclude all other data values. Data values ‘101’ through ‘109’ could be other plausible device types that are recognized within the organization, while ‘999’ would simply not refer to a recognized real-world object and is therefore not truthful.

An example applied at the organization during the research period was the assessment of free IP addresses within the DHCP range for Wi-Fi-connections across multiple floors. Three IP address blocks, each holding 1024 addresses, were available for all eleven floors at the site in Brussels. Using the presentation layer, it was easy to compute the number of IP addresses being in either active lease, conflict, being reserved, or free. It was quickly concluded from these numbers that the DHCP lease scope for one of these IP address blocks was almost fully utilized. Some of the changes applied to achieve a more flexible use of address space were to cut the original DHCP lease duration of eight hours to four hours and to allow for devices to receive an IP address regardless of current office floor.

It has thus been mentioned how with IPAM it is possible to retrieve the association between an IP address and the device currently leasing the IP address, though it turns out that even inspecting unused IP addresses return data values about the name of a device and its accompanied MAC address and OS fingerprint. However, when discussing data quality, it is important to know what data values can be extracted via the monitoring tool, and whether the data within a given scope, such as the IP address block of 256 addresses, is consistent, complete, and timely, just to name a few data dimensions. Taking the example again from an IP address being in ‘active lease’ by a workstation, a few attributes can be discovered. It has been mentioned already that the monitoring tool is capable of showcasing when a device has been last discovered leasing a particular IP address. This may comply with the timeliness data dimension, meaning that there is a verification of when the data about the device were fetched.

Other data attributes that could be fetched, besides the associated IP address, are the MAC-address of the device and which operating system it is using. This complies with one of the discovery protocols described by Reddy et al. (2014), it being the TCP discovery type by monitoring TCP/IP packets. IP and MAC addresses can also be fetched via ICMP echo requests, or ‘pings’. While this provides useful information about the device itself, it may still be unclear how the device is labelled. This is related to the importance of *naming convention*: if devices or servers are not named in the same style, not only does it deteriorate consistency in labelling, but it may be unknown where these devices are located. For the organization however, multiple leaf networks are labelled with floor numbers for their offices to keep track of IP address availability for devices on each office floor. This does not mean however that it can be derived whether a workstation assigned to a particular IP address labelled with *Floor Three* is also present at the third floor, to give an example.

Although the monitoring tool cannot improve data quality about the attributes of detected devices itself, it could give an indication of where weaknesses in data quality lie. Does data completeness underperform within a given network range? Is the formatting for all data attributes consistent within a given network range? Is the fetched data pulled from IP-packets considered to be outdated? These questions are related to the concern of data quality improvement and are raised upon using network monitoring tools and being exposed with the presented data. A quick inspection on just a few 256-IP address blocks shows that not all three data attributes can be fetched related to an IP-address which is under active lease. The naming convention could change without exactly understanding whether the labelling was intentional or simply random.

The question now is where stakeholders have identified the ‘*clutter*’, meaning the improper data quality caused by either improper naming convention made responsible by human input, or the incompleteness of data attribute values due to the lacking capability of the monitoring tool. Looking through every IP-address block available would be a monotonous, time-intensive, and simply illogical task, due to the status of workstations constantly changing and devices being introduced, moved, or even removed every day. This is why Infoblox IPAM is also equipped with a reporting tab for summarized results for DNS, DHCP, and IPAM which may be refreshed daily or hourly. However, not all stakeholders are equally knowledgeable about network technology and it would be a considerate challenge to explain an action plan of improving data quality without avoiding mentioning technical concepts such as DDI. It would therefore be better to first understand where the stakeholders are at related to their thought to network data quality and where they believe the current problem is today. Only through an assessment of describing in what current situation stakeholders think the organization is at concerning the problem can an action plan be developed while keeping in mind the concerns of different groups of stakeholders.

4.5 First Identified Data Quality Issues

Several data quality issues that can be identified simply by examining the list of IP addresses and its available data from the example IP address block are now going to be explained. These data quality issues are mainly related to the data incompleteness dimension, and this section discusses not only minor data incompleteness issues, but also instances where discovered data of NetMRI as a network scanner is unavailable for all data attributes. Also, the organization applies a method where data attributes such as country and site can be filled in automatically based on the associated leaf network.

4.5.1 Data Incompleteness Issues

While it has been explained that IP addresses labelled as '*conflict*' are related to devices that have not been detected for several months or years due to mismatched MAC addresses or DCHP leases, other IP addresses that are considered '*unused*' may contain device information as well. This seems odd as the DHCP server recognizes the lease state of the device as '*free*', meaning that a new device could be assigned to the IP address in the future. Even more inconvenient is the lack of a timestamp for these unused IP addresses, meaning that there is no information available about any activity of said device. Chances are that data could be overwritten when a new device would use that given IP address, accompanied with a timestamp for verification. To make things even more complicated, for some IP addresses the OS fingerprint cannot create a match and no name for the device is stated, while a MAC address has been found and a timestamp is included. Interestingly, for IP addresses which are supposed to cause a conflict for DHCP leasing, the device name and MAC address can be fetched, accompanied with a timestamp for last discovery, but the fingerprint for the operating system is missing. This is odd as for IP addresses assigned to a device, either in active lease or unused, seem to have OS fingerprints in most cases. This indicates that data completeness is underperforming in multiple scenarios, meaning that, for this IP address block at least, there is no IP address data available that provides the name, MAC address, OS fingerprint, and timestamp all together. A quick search on another IP address block did showcase a few data records where all four data attributes were present. Nevertheless, a desired goal would be to have all data attributes available for all IP address either currently on active lease or are currently reserved for a device.

A possible reason for a device to not include any data values about its OS fingerprint could be that the DHCP client identifier has not been detected. To give a simplified explanation, the client identifier acts as a value for which DHCP servers can then determine which value is associated with which client. Relating the DHCP client identifier with a data quality dimension, two mechanisms are in place to prevent data deduplication. Multiple IP addresses cannot share the same MAC address nor the DHCP client identifier in the same network. A DHCP server identifying an IP address holding an identical DHCP client identifier will reply by assigning the same IP address on condition that the original lease is still active. However, if such an identifier is lacking, one can be generated based on the MAC address and the type of hardware installed on the device. Since the hardware has to run on an operating system, the DHCP client identifier then also contains information about the operating system fingerprint derived from the hardware specifications of the device. It would therefore seem logical to find records where the DHCP client identifier for a workstation is missing and no match is found for the operating system; however, the MAC address may still be present. However, the DHCP footprint of an OS is merely an estimation. It could be that an exact assessment of the hardware of the workstation cannot be made. Another legitimate explanation could be that it is not related to a workstation but to a printer, for which the operating system is different and is a fixed, installed device not specifically associated to a user. Strangely enough, when the DHCP client identifier and the MAC address are present associated with an IP address, it could be that the result for the identified operating system delivers '*no match*'. This usually is the case when no device name is assigned to the given IP address, but it could occur with the device name present. This means that there is no guarantee that the OS fingerprint can be fetched even with the presence of the device name, MAC address and DHCP client identifier, and leads to a significant problem regarding data completeness.

A list of available standard fingerprints can be found within Infoblox DHCP data management, ranging from operating systems for workstations to projectors and game consoles. The main philosophy is that subtle differences in DHCP packets can determine on which OS a client is running on. Any achieved connectivity of these devices within the internal organizational network environment should therefore be recognized through the user interface under the fingerprints data attribute for IP addresses. Of course, under this list of fingerprints are the OSs of the workstations for which the issues related to data completeness have been discussed. Therefore, it is odd that a '*no match*' result in fingerprinting is achieved with most of the device attributes known. While the organization does not apply custom fingerprints, a remaining question could be whether OSs cannot be determined due to the lack of custom fingerprints, due to the lack of subtle differences in the DCHP packets, or due to another reason related to data collection issues.

For those IP addresses which are under active lease and are recognized to be paired with a known device, one data attribute describes the ‘discoverer’ of the data, which is always labelled with the name ‘NetMRI’. Extra data attributes that can be pulled from NetMRI are network port details of the device, the device model, and its type. In the case of a virtual network environment, details of the virtual host and its paired virtual cluster could be fetched if present. The integration between the DDI tool and NetMRI is realized via a *switch port manager* (SPM) product allowing for network discovery, topology, and granting users a historical view of data attributes such as IP and MAC addresses. Designed to create a simplistic overview of network capacity by determining the percentage of free ports, it can be synced with NetMRI to give a detailed overview of the scanned devices. To add to that, the extra available data from SPM can be integrated to the DDI tool, promoting data completeness about an entity, in this case a device connected to a network.

For IP addresses under active lease, whether by a workstation, printer, or network switch, just to name a few device types, critical information regarding network device discovery is included and obtainable via the DDI tool. It concerns the port number and port name of the attached device, allowing for easy tracing of these devices. Unfortunately, additional data attributes that would seem to come in handy to identify the attached devices in more detail are left blank in the case of workstations and printers, such as the device name and type. However, the presence of port information while the device name or type is lacking means that network administrators can still trace the device and its activity. A possibility could be that the device type can be deducted from the OS type successfully identified. An existing, worse scenario would be inspecting IP addresses associated with unmanaged devices showing that port information is lacking. Another observation is that in some occurrences of inspecting IP addresses under active lease, none of the data attributes are filled in. It is possible however to identify the name of the related object by inspecting DHCP lease details or address records, and there have been instances where the device name would suggest it is a familiar workstation unit within the organization, matching current naming convention.

4.5.2 Inconsistent Availability of Discovered Data via NetMRI

There are, however, many inconsistencies in the data when taking into account the factors whether or not discovered data via NetMRI is present of an IP address under current active lease and whether the timestamp of the first and last data discovery make sense. To give a simple example, on April 6th, the data of the IP address of the workstation that was provided by the organization for the purposes of this research stated that the device was first discovered on the same date, almost forty-five minutes after having logged in on that particular date. As a researcher, the question could be asked whether historical analysis of either the IP address or of the device in question before the stated first discovery date has been made impossible. This is because the IP address of the workstation had been recognized as being under active lease before April 6th and was paired with the correct workstation. If it is true that historical analysis before April 6th has been made impossible, then another question would be how long historical data is kept for such analysis. However, historical data of the workstation is available via DHCP lease history data reports, giving doubt to the inclusion of the first discovered timestamp generated by NetMRI.

It seems that the data related to the timestamps for both first discovery date and last discovery date can also raise some doubt for data quality. For instance, two IP addresses residing in the same IP block were under active lease. The first IP address had a recent *last discovered* timestamp (within six hours) but lacked any other discovered data. The second IP address has a *last discovered* timestamp just two hours earlier of that of the first, but discovered data from NetMRI was available. On the contrary, one IP address was discovered which had an active lease status for only four minutes, and discovered data from NetMRI was lacking. Therefore, there is no guarantee of knowing whether or not an IP address under active lease is paired with data fetched from NetMRI by only examining the timestamps of the DHCP lease and last discovery by NetMRI.

One remarkable finding was an IP address under active lease for which NetMRI provided the *last discovered* timestamp set on March 31st. Other IP addresses under active lease in the same IP address block have been discovered after March 31st, and all hold the exact same timestamp of April 6th, 10:18 AM. Interestingly enough, the IP address in question gained its DHCP lease on April 6th, 10:24 AM, six minutes after the NetMRI discovery process identified the other devices holding an IP address on the same IP address block. Whether this explains the fact that there is no discovered data available from NetMRI still has to be determined. Other IP addresses which were discovered under the conditions that they were displayed to be under active lease and were not paired with discovered data from NetMRI would hold a DHCP start lease date from over one to nearly two years ago. On one discovered instance, the DHCP lease date was set to ‘*never*’.

4.5.3 Extensible Attributes

The organization has also introduced extensible attributes for its administration purposes besides relying on the data collected via discovery processes using packet tracing. Examples of these attributes could be the name of a country, building, or site, even homing down to including the rack where a relevant device is held in a data room. Currently, the organization has only allowed for two extensible attributes to be inheritance enabled: '*country*' and '*zone*'. An attribute being inheritance enabled means that a network administrator needs to spend less time to assign the correct attribute to different subnets. Take the example of '*Belgium*' being the country where the network container 0.0.0.0/8, meaning the IP address range of 0.0.0.0 to 0.255.255.255 and all its underlying network containers and leaf networks, needs to be associated with. However, there is one network container, let us say 0.128.0.0/16, where '*Netherlands*' is the associated country. The challenge therefore is to enable that only the subnets under 0.128.0.0/16 are excluded, despite it falling within the range for where the country attribute of Belgium is going to be applied. Instead, the country attribute of Netherlands has to be applied for all subnets falling under this particular network container.

While their functionality seems simple and quite humble, it is a step forward regarding data quality in complex network topologies. Referring back to the simple example, one may suggest to move the network container allocated to The Netherlands over to a more appropriate large network container; however, this would introduce a grave risk of migrating IP addresses over and reconfiguring DHCP settings for all devices included in the network container, costing precious time and money. Being empowered to assign attributes to a subnet containing a wide array of IP addresses in possible follow-up subnets allows for easier management of standard relevant attributes throughout the organization's network.

4.5.4 Overview of First Identified Data Quality Issues

The following list is intended to give an overview of the first identified data quality issues and an attempt is made to pair these data quality issues with the KPIs established by the organization during implementation of the network management system. An estimation is also made to what level of severity the data quality issues can be categorized under, as some may oppose a threat to business continuity. Suggestions for resolving each data quality issue are also presented in the list. Note that one of the data quality issues included in the table is not fully described in the thesis because of its low level of severity on any of the organization's KPIs. Nevertheless, such minority data quality issues should still be dealt with in order to keep a complete, accurate, and consistent view of the available data.

- [Data incompleteness] – Inspection of IP addresses under active lease face multiple data incompleteness issues such as missing MAC address, OS fingerprint timestamp, and/or name of related device, sometimes even all data attributes. Port numbers and port names are, in few occasions, also unavailable.

  - KPIs possibly affected: % error rate, business continuity, criticality of survival. Not being able to identify a rogue device when data attributes are missing of an IP address lead to security risks in continuity.
  - Level of severity: significant to critical
  - Possible solution: Make further attempts to improve the frequency of including the DHCP client identifier, or combine different network discovery protocols as suggested by Reddy et al. (2014).
- [Data incompleteness] – Even when most of the device attributes are known, such as MAC address, OS fingerprinting can still result in '*no match*', while it should fall under a fingerprint template.
  - KPIs possibly affected: % error rate
  - Level of severity: Normal/Minor
  - Possible solution: Update the fingerprint template list of OS to identify more cases, otherwise check the performance of the currently used network discovery protocol for improved packet tracing.
- [Data consistency] – Presentation of naming convention of workstations; while each workstation has a consistent naming convention to determine whether it belongs in production, development, or acceptance, often duplicate records of the same related object will be present, such as the lease type and A/PTR-records together.
  - KPIs possibly affected: not necessarily related to % error rate as the data is available and assumed to be correct, only not consistent in its format.
  - Level of severity: low
  - Possible solution: For the discovered data within IPAM, force it that the name of the device is identical to the host name as derived from the DHCP lease actions.
- [Data timeliness] – Timestamps of the data attributes *'first discovered*' and *'last discovered*' for IP addresses under conflict are identical, offering almost no meaningful information of DHCP lease history.
  - KPIs possibly affected: % error rate, business continuity. Again, details of when conflicts or rogue activity arises and end are needed for the assessment of the level of security in place.
  - Level of severity: normal
  - Possible solution: Borrow the most recent timestamp from DHCP lease history paired with the device having originally caused the IP address conflict.
- [Data timeliness] – Discovered data of IP addresses under active lease would have a *first discovered* and *last discovered* timestamp older than three days while

DHCP lease actions after the last discovered timestamp have occurred, causing data timeliness issues.

  - KPIs possibly affected: % error rate, since the percentage of present time cannot be computed with outdated data.
  - Level of severity: significant
  - Possible solution: For data timeliness purposes, it is important to constantly update the *last discovered* timestamp based on DHCP lease history data records or by hourly scans of NetMRI.

- [Data timeliness] – The *first discovered* timestamp of an IP address, regardless of its status, can be quite recent (present day or the day before) while further history of the IP address is available via DHCP lease history, giving doubt to the inclusion of such *first discovered* timestamp.
  - KPIs possibly affected: % error rate; it is understandable to not keep data history from the very beginning due to storage limitation reasons, but data analysts have little to no use of this data attribute if it is irrelevant to the start of a new original DHCP lease or another unique activity.
  - Level of severity: normal/low
  - Possible solution: A suggestion would be to use the timestamp of when the current (or last) lease originally started based on DHCP lease history data records. This allows for data analysts to see for those IP addresses having a rich DHCP lease history of multiple devices the start data of the latest original lease. This also gives more information on how long a lease was held until a possible conflict arrived, when combined with the solution of the previous data quality issue.

4.6 Data Conflict Resolution

Updating the data is done by performing network discovery practices as described in section 2.6. Newly discovered data will either overwrite unmanaged data or will be merged with existing data within the database, and is provided via the NetMRI tool. To clarify, unmanaged data is present pre-discovery and is not properly configured for either DNS or DHCP. For instance, an unmanaged IP address is not associated with a lease, host or with any address record. Resolving such problem could be done by either releasing the unmanaged status, allowing the IP address to be presented as '*unused*', or by pairing it with a host or address. In post-discovery though, new conflicting addresses may be presented which were globally described in the previous section. The reasoning for focusing on these IP address conflicts occurring within the organization is that it could be seen as a proxy for data quality within the scope of IP address blocks. A conflict indicates that the data from the IP address is not similar to the discovered data of that same IP address, which could be easily related to the topic of data quality. The problem with it being eligible as a proxy is that data incompleteness for several attributes may still occur without the IP address being labeled as having a conflict, as has been described earlier. Still, data incompleteness does not directly suggest that data values are conflicting and thus the overall quality of available data attributes may still be accurate. Nevertheless, investigating IP address data conflicts may still give an indication of how well network discovery data is managed and which type of conflict occurs the most.

According to the administrator guide of Infoblox (2012), there are four types of IP address conflicts that could occur. The first type is a DHCP lease conflict, where the MAC-address of an IP address in active lease via DHCP within the discovered data does not match the previously related MAC-address, causing the IP address to have this type of conflict. Three possible resolutions are available, where first the decision has to be made whether to hold the DHCP lease or to release it. By holding it, the conflict is ignored and the DHCP lease for the IP address is kept. By releasing it, it would be suggested from the discovered data to create either a fixed address or a reservation for the IP address. The second type of conflict is not a lease conflict per se, but rather a fixed address conflict. With a fixed address, the DHCP server has to identify the MAC address of the device in order to assign the same IP address for every request initiated. Mismatching MAC addresses between existing and newly discovered data causes a conflict, where the network administrator has the option of either clearing the discovered data and maintain the fixed address, or to update the fixed address with the new MAC address.

DHCP range conflict is a third type of conflict. Rather similar to the first type, in this scenario an IP address resides within a given DHCP range but is not associated with an existing DHCP lease or address. The discovered data however presents that the IP address is in an active state, where the network administrator is forced to make the decision to either neglect the discovered data or to create a fixed address or reservation concerning the IP address. Finally, the fourth type of conflict is quite similar to the second type, except that the IP address is considered to be belonging to a host record. The same conflict still applies concerning the MAC addresses of both the existing data and the discovered data, and the options are to either neglect the discovered data or to update the host record with the new MAC address.

Critical to understand from this explanation is that newly discovered data should not immediately be considered as new, ‘true’ data. Instead, in the scenario of a conflict network administrators are still empowered with the choice to either accept or neglect the discovered data. This leaves room for discussion whether or not the discovered data presents ‘improved’ data for all occasions. Of course, the network administrator has to assess the accuracy of the discovered data in order to select its choice of resolution. This is why the user interface does not provide an option for ‘*bulk resolving*’, meaning the resolution of multiple conflicts simultaneously using the same resolution method, as each conflict should be respected on its own to promote data quality importance. In order to gain more understanding of their assessment, it is suggested to inspect on which type of conflict occurs the most within a given scope of IP address blocks and which type of resolution has been performed on earlier identified conflicts.

The following is an example of IP address conflict resolution activity. Near the end of March, the IP address conflicts in the exemplary IP address block as shown in the previous section were resolved. Likely to have been DHCP lease-type conflicts, the lease state of these IP addresses have been cleared and their current state is set to *'free'*. Interestingly enough, one of the IP addresses that used to be in active state had now been identified as a DHCP lease conflict between the period of March 31st and April 4th. Afterwards, it regained its status of being under an active DHCP lease, possibly due to the prior absence of the employee linked to the earlier detected device associated with that particular IP address, or due to a ping sweep.

During the timeline of the case study research, the organization in question has raised concern about how alerts are not generated from duplicate IP address events, caused by one of the four types of conflicts explained earlier. Ideally, these alerts are to be presented by a separate event management system, OMNIbus by IBM, and seen by members of the MCR. The organization has also excluded the possibility to forward such events via Splunk, a big-data analysis tool which is also currently used. Interestingly enough, recent versions of Infoblox Grid Manager an engine based on Splunk for generated reports and analysis. This adds the challenge however to integrate a solution for NetMRI to forward its discovered events, probably over a system log, over to a separate event management system who can then read and interpret the data and produce the event alerts for MCR to react upon. This shows that even though the network monitoring tool was given time to mature within the organization, not all data-related issues are forwarded to those responsible to allocate and/or resolve them, leaving a risk that some of them may be unnoticed.

4.7 Data Quality Problems in Generated Reports

The following section shall now discuss possible data quality issues found when inspecting the generated data reports available via the presentation layer of the Infoblox DDI solution base off data from NetMRI. These are mainly focused on the data accuracy and incompleteness dimensions. Six of these issues have been identified and will be reviewed at the end of the section in similar fashion as shown in section 4.4.4. Due to the various limitations of the thesis, as will be acknowledged in the final chapter, only three data reports have been carefully chosen for further inspection. Examples of data records available in these reports have been altered in order to confirm with confidentiality policies.

4.7.1 IPAM Network Usage Statistics

As mentioned earlier, the Infoblox DDI tool allows for the user to retrieve generated reports on multiple network topics. The first one being discussed is *IPAM Network Usage Statistics* and contains data about the DHCP range utilization on the last scanned networks presented on a table. By default, the timeframe of these last scanned networks is set to one hour. Each row on the data table contains attributes of the timestamp, network zone, its IP address and bit-length, the percentage of DHCP covering the IP address range on the network, the number of IP addresses allocated, reserved, assigned or unmanaged, and the percentage of utilization of IP addresses, not necessarily being within the DHCP range. This data gets uploaded hourly for all networks included in the IP address management module.

Interesting to note from the available data is that it only concerns leaf networks, with the largest registered leaf networks holding 65536 IP addresses (or an accompanied bit-length of 16). This means that these IP addresses are not divided into further subnets, creating this large IP address block with the first and last IP address typically being reserved. This means that data redundancy has been eliminated giving the idea that networks containing multiple leaf networks are not considered to be included in the generated data reports. One disadvantage of this is that a more basic overview of a large network container cannot be fetched directly from the data, but has to be computed by analysing the data of all related leaf networks. On the other hand, this eliminates the duplication of data errors if one of the data records describing a leaf network would contain dirty data. This issue will be discussed shortly using an observed, yet simple data error.

Upon first glance of the first ten thousand records covering leaf networks ranging from size between one and 8196 IP addresses, all falling under the exact same timestamp, it shows that the organization has put effort into eliminating instances of unmanaged IP addresses by introducing DHCP ranges that would cover all possible IP addresses within a given network for devices. However, going through historical data it does not take considerable effort to retrieve data records of leaf networks containing unmanaged devices. A remarkable finding is that a 256-IP address block held the record for having the most detected unmanaged IP addresses, meaning that devices have been paired with an IP address falling outside the designated DHCP lease within said IP address block, during the period January 26th till February 7th with an average of 104 unmanaged IP addresses on each recorded day, or over 40% of the total IP address block.

While such data observations may raise concern to some, the main purpose of the data is to show the allocation of IP addresses across the leaf networks, and to determine whether the number of IP addresses for lease assignment needs to be increased. However, it is not directly clear how the percentage of IP address utilization within a given network is computed by simply looking at the presented table. It has to be known beforehand that the formula applied to determine the percentage of IP address utilization in a given network is the following:

*# of allocated addresses / (# of total addresses - # of reserved addresses) * 100%.*

The reserved IP addresses would refer to the first and last IP address in a block, which are suggested to be the IP address for the router and the broadcast IP address, as has been explained earlier. While not directly stated via the web interface, reserved IP addresses have to be deducted from the total number of IP addresses available for the possible DHCP range, as reserved addresses fall outside the range. While the formula is sound for data covering networks with IP address blocks ranging from 4 to 8192 total included IP addresses, the formula does not apply for networks with IP address blocks including either one or two IP addresses. This is because the number of reserved addresses has been unchanged and is set to ‘2’. Of course, there would be a data conflict for those data rows where the size of the network is only one IP address, and applying the formula would suggest that IP address utilization percentage would either be 0% (in the case of the single IP address not being allocated) or -100% (in the case of the single IP address being allocated, and divided by -1 (1 minus 2)). Thankfully, the value of IP address utilization is unaffected despite this data conflict, and for networks holding one IP address the only registered values are 0% and 100%, which complies with mathematical logic. Similar to networks holding two IP addresses, applying the formula would mean that regardless of the number of allocated IP addresses, the number of total IP addresses minus the number of reserved IP addresses would always be zero, making it unable to compute a percentage of IP address utilization. Again, the actual percentages are displayed correctly in the data table (0%, 50%, or 100%).

Closer inspection on such one or two IP-address block networks via the IPAM mapping functionality shows that there are no reserved IP addresses included, therefore making the data value concerning the number of reserved IP addresses incorrect for such networks. The impression is therefore that the value ‘2’ has been manually included as a constant rather than it being detected by NetMRI. Such dirty data could be avoided by applying a second rule to introduce the value ‘0’ for those networks holding either one or two IP addresses; however, those responsible for applying such constant values should be absolutely sure that no deviations in the number of reserved IP addresses exist in all other networks, regardless of size. Given the network size of the organization, this seems to be a risky assumption. If possible, a preferable alternative would be to allow for NetMRI to check the number of reserved addresses in all instances.

Another data quality issue related to the single IP-block leaf networks is the reporting of the top utilized networks based from the hourly data scans of IPAM network statistics. Requesting a list of top utilized networks since the beginning of the hourly scans on January 25th will result in ten randomly selected single IP-block address networks. Included in the generated data table is the utilization rate given in percentage, the total number of

IP addresses, and the number of IP addresses assigned, reserved, and unmanaged. What is missing from the previous generated data report (IPAM Network Usage Statistics) is the number of IP addresses allocated. Referring back to the formula to compute the percentage of network utilization, the number of allocated addresses is needed to compute the number of currently used IP addresses before it gets divided by the number of total IP addresses minus the standard two reserved. While it has been explained that the two reserved IP addresses do not apply for leaf networks holding either one or two IP addresses, the utilization percentage is still computed correctly.

Not surprisingly, it seems that the top ten utilized networks are single-IP address leaf networks with an allocated IP address, therefore scoring 100% on network utilization. In fact, there are sixteen registered single-IP address block leaf networks. The following twenty-six leaf networks in the list hold two allocated IP addresses, also scoring 100% on network utilization. If we expand the top *N* list to 500, place #43 to #500 are all taken by four-IP block leaf networks holding two reserved IP addresses and two allocated IP addresses. The main point emphasized is that the ranking of top utilized networks by utilization percentage becomes a poor variable given the existence of such small leaf networks congesting the top *N* list. A better variable to rank top utilized networks would be the number of allocated IP addresses as it allows for larger networks to showcase their actual quantitative utilization rather than the utilization in proportion to their size. To give an example, a leaf network holding 256 IP addresses of which a hundred are allocated imposes a larger utilization rate than a single IP-address leaf network with an allocated IP address. This means that the data attribute stating the number of allocated IP addresses needs to be included in the data table.

4.7.2 DHCP Lease History

The second important generated data report to be discussed in this thesis is the *DHCP Lease History*. The data table includes a timestamp of when DHCP lease activity occurred, the name and IP address of the grid member associated with the event, the type of action concerning the DHCP lease, the lease IP address for which the action is related to, the MAC address, OS fingerprint, and host name of the device in question, and finally the timestamps of when the lease starts and when it is terminated. DHCP lease history is updated in real-time, and thus multiple data records can be added every second. The combination of such data attributes allows the user to detect in which grid member which DHCP lease action was performed to which device at what time in the case of data completeness and accuracy.

First observations of data records show that the majority of DHCP lease activities take place in multiple Belgian grid members, followed by grid members in the United Kingdom at the site in London. This should not come as a surprise since it has been explained earlier how the majority of the DHCP message rate trend takes place at these grid members. For DHCP leases going through the grid members in Belgium, the length of the lease may vary between five minutes to three days maximum, whereas in London some DHCP leases may be valid for a week. Other possibilities in Belgium are eight hours, two hours, one hour, or half an hour. An explanation for the different DHCP lease durations could be the policy of the organization to apply different DHCP lease times for different types of devices. For instance, while workstations or printers owned by the organization may hold a relatively long DHCP lease duration, devices that are likely to move around or are not privately owned by the organization (such as with BYOD) may hold a shorter DHCP lease time. However, the organization seems to apply different DHCP lease duration protocols for their different grid members. Analysing DHCP lease times over the three Belgian grid members installed at the workplace showed that for two of the three grid members, a DHCP lease of three days was assigned regardless of device type and whether an IP address was issued, renewed, or fixed. The third grid member held a flexible protocol, where different DHCP lease lengths were applied, rather independent on device type or type of action applied for the IP address. For instance, for the action to make IP addresses fixed to a printer, DHCP leases varied from being five minutes to three days long. Similar story can be said for IP addresses being renewed to printers.

However, not all DHCP end dates are set to maximum three days after their start. Sorting DHCP lease history data from the last twenty-four hours by descending DHCP lease end timestamp shows that there are data records available of abandoned IP addresses still holding an unusual DHCP lease end date. Between March 2014 and April 2016 148 records have been found where an IP address was issued to an unknown device holding an DHCP lease end date of January 19th, 2038, strongly related to a meaningful state data quality problem as coined by Rodriguez et al. (2009). To clarify the term ‘*abandoned*’ in this context, it could be that a DHCP server tries to issue a lease of a particular IP address to a client, but upon sending an echo to check whether the IP address is indeed free, it receives an echo reply (or ping response) from the IP address. It could also be that the client rejects the proposed IP address, therefore abandoning the proposed DHCP lease. Therefore, there is a possible chance that IP addresses with abandoned DHCP leases are still occupied by other devices. For instance, Princeton University (2014) once released a document where it was stated devices running on the Android operating system would keep using IP addresses even after the DHCP lease end date, causing network administrators to notice data records of abandoned IP addresses still being able to deliver a ping response.

A powerful feature of the DHCP Lease History generated reports is that one can apply a filter to only search for data records on a particular lease IP address to gain insight on the lease actions associated to that particular IP address. This is useful to inspect IP addresses where the IPAM IP address map shows an IP conflict in order to deduce the cause of the conflict, and possibly the best course of action to resolve it. The previous mentioned issue of the abandoned IP addresses can also be inspected further by filtering the search results on the associated IP address. An example is now given of such investigation on the behaviour of a now abandoned IP address before moving on with inspecting IP addresses under some type of conflict.

Table 6 displays a typical sequence of DHCP lease behaviour of a single IP address to a workstation but gets abruptly disturbed with the issue of the IP address being converted to the *abandoned* state. The typical sequence consists of the IP address being renewed for each time the workstation is booted up at the start of the working day. Since the lease was originally set for three days and the time that the workstation is offline due to it not being used after working hours is less than three days, allowing for the DHCP to renew the lease for another period of three days during boot-up, basically the three day period is shifted to the next day. It could be that the workstation is not used for longer than three days, at which point the lease IP address is freed. It could be however that on the following day after that, the same workstation gets (re-)issued with the same IP address due to the DHCP server retrieving the information of earlier leases. This process of renewing, freeing and issuing the lease can go on indefinitely in the best case. Searching for IP addresses holding the abandoned status however shows that the status changes to abandoned typically seconds after the last renewal of the lease for three days. Strangely enough, the abandoned lease IP address still holds a DHCP lease start and end date, the latter being set at the earlier mentioned date of January 19th, 2038.

A second observed scenario could be that the lease IP address switches between the issued and abandoned state multiple times within seconds before a new stable lease is given. To give one example, on April 12th, 2016 between 2:42:26 and 2:42:31 PM four devices attempted to hold the lease IP address until one other device became successful in holding and reissuing the lease. Investigating the IP lease history of the four devices which failed to issue the IP address in question shows that in three out of four cases, the sequence of lease actions seem to be illogical. These devices would already have their designated IP address from their original lease and would, similar to the previous table, often be renewed or freed until the next lease would be issued. However, the chronological sequence of lease actions of these three devices, as can be observed from the presented data on the Infoblox DDI tool, is represented in Table 7.

Table 6 Typical Sequence of DHCP Lease Behavior Ending With Abandonment

Time	**Action**	**Lease IP**	**MAC Address**	**Host Name**	**Lease Start**	**Lease End**
04/08/2016 07:36:25	Issued	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-08 07:36:25	2016-04-11 07:36:25
04/09/2016 07:35:38	Re-newed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-09 07:35:38	2016-04-12 07:35:38
04/10/2016 07:37:19	Re-newed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-10 07:37:19	2016-04-13 07:37:19
04/13/2016 07:37:19	Freed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-10 07:37:19	2016-04-13 07:37:19
04/15/2016 07:38:00	Issued	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-15 07:38:00	2016-04-18 07:38:00
04/16/2016 07:32:50	Re-newed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-16 07:32:50	2016-04-19 07:32:50
04/16/2016 07:32:54	Aban-doned	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-16 07:32:54	2038-01-19 04:19:00

Table 7 Data Quality Issue #1 - Issued to Renewed for Different IP Addresses

Time	**Action**	**Lease IP**	**MAC Address**	**Host Name**	**Lease Start**	**Lease End**
04/12/2016 14:30:31	Re-newed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123ABC*	2016-04-12 14:30:31	2016-04-15 14:30:31
04/12/2016 14:42:30	Issued	5.6.7.8	44:dd:55 :ee:66:ff	*Workstation 123ABC*	2016-04-12 14:42:30	2016-04-12 22:42:30
04/12/2016 14:42:31	Aban-doned	5.6.7.8	-	-	2016-04-12 14:42:31	2038-01-19 04:19:00
04/12/2016 17:23:37	Re-newed	1.2.3.4	11:aa:22: bb:33:cc	*Workstation 123 ABC*	2016-04-12 17:23:37	2016-04-15 17:23:37

To clarify, the row for which the cells are filled with a coloured shade is not originally present in the data report when filtering on the records for a particular device only. This means that the row is not part of the sequence presented in the tool. The inclusion of the row will be explained shortly. Notice that the length of the original lease lasts for three days while the new issued IP address would only hold a lease of eight hours. Examining these data records without knowing the IP address history of, in this case, 5.6.7.8 (the coloured row) would leave the data user puzzled as of why the device would issue a new IP address during lease renewal. Another issue is that the state of the new IP address after the device received its original IP address lease renewal is unknown unless the user filters the search to find lease history records of only that particular new IP address, hence the inclusion of the extra row in the table. Though, it is logical that the data record for when the DHCP lease gets abandoned regarding IP address 5.6.7.8 is lacking in the previous table, given the fact that the data attribute of the device name is not present. If it were present, it would explain the jump between issuing a new IP address lease and renewing the existing one after that, and it would have been included in the table when filtering on device name only.

This should be considered a data quality flaw related to data incompleteness since the inclusion of the device name for every data record in which the DHCP lease action is set to '*abandoned*' would eliminate this inconvenience. Unfortunately, this is not always the case, and the previous example is one where the host name is lacking on the data record where it is stated that the lease for the IP address 5.6.7.8 was abandoned. This is illustrated by the inclusion of an additional row. The cell for the hostname data attribute within this row has been filled in with a darker coloured shade to illustrate the data incompleteness problem and the cause for why this row is not present when filtering on the inclusion of a device name only.

The fourth device that issued a lease on IP address 5.6.7.8 showed the following data records around the time of the lease attempt, presented on the following table. At first, it seems logical for the device to attempt a DHCP lease for the IP address 5.6.7.8. Again, the coloured row in the table refers to the explained problem that this data record was not visible in the table displaying the DHCP lease history of the device, but that of the IP address itself. Thus, in the data table for the DHCP lease history of the device, the user sees two instances in a row where two different IP addresses are issued, which should be interpreted as either a data error or data incompleteness. Exclusion of the data record showing the IP address being abandoned one second after it being issued suggests an incomplete representation data quality problem as mentioned by Rodriguez et al. (2009).

Table 8 Data Quality Issue #2 - Issued to Issued for Different IP Addresses

Time	**Action**	**Lease IP**	**MAC Address**	**Host Name**	**Lease Start**	**Lease End**
04/12/2016 14:08:56	Freed	2.3.4.5	22:aa:33:bb:44:cc	*Workstation 456DEF*	2016-04-12 14:03:56	2016-04-12 14:08:56
04/12/2016 14:42:31	Issued	5.6.7.8	44:dd:55:ee:66:ff	*Workstation 456DEF*	2016-04-12 14:42:31	2016-04-12 22:42:31
04/12/2016 14:42:32	Aban-doned	5.6.7.8	-	-	2016-04-12 14:42:32	2038-01-19 04:19:00
04/12/2016 16:14:23	Issued	2.3.4.5	22:aa:33:bb:44:cc	*Workstation 456DEF*	2016-04-12 16:14:23	2016-04-12 16:19:23

However, these observations were simply found by chance after first globally inspecting the plausible values of the data in order to gain a better understanding of assessing whether the data is correct or accurate. More motivation to purposefully inspect an IP address could be that the IP map shows it is under conflict. Let us take an example where an IP address is in a conflict state and has last been discovered by NetMRI on March 31st, 2016 on 10:11:58 AM. Combining the DHCP lease history data of both the IP address and the associated host name delivers the following table. It is understandable how the conflict on the IP address 1.2.3.4 has been established in this example. The workstation has been freed from the IP address and was immediately issued a new one, leaving the old one in a free state. NetMRI then reports that the device has not been paired with the IP address ever since, and pinging the IP address dot not return a response. The description of this type of conflict states that the discovered MAC address is included within the DHCP range, but that there is no matching existing lease for that MAC address.

Table 9 Data Quality Issue #3 - No Matching Existing Lease based on MAC Address

Time	**Action**	**Lease IP**	**MAC Address**	**Host Name**	**Lease Start**	**Lease End**
03/31/2016 10:02:43	Re-newed	1.2.3.4	11:aa:22:bb:33:cc	*Workstation 123ABC*	2016-03-31 10:02:43	2016-04-03 10:02:43
03/31/2016 11:28:26	Freed	1.2.3.4	11:aa:22:bb:33:cc	*Workstation I23ABC*	2016-03-31 10:02:43	2016-03-31 11:28:26
03/31/2016 11:28:27	Issued	5.6.7.8	44:dd:55:ee:66:ff	*Workstation 123ABC*	2016-03-31 11:28:27	2016-04-03 11:28:27
03/31/2016 11:39:23	Re-newed	5.6.7.8	44:dd:55:ee:66:ff	*Workstation 123ABC*	2016-03-31 11:39:23	2016-04-03 11:39:23

It is not a prerequisite of an IP address to be issued by another device in order for it be under the conflict state. Instead, simply freeing the IP address without further activity shows the IP address in conflict within the IPAM data management section. Thus, it could be that regardless of whether the IP address has been issued by a different workstation or not after being freed from its previous lease, the state associated with that IP address on the IPAM data management section would be placed to '*conflict*'. IP addresses that have been abandoned till present time since the last discovery data are also labelled under conflict. The description then reads that the discovered address is in conflict with an existing lease, which happens to be the illusive lease till January 19th, 2038 given to those IP address which are labelled under the abandoned state. IP addresses under a conflict state that were initially fixed addresses for printers would be immediately freed from its fixed state and issued to a workstation, as both actions would take place on the same timestamp.

Data analysis on DHCP lease history has also been performed for those IP addresses holding a conflict where the discovered MAC address would be conflicting with the existing host paired to the IP address. Using a combined filter of the IP address in question with the newly discovered address would more than often search results of data records, but this is not always the case. For those IP address conflicts where no search results are produced via the filter, looking up the newly discovered MAC address could end up in finding records unrelated to the timeframe of the conflict, which leaves the question why it would be mentioned in the details of the conflict, related to the meaningless state data quality problem of Rodriguez et al. (2009) suggesting doubt whether or not the conflict has actually occurred.

In conclusion, the DHCP lease history data records do, to some extent, help the user with investigating the possible root cause of the IP address conflict in order to determine the best course of action regarding resolution. However, there are some data quality problems that have been detected, mostly under the category of data incompleteness, and may hinder the root-cause analysis of the IP address conflicts. Data incompleteness for specific data attributes, most notably within the data records where the IP action state is set to abandoned, may lead to users spending extra time to combine data records from both the device lease history and the IP address lease history in order to obtain a full sequence of actions within the timeframe of the conflict.

4.7.3 DHCP Top Lease Clients

One of the many other generated reports available is the *DHCP Top Lease Clients* showing the top-*n* MAC-addresses with the most activity, set standard over the last twenty-four hours. Activity is measured by counting the total number of instances where a MAC-address is issued, renewed, and freed. Taking the top 100 MAC-addresses as lease clients for the period between January 1st, 2016 and April 30th, 2016, it is possible to categorize the lease clients into three types, labelled as A, B, and C for this section. The following table represents some of the data records within the produced top-100 list for which all three types are visible; the point is that the data within the table depicts different lease behaviours of the MAC addresses.

Table 10 Multiple Data Records of Top 100 DHCP Lease Clients between January 1 - April 30, 2016

Rank	**Type**	**# Issued**	**# Renewed**	**# Freed**	**# Total**
1	C	73414	239270	72957	385641
2	A	184748	1	184747	369496
3	A	184626	3	184625	369254
11	B	5	41781	0	41786
12	B	4	36896	4	36904
13	B	4	36894	4	36902
28	C	7431	8640	3742	19813
37	A	7923	0	7923	15846

Type A shows that an MAC-address is frequently issued with and freed from a DHCP lease, but almost never renewed. For the three instances labelled with type A in the table, the MAC-addresses in all three data records were associated with a wireless access point. Its three-day issued static lease would usually last about fifty-one seconds until freed. Around five seconds after being freed, a new three-day lease for the same IP address is issued and would again last around fifty-one seconds, repeating the cycle. To put this into perspective, for the second and third-ranked top lease client, this cycle would repeat itself over 1500 times a day (184748 times issued / 121 days, or (24 hours * 3600 seconds / (3600 seconds / 56 second cycle duration). For rank #37, the time between the lease being freed and re-issued was twelve seconds, with the same fifty-one second-long issue period, and took place between January 8 and January 13, 2016, repeating its cycle for over 1600 times a day (7923 / (116,05 hours / 24 hours a day). Since the IP address was static for all three cases, there is a case of conflict regarding similar MAC addresses in the same network. To be reminded, MAC addresses should be unique within each leaf network, but technically the same MAC address can be used on multiple networks.

Type B shows that DHCP leases are frequently renewed for the same device, in this case workstations for development. For rank 12 and 13 in the table, this can be seen as a typical result: inspecting the DHCP lease history on both MAC addresses show that the lease duration was set on five minutes, forcing such frequent renewals of the DHCP leases. For rank 11 however, the lease was set on three days, and would typically be renewed within ten minutes, but most often after three minutes and thirty seconds. This seems odd, even more so when five-minute DHCP leases would usually be renewed on the halfway point, being two minutes and thirty seconds. This means that there is doubt whether this was purposefully configured or whether DHCP lease durations were misconfigured for the device associated with the eleventh ranking on the DHCP top lease client list. In either case, it creates a tremendous amount of data records and ‘clutter’ within the DHCP lease history generated reports.

Finally, Type C shows very erratic behaviour as DHCP leases could be issued, renewed, or freed without a clear pattern. Inspection on the number one-ranked top client could show a sequence of events similar to the following table, for which a five-minute window on May 2nd, 2016 has been monitored between 4:54 PM and 4:59 PM. Both the associated lease IP address and the host name are unique for all fifteen recorded data records in the table. It should be noted that the earlier discussed data quality problem of incompleteness regarding the lack of data records showing DHCP leases being abandoned has not been resolved in this table, and thus it is possible that for two consecutive data records holding the action ‘*issued*’ a data record showing the previous lease being abandoned could be in place in case the name of the workstation was included in that particular data record. Of course, as discussed earlier, the main problem of these data records is that the name of the workstation lacks most of the time for data records showing DHCP leases being abandoned.

Nevertheless, the sequence of DHCP actions included in the table is unusual. First off, from the fourth till the sixth data row, DHCP leases get renewed for three different IP addresses, two of which belong to the same leaf network (0.0.0.3 and 0.0.0.4). As mentioned earlier, technically this should be impossible as MAC addresses are to be unique within the scope of a (leaf) network. The same technical issue is visible throughout the entire sequence of the fifteen data records; twelve of the fifteen IP addresses belong to the same leaf network 0.0.0.x, and the remaining three belong to the leaf network 0.0.1.x (2) and 0.0.2.x (1). As the MAC address hops from one workstation to the other, there seems to be a conflict regarding MAC address duplication on the same leaf network attempting to correct itself constantly, causing a tremendous amount of clutter within the data records. Furthermore, the pattern of DHCP actions cannot be deduced and appears to be random, unlike in the previous explained types. This type should therefore be considered holding a non-logical and theoretically incorrect DHCP lease behaviour, and should raise concern to network administrators. Given the total number of DHCP actions for the top lease client (385,641) for the first four months of 2016, with rough calculations this comes down to an average of 2.2 DHCP lease actions per minute. As long as the associated MAC address cannot hold a more stable DHCP lease under condition that no MAC address duplication conflict arises within the same (leaf) network, this cycle could go on indefinitely creating many meaningless data records and wasting database storage space, making it harder for data analysts to filter through the data to possibly find meaningful results.

The second type-C entry in the top-100 lease client table is related to a printer which in 2016 had two notable periods. Starting from January 8th, the printer was ‘juggling’ between issuing, renewing and freeing two different IP addresses (e.g. 1.2.3.4 and 5.6.7.8), causing a DHCP action data record to be generated every five to ten seconds in the lease history reports. Eventually, a third IP address got involved (9.10.11.12) and the juggling continued between one of the original associated IP addresses (1.2.3.4) and the newly introduced one. Again, in both instances, lease duration was originally set to three days. This continued till 4:44 PM on January 11th, and no further lease history of the MAC-address was registered until April 14th, where it would issue its lease on the IP address 1.2.3.4 with a duration of half an hour and renew its lease every fifteen minutes. This seemed to have fixed the problem: when the printer would be freed from this IP address, the next DHCP lease issued would be on the same IP address, showing a logical cycle of the IP address being freed, re-issued, renewed, freed and re-issued again.

These instances should be considered a data quality problem not only for previously mentioned reasons, but because the data should be considered as inaccurate. Even with perfect scores on data timeliness, a data record stating that the MAC address is renewed on one IP-address (0.0.0.3) and fourteen seconds later renewed on another IP-address on the same network (0.0.0.4) would indicate that one of the two data records would be impossible, since neither an action of abandoning nor issuing a DHCP lease is included between the two data records. At least one of the two data records here should therefore be false, impacting data quality negatively on the accuracy dimension.

Table 11 Five-Minute Snippet of DHCP Lease Records of Top DHCP Lease Client between January 1 - April 30, 2016

Time	**Action**	**Lease IP**	**MAC Address**	**Host Name**	**Lease Start**	**Lease End**
05/02/2016 16:54:31	Freed	0.0.0.1	11:aa:22:bb:33:cc	*Workstation 001*	2016-05-02 16:39:13	2016-05-02 16:54:31
05/02/2016 16:54:53	Renewed	0.0.1.1	11:aa:22:bb:33:cc	*Workstation 002*	2016-05-02 16:54:53	2016-05-02 18:54:53
05/02/2016 16:55:10	Issued	0.0.0.2	11:aa:22:bb:33:cc	*Workstation 003*	2016-05-02 16:55:10	2016-05-02 18:55:10
05/02/2016 16:56:18	Renewed	0.0.0.3	11:aa:22:bb:33:cc	*Workstation 004*	2016-05-02 16:56:18	2016-05-02 18:56:18
05/02/2016 16:56:32	Renewed	0.0.0.4	11:aa:22:bb:33:cc	*Workstation 005*	2016-05-02 16:56:32	2016-05-02 18:56:32
05/02/2016 16:56:41	Renewed	0.0.2.1	11:aa:22:bb:33:cc	*Workstation 006*	2016-05-02 16:56:41	2016-05-02 18:56:32
05/02/2016 16:57:23	Issued	0.0.0.5	11:aa:22:bb:33:cc	*Workstation 007*	2016-05-02 16:57:23	2016-05-02 18:57:23
05/02/2016 16:57:43	Renewed	0.0.1.2	11:aa:22:bb:33:cc	*Workstation 008*	2016-05-02 16:57:43	2016-05-02 18:57:43
05/02/2016 16:57:47	Issued	0.0.0.6	11:aa:22:bb:33:cc	*Workstation 009*	2016-05-02 16:57:47	2016-05-02 18:57:47
05/02/2016 16:57:58	Issued	0.0.0.7	11:aa:22:bb:33:cc	*Workstation 010*	2016-05-02 16:57:58	2016-05-02 18:57:58
05/02/2016 16:58:05	Freed	0.0.0.8	11:aa:22:bb:33:cc	*Workstation 011*	2016-05-02 16:33:51	2016-05-02 16:58:05
05/02/2016 16:58:32	Issued	0.0.0.9	11:aa:22:bb:33:cc	*Workstation 012*	2016-05-02 16:58:32	2016-05-02 18:58:32
05/02/2016 16:58:46	Freed	0.0.0.10	11:aa:22:bb:33:cc	*Workstation 013*	2016-05-02 16:51:56	2016-05-02 16:58:46
05/02/2016 16:58:55	Renewed	0.0.0.11	11:aa:22:bb:33:cc	*Workstation 014*	2016-05-02 16:58:55	2016-05-02 18:58:55
05/02/2016 16:59:01	Issued	0.0.0.12	11:aa:22:bb:33:cc	*Workstation 015*	2016-05-02 16:59:01	2016-05-02 18:59:01

As has been explained earlier in the case study, the organization recently discovered that events for duplicate IP address conflicts would generate an alert to the MCR where members hold the authority of delegating such resolution tasks over to other network operations groups. Such alert should also be generated for the event that the same MAC-address is used over multiple IP addresses in the same leaf network, or when an unusual amount of DHCP actions associated with a particular MAC-address or workstation is registered via NetMRI. The description of the second so-called type-C entry shows that it took around three days to stop any DHCP lease activity associated to that MAC-address until a stable fix was found. The generated DHCP lease history data records over these three days are not meaningful to any data analyst other than observing a major network configuration flaw being present, and should be alerted and resolved much sooner.

4.7.4 Overview of Data Quality Issues in Generated Network Data Reports

The discovered data quality issues within generated data reports based on NetMRI's discovered data, along with affected KPIs and levels of severity, are now summarized:

- [Data accuracy] – For IP address utilization, the number of reserved addresses is always set to '2', even for leaf networks holding one IP address. Thankfully, the percentage of IP address utilization is still calculated correctly, otherwise negative percentages would have been computed.
  - KPIs possibly affected: % error rate, operational expenditures; the latter would have been possibly affected if the computed percentage of IP address utilization was proven to be false, which could have resulted in network capacity investment driven by dirty data.
  - Level of severity: Normal
  - Possible solution: Since percentage of IP address utilization is still computed correctly, for accuracy purposes a CFD should be introduced that for leaf networks holding either one or two IP addresses, the number of reserved IP addresses should be set to zero.
- [Data accuracy] – Top 500 utilized networks are dominated by one, two, or 4 IP address block leaf networks, making the generated report less useful.
  - KPIs possibly affected: operational expenditures; inaccurate or incomplete representation of the top utilized networks can lead stakeholders to think that there is a network capacity issue and more investments should be made.
  - Level of severity: Low
  - Possible solution: Change the ranking of top utilized networks from percentage of IP address utilization to the absolute number of allocated IP addresses. This way, larger leaf networks holding, for instance, a hundred allocated addresses are likely to be presented in the top utilized networks.
- [Data incompleteness] – Host name and MAC address data attributes are missing for DHCP lease history data records with DHCP action '*abandoned*'.
  - KPIs possibly affected: % error rate, business continuity; more effect is needed to extract all possible information of which device is referred to. Abandonment of a lease related to a critical component of the network without the data attributes describing the device would go unnoticed, causing further IPAM problems.
  - Level of severity: Significant
  - Possible solution: Deduce the MAC address and host name by querying the most recent DHCP lease history data record which matches on the *Lease IP* data attribute. This also eliminates for data analysts to perform

multiple queries on DHCP lease history of either an IP address or a hostname/MAC address.

- [Data accuracy] – *Lease end* data attribute of DHCP lease history data records with DHCP action '*abandoned*' have a fixed timestamp of 2038-01-19.
  - KPIs possibly affected: % error rate; filtering all DHCP lease history data on *lease end* timestamps with most recent on top would show mostly those data records where DHCP leases were abandoned.
  - Level of severity: Low
  - Possible solution: Remove *lease end* timestamp for such records as it is illogical to present a timestamp of when an IP address is abandoned for lease. Instead, use only the timestamp of abandonment on the *lease start* data attribute.
- [Data accuracy] – *Lease end* data attribute of DHCP lease history data records with DHCP action '*issued*' or '*fixed*' have a *lease end* timestamp of 2038-01-19, which violates the DHCP lease duration policy.
  - KPIs possibly affected: % error rate; IP address is continuously reserved which has an effect on IP address utilization figures. Furthermore, device in question would remain access to Internet even when rogue.
  - Level of severity: Significant to critical
  - Possible solution: Since the maximum DHCP lease duration is set to three days, such occurrences are in violation with DHCP lease policy and alerts should be triggered on the basis that any end lease timestamps referring to a point in time beyond three days in the future (but most likely only those holding 2038-01-19) are recognized and immediately abandoned.
- [Data accuracy] – Multiple MAC addresses are associated in an extreme amount of DHCP lease history data records showing erratic behavior of issuing, renewing, and freeing DHCP leases, often generating more than two data records per minute. This creates a lot of clutter within the database. This is mainly the result of identical MAC addresses being present in the same network.
  - KPIs possibly affected: % error rate, business continuity, latency; critical DNS servers could be in jeopardy when identical MAC addresses are juggling from one IP address to the other within the same leaf network, but thus far it has been identified for workstations only.
  - Level of severity: Significant to critical
  - Possible solution: Alerts should be generated for such occurrences where identical MAC addresses are trapped in multiple IP addresses within the same leaf network. The sooner such instances are identified, the less data clutter there will be in the data record history logs, and the less workload the network scanner has to endure.

5 DISCUSSION & CONCLUSION

This final chapter of the thesis will discuss the case findings and based on these findings paired with the theoretical background described in Chapter 2, a conclusion will be made in which the research question introduced in Chapter 1 will be answered as complete as possible. Of course, limitations have to be acknowledged regarding internal and external validity and the scope of the research performed for this thesis. Furthermore, possible ideas for future research on the topic of network automation systems and data quality assessment should be mentioned.

5.1 Discussion

The motivation of the organization to merge the network grids of both the Belgian and the Dutch site can be seen as a practice of eliminating business silos as was discussed earlier using the work of Redman (1998). The impacts of poor data quality discussed in the case study were mostly categorized on the operational and tactical level given the established KPIs of cutting costs on operational expenditures and having a higher level of visibility to make network capacity and workflow decisions. Furthermore, the established KPIs in the case study were mainly process- and outcome-based (Masayna et al., 2007). While the number of IP address conflicts may not necessarily be seen as a legitimate proxy to measure the level of data quality within IPAM-related practices, the organization does have the capability to use the network monitoring system to perform data triangulation to validate the best, 'true' candidate to hold the IP address, fulfilling one of the needs addressed by Liebchen & Shepperd (2008). The benefits of network automation practices were shown in the example where IP address blocks for Wi-Fi connections on different office floors were rearranged, showcasing that this decision was based on proper, timely data (Chengalur-Smith et al., 1999). This also shows that Lee et al.'s (2014) cycle of monitoring, design, and deployment for network operations can be seen using this example. Using Fan & Geerts' (2012) breakdown of data quality dimensions, we have seen that information completeness and accuracy are the most stressed dimensions given the problems identified within the case study. Two of the three data quality problems described by Rodriguez et al. (2009) have been identified with the given examples of data quality issues in available generated reports. It should be noted however that the combination of the Infoblox DDI system for network operations and NetMRI for network scanning leads to data holding the Closed World Assumption (Abiteboul, 2006), meaning that all data of the network entity has been scanned and no other explanation outside that of the properties of the database can be given for data incompleteness issues. As for accuracy, it seems more attention has been given to semantic rather than syntactic data accuracy (Batini et al., 2009; Zaveri et al., 2015), meaning the correct state of an entity should be represented rather than a likely accepted value. Especially with data accuracy issues within the topic of DHCP leases, the semantic type is strongly addressed. Another issue is the non-uniform use of timestamp to conform to data currency, which was addressed by Dong et al. (2009) and discussed in the theoretical background. As for the problem with the number of reserved addresses on a one-IP address block, it could be interpreted as biased data based on a wrongly implemented methodology (Wesslén, 2000).

The typical data quality methodology steps of Batini et al. (2009) being state reconstruction, assessment, and improvement, can be applied within the scope of the case study, holding a data-driven strategy where the main focus rests on data values. Improvements on data collection techniques could be done by applying conditional functional dependencies based on the work of Fan (2008). For instance, a CFD can be developed and implemented to resolve the data incompleteness issue for when the host name of a workstation is missing from a DHCP lease history data record where the DHCP action is set to ‘abandoned’. Configuring the CFD to add the host name previously included in the last registered DHCP lease history data record sharing the same IP address before the instance of abandonment should resolve this particular data quality issue. This proposed solution is supported by Fan’s (2008) statement that CFDs are to be used for data repairing purposes and to impute new data merely to improve on the original data databases. As for the second methodology step by Batini et al. (2009), the usage of a network management system solution like Infoblox DDI allows for earlier assessment of data quality and root-cause analysis of quality issues or conflicts. Also, as discussed earlier, data-driven techniques are in favour when focusing on one-time solutions or cost optimization on the short term. This is applicable for the discussed case study given its timeline and the inclusion of an economic KPI regarding operational expenditures.

Batini et al. (2009) discuss how data quality improvement should be done via its assessment using a methodology concerning audit, operational, and economical issues. Using their breakdown of information systems types for data quality monitoring practices, the combination of the Infoblox grid members having their own assigned tasks with NetMRI functioning as a network scanner can be categorized as a distributed and cooperative information system. To add to that, Infoblox presents its DDI-solution to achieve consistency between service and management views of data, meaning no extra redundancy is needed to prevent risks in inconsistent, simultaneous updating, which for a distributed and cooperative information system could have been an added challenge. Interestingly, from the case study it is not directly assumed that further data quality activities were present, given the fact that KPIs were already established. Taking Masayna et al.’s (2007) framework and the case background information, both organizational and external influences already placed pressure on data quality to develop and accurately measure KPIs. However, the data quality in question here would be heavily dependent on the performance of NetMRI as a network discovery system using its implemented discovery protocols (Reddy et al., 2014).

Current concerns of the organization that there is a lack of alert notification on duplicate IP address events means that the analysis layer in monitoring operations systems needs to be reconfigured in order to be integrated with other event management systems (Lee et al., 2014). Within the analysis layer, both fault identification and fault localization are possible using Infoblox DDI (Lee et al., 2014), but it leaves room for more automation. This is because for now, each conflict is sent to the master control room only for it to be reassigned to a different team to resolve it. Furthermore, Lee et al.'s (2014) discussion of active versus passive polling is not directly applicable in the case study, since a hybrid of the two is available depending on the type of generated report used. Thus, the monitoring system in place allows for a balance between accuracy and efficiency. However, aggregated data flow techniques are used in multiple generated reports (e.g. DHCP Lease History) and create more risk in data incompleteness (Lee et al., 2014), which has been illustrated in the case study. Finally, while the proposed passive network appliance monitoring system of Schultz et al. (2011) combines real-time and summary logs of packet data flow, the monitoring system used at the organization utilizes a similar approach where in some data reports, snapshots can be created of either real-time or summary event data logs. This allows for easier interpretation of network data and simplifies the process of visualizing network performance in historic points.

5.2 Conclusion

There might be no ambiguous answer to the question how data quality could be improved within centralized network management systems given the different short-term and long-term perspectives and the variety of data quality problems, but in order to achieve more automation within the cycle of monitoring, configuration, and deployment, data quality assessment practices should be in place nonetheless with a stronger focus on the long-term perspective. Although it was already predicted that network management systems themselves are not intended to document data quality problems and repair them, the accuracy of the representation of the network's capability and behavior is dependent on data quality on the dimensions of completeness, timeliness, and accuracy. The benefits of improving (network) data quality are in conformance with the theoretical background, being better decision-making to stakeholders to conform with business strategy to achieve alignment on the operational, tactical, and strategic level, and to eliminate the possibility of business silos. The latter can be achieved by using such centralized platform combined with multiple access rights levels to achieve segregation of duties to mitigate the risk of unauthorized data manipulation while maintaining a high level of visibility. The most notable example was how members in the Master Control Room held the authority to assess the criticality of each event alert and delegate it to other teams to resolve the issue related to that event.

Sending alerts of data quality-related issues over to an event management system allows for a more direct and effective approach of assessing overall data quality. However, as shown with the organization's concern of not being alerted of duplicate IP address conflicts, there is no guarantee that all data quality-related events get transferred via the system log of the network scanner from one system to another, in this case NetMRI to a separate event management system. This poses a risk not only in terms of data quality in itself, but also in security auditing. This means that automation of data quality assessment is still hindered, forcing organizations to either develop data repairing techniques to be implemented in these systems using any of the discussed data quality improvement methodologies, or to perform manual work on data quality assessment and discover data quality issues that are either new or were initially undetected. Nevertheless, the introduction of a centralized network management system is a significant step toward network process automation since no manual updates of the data are needed for, to give an example, IPAM that was initially stored and updated via spreadsheets. It should also be noted that since the network management system was already in place at the Dutch site several years ago and integration of both the Belgian and Dutch grid members was achieved relatively fast, the system has had the opportunity to mature in terms of monitoring and long-term goals of achieving automation via improved data quality assessment seems therefore fitting.

Data quality is going to be as good as the performance of the discovery protocols applied by the network scanner, in this case NetMRI. Data quality and automation therefore are related to each other: proper data quality is needed to ensure that processes can be automated with a high level of certainty and, ideally, automated process should have the capability to determine whether there are possible data quality issues, such as in data incompleteness, leading to a direct alert notification to MCR and cancellation of further processes to avoid the scenario of stakeholders being presented with KPIs based on dirty data. In order for organizations to benefit from increased speeds of network maintenance and monitoring processes working towards the possibility of automation, improper data quality can be seen as one of the major, if not the biggest hurdle to that objective. Of course, the data quality issues that have been discovered and documented in the case study were the result of manual analysis of the available generated data reports. Such documentation is still needed rather than having an automatically generated report of the event management system describing the problem in static terms. This is because human interpretation of data quality issues can be communicated in such a way that multiple groups of stakeholders can understand the impact of the data quality problem for their related KPIs. Also, the organization can gain a better insight of the possible weaknesses of the discovery protocols used to fetch the data and where the challenges of integration network management systems with event management systems are to improve automation in IT. This confirms with Batini et al.'s (2009) research that a process-driven data quality improvement methodology is preferred to understand the cause of network data quality problems in the long term, as automation should be considered a goal in the long term.

It should be noted that the term ‘automation’ is interpreted in two ways. Automation is used in the sense that network data can be fetched and updated automatically using network discovery protocols by NetMRI and transferred over to a presentation layer for data analysis purposes. While this cuts down operational expenditures on data fetching and eliminates manual updating, automation should also be interpreted in the sense that organizations attempt to promote automation in terms of data analysis, and indirectly to network security monitoring. In conclusion, this work has shown that there are possible challenges of data quality in network management systems in terms of knowledge, visibility, and criticality paired with achieving process automation.

5.3 Limitations & Future Research

It should not be assumed that the case study findings can be generalized to other organizations using network management systems similar to the one described in this thesis, thus acknowledging a limitation that this research was primarily focused on internal validity and underperforms in terms of external validity. This is primarily due to the fact that the thesis focuses on a single, intensive case study. Direct results of resolving identified data quality issues were not measurable given the time constraint, and also because the organization proposes an action plan in the future. Furthermore, due to the scope of the research, other topics such as data governance, security compliance, and cost analysis of data quality improvement have been left out and are to be considered relevant topics when discussing data quality improvement practices in general. Further research could include the full appliance of a complete data quality improvement methodology as coined by Batini et al. (2009), and possibilities of a passive network appliance monitoring system, as described by Schultz et al. (2001), could be investigated further having in mind the challenges of identifying and transferring data quality-related issues over from a network management system to an event monitoring system. The purpose of this thesis was merely to explore data quality challenges within the field of network operations and to establish a link how data quality assessment policies allow organizations to have a better chance at automating processes with economic benefits on the long-term.

REFERENCE LIST

Abiteboul, S., Segoufin, L., & Vianu, V. (2006). Representing and querying XML with incomplete information. *ACM Transactions on Database Systems (TODS), 31*(1), 208-254.

Alles, M., Brennan, G., Kogan, A., & Vasarhelyi, M. A. (2006). Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. *International Journal of Accounting Information Systems*, *7*(2), 137-161.

Ashford, S. J., & Black, J. S. (1996). Proactivity during organizational entry: The role of desire for control. *Journal of Applied psychology*, *81*(2), 199.

Batini, C., Barone, D., Mastrella, M., Maurino, A., & Ruffini, C. (2007). A Framework And A Methodology For Data Quality Assessment And Monitoring. In *ICIQ* (pp. 333-346).

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR), 41*(3), 16.

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

Bellovin, S. M. (1995, June). Using the Domain Name System for System Break-ins. In *USENIX Security*.

Bitektine, A. (2007). Prospective case study design: qualitative method for deductive theory testing. *Organizational Research Methods*.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Englewood Cliffs, N.J.: Prentice-Hall.

Caballero, I., Verbo, E., Calero, C., & Piattini, M. (2007). A Data Quality Measurement Information Model Based On ISO/IEC 15939. In *ICIQ* (pp. 393-408).

Caro, A., Calero, C., Caballero, I., & Piattini, M. (2006). Defining a data quality model for web portals. In *Web Information Systems–WISE 2006* (pp. 363-374). Springer Berlin Heidelberg.

Chengalur-Smith, I. N., Ballou, D. P., & Pazer, H. L. (1999). The impact of data quality information on decision making: an exploratory analysis. *Knowledge and Data Engineering, IEEE Transactions on*, *11*(6), 853-864.

CICA/AICPA (1999). *Continuous auditing: research report*. Canadian Institute of Chartered Accountants. Toronto, Canada.

CobiT (2004). Control objectives for information and related technology (CobiT). 4th Edition.

Control [Def. 1.1, 1.2, 1.4] (n.d.). In Oxford Dictionaries. Retrieved on March 4th, 2016 from <http://www.oxforddictionaries.com/definition/english/control>. Oxford University Press.

Daniel, P., & Keegan, R. (1989). Are your performance measures obsolete?. Management accounting, 45.

Danzig, P. B., Obraczka, K., & Kumar, A. (1992). An analysis of wide-area name server traffic: a study of the Internet Domain Name System. *ACM SIGCOMM Computer Communication Review*, *22*(4), 281-292.

De Veaux, R. D., & Hand, D. J. (2005). How to lie with bad data. *Statistical Science*, 231-238.

Denzin, N. (1970). Strategies of multiple triangulation. *The research act in sociology: A theoretical introduction to sociological method*, *297*, 313.

Dimension [Def. 2] (n.d.). In Oxford Dictionaries. Retrieved on March 4th, 2016 from <http://www.oxforddictionaries.com/definition/english/dimension>. Oxford University Press.

Dong, X. L., Halevy, A., & Yu, C. (2009). Data integration with uncertainty. *The VLDB Journal—The International Journal on Very Large Data Bases*, *18*(2), 469-500.

Dubois, A., & Gadde, L. E. (2002). Systematic combining: an abductive approach to case research. *Journal of business research*, *55*(7), 553-560.

Dyer, W. G., & Wilkins, A. L. (1991). Better stories, not better constructs, to generate better theory: A rejoinder to Eisenhardt. *Academy of management review*, *16*(3), 613-619.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review*, *14*(4), 532-550.

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc..

Eriksson, P., & Kovalainen, A. (2008). Qualitative research evaluation. *Qualitative methods in business research*.

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

Fan, W. (2008, June). Dependencies revisited for improving data quality. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 159-170). ACM.

Fan, W., Geerts, F., & Jia, X. (2009). Conditional dependencies: A principled approach to improving data quality. In *Dataspace: The Final Frontier* (pp. 8-20). Springer Berlin Heidelberg.

Fan, W., & Geerts, F. (2012). Foundations of data quality management. *Synthesis Lectures on Data Management*, *4*(5), 1-217.

Fisher, C. W. (1999). An empirically based exploration of the interaction of time constraints and experience levels on the data quality information (DQI) factor in decision making. *University at Albany, Albany, NY*.

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management*, *39*(2), 109-116.

Foss, S., & Waters, W. (2007). Destination dissertation. *A Traveler's Guide to A Done Dissertation. Langham, Maryland: Rowman and Littlefield Publishers*.

Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York: Basic Books.

Ghose, A., & Koliadis, G. (2007). *Auditing business process compliance* (pp. 169-180). Springer Berlin Heidelberg.

Glaser, B. S., & Strauss, A. (1967). The discovery of grounded theory. *Strategies for qualitative research. London: Weidenfeld and Nicolson*.

Hand, D. J. (1998). Reject inference in credit operations. *Credit risk modeling: Design and application*, 181-190.

Hanemann, A., Liakopoulos, A., Molina, M., & Swany, D. M. (2006). A study on network performance metrics and their composition. *Campus-Wide Information Systems*, *23*(4), 268-282.

Herrera, Y. M., & Kapur, D. (2007). Improving data quality: actors, incentives, and capabilities. *Political Analysis*, *15*(4), 365-386.

Hey, N. (2006). A process-centric approach for selecting critical information domains for targeted data quality improvement. Proceedings of Strategic Data Quality Management.

Infoblox (2012). Infoblox Administrator Guide - NIOS 6.3 for Infoblox Core Network Services Appliances. Retrieved from <https://www.infoblox.com/support/tech.../NIOS_AdminGuide_6.3.pdf> on April 1st, 2016.

Infoblox (2015). Infoblox Enterprise-grade DDI. Retrieved from <https://www.infoblox.com/sites/infobloxcom/files/resources/infoblox-datasheet-infoblox-enterprise-grade-ddi_0.pdf> on February 22, 2016.

Infoblox (2015). The Infoblox Grid. Retrieved from <https://www.infoblox.com/sites/infobloxcom/files/resources/infoblox-datasheet-the-infoblox-grid.pdf> on February 22, 2016.

ISO/IEC. (2000). *ISO/IEC 15939. Information Technology – Software Management Process*.

Jack, E. P., & Raturi, A. S. (2006). Lessons learned from methodological triangulation in management research. *Management Research News*, *29*(6), 345-357.

Johnson, P. M., & Disney, A. M. (1999). A critical analysis of PSP data quality: Results from a case study. *Empirical Software Engineering, 4*(4), 317-349.

Kaplan, R. S., & Norton, D. P. (1996). The balanced scorecard: translating strategy into action. Harvard Business Press.

Kind, A., Dimitropoulos, X., Denazis, S., & Claise, B. (2008). Advanced network monitoring brings life to the awareness plane. *Communications Magazine, IEEE, 46*(10), 140-146.

Kogan, A., Sudit, E. F., & Vasarhelyi, M. A. (1999). Continuous online auditing: A program of research. *Journal of Information Systems*, *13*(2), 87-103.

Lee, J., Lee, D., & Kang, S. (2007). An overview of the business process maturity model (BPMM). In *Advances in web and network technologies, and information management* (pp. 384-395). Springer Berlin Heidelberg.

Lee, S., Levanti, K., & Kim, H. S. (2014). Network monitoring: Present and future. *Computer Networks*, *65*, 84-98.

Lichiello, P., & Turnock, B. J. (1999). *Guidebook for Performance measurement*. Turning Point.

Liebchen, G. A., & Shepperd, M. (2008, May). Data sets and data quality in software engineering. In *Proceedings of the 4th international workshop on Predictor models in software engineering* (pp. 39-44). ACM.

MacMillan, L. (2007). *Strategies for Successful Scorecards: Key to Performance Management Initiatives*.

Madnick, S., & Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering, 59*(2), 460-475.

Masayna, V. (2006). *A Framework for Linking Data Quality Efforts to Organisational Key Performance Indicators* (Doctoral dissertation).

Masayna, V., Koronios, A., Gao, J., & Gendron, M. (2007, June). Data quality and KPIs: a link to be established'. In *The 2nd World Congress on Engineering Asset Management (EAM) and The 4th International Conference on Condition Monitoring*.

Mendes, E., & Lokan, C. (2008). Replicating studies on cross-vs single-company effort models using the ISBSG Database. *Empirical Software Engineering, 13*(1), 3-37.

Monitor [Def. 1, 1.2] (n.d.). In Oxford Dictionaries. Retrieved on March 4th, 2016 from <http://www.oxforddictionaries.com/definition/english/monitor>. Oxford University Press.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst*, *10*, 26.

Perkins, C. E., & Luo, K. (1995). Using DHCP with computers that move0. *Wireless Networks*, *1*(3), 341-353.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

Pollack, A. (1999, October 1). Two Teams, Two Measures Equaled One Lost Spacecraft. *The New York Times*. Retrieved from http://partners.nytimes.com/library/national/science/100199sci-nasa-mars.html

Pras, A., Schönwälder, J., Burgess, M., Festor, O., Perez, G. M., Stadler, R., & Stiller, B. (2007). Key research challenges in network management. *Communications Magazine, IEEE*, *45*(10), 104-110.

Princeton University. (2014, December 12). *Android 2.1 - 4.1.1 Allows DHCP Lease to Expire, Keeps Using IP Address*. Retrieved from <https://www.net.princeton.edu/android/android-stops-renewing-lease-keeps-using-IP-address-11236.html> on April 12, 2016.

Reddy, R. A., Swamy, J. N., & Reddy, R. G. (2014). Detecting Embedded Devices using Network Discovery.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79-82.

Roberts, P. A., & Challinor, S. (2000). IP address management. *BT Technology Journal*, *18*(3), 127-136.

Rodríguez, C., Daniel, F., Casati, F., & Cappiello, C. (2009). Computing Uncertain Key Indicators from Uncertain Data. *ICIQ*, *9*, 106-120.

Saaty, T. L. (2004). Decision making—the analytic hierarchy and network processes (AHP/ANP). *Journal of systems science and systems engineering*, *13*(1), 1-35.

Sauser, B. J., Reilly, R. R., & Shenhar, A. J. (2009). Why projects fail? How contingency theory can provide new insights–A comparative analysis of NASA's Mars Climate Orbiter loss. *International Journal of Project Management*, *27*(7), 665-679.

Schultz, M. J., Wun, B., & Crowley, P. (2011, October). A Passive Network Appliance for Real-Time Network Monitoring. In *Architectures for Networking and Communications Systems (ANCS), 2011 Seventh ACM/IEEE Symposium on* (pp. 239-249). IEEE.

Silverman, D. (2000) Doing qualitative research: A practical handbook. Thousand Oaks: Sage Publication.

Sohn, S. Y., & Shin, H. W. (2006). Reject inference in credit operations based on survival analysis. *Expert Systems with Applications*, *31*(1), 26-29.

Stake, R. E. (1995). *The art of case study research*. Sage.

Stoecker, R. (1991). Evaluating and rethinking the case study. *The sociological review, 39*(1), 88-112.

Sufi Abdi, B. (2013). Framework for Measuring Perceived Quality in Technical Documentation.

Tarokh, M. J., & Nazemi, E. (2006). Technical Note: Performance measurement in industrial organizations, case study: Zarbal Complex. Journal of Industrial Engineering International Islamic Azad University, 2(3), 54-69.

Vasarhelyi, M. A., Alles, M. G., & Kogan, A. (2004). Principles of analytic monitoring for continuous assurance. *Journal of emerging technologies in accounting, 1*(1), 1-21.

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

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

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5-33.

Wasserman, G. S., & Lindland, J. L. (1996). A case study illustrating the existence of dynamics in traditional cost-of-quality models. *Quality Engineering*, *9*(1), 119-128.

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All---A Contingency Approach to Data Governance. *Journal of Data and Information Quality (JDIQ)*, *1*(1), 4.

Weill, P. (2004). Don’t just lead, govern: How top-performing firms govern IT. *MIS Quarterly Executive, 3*(1), 1-17.

Wende, K. (2007). A Model for Data Governance-Organising Accountabilities for Data Quality Management. *ACIS 2007 Proceedings*, 80.

Wende, K., & Otto, B. (2007). A Contingency Approach To Data Governance. In *ICIQ* (pp. 163-176).

Wesslén, A. (2000). A Replicated Empirical Study of the Impact of the Methods in the PSP on Individual Engineers. *Empirical Software Engineering*, *5*(2), 93-123.

Wetzstein, B., Leitner, P., Rosenberg, F., Brandic, I., Dustdar, S., & Leymann, F. (2009, September). Monitoring and analyzing influential factors of business process performance. In *Enterprise Distributed Object Computing Conference, 2009. EDOC'09. IEEE International* (pp. 141-150). IEEE.

Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for linked data: A survey. *Semantic Web, 7*(1), 63-93.

Zseby, T., Hirsch, T., & Claise, B. (2008). Packet sampling for flow accounting: Challenges and limitations. In *Passive and Active Network Measurement* (pp. 61-71). Springer Berlin Heidelberg.

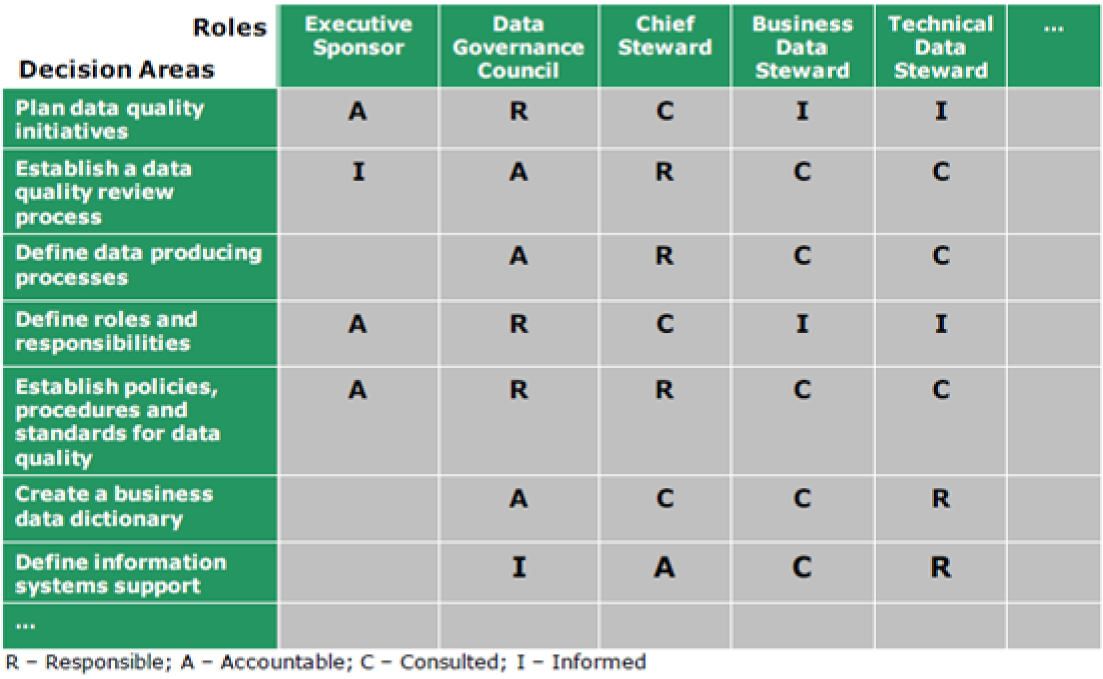
APPENDIX

A. Data Governance

While the importance of data quality and possible methodologies for improving data quality has been discussed, one of the questions remaining is who should be held responsible for managing data quality and the attempts for its improvement. Over the last ten years, the term '*data governance*' has been used more frequently with an emphasis on collaboration between the IT department and other organizational bodies. When thinking about the term 'data governance', one may ask whether or not it would be similar or fall under the category of IT governance. In the paper of Wende (2007), a possible data governance model is discussed. The author uses the definition of IT governance from Weill (2004), it being the goal of achieving desirable behavior of using IT by specifying a framework for accountability and decision rights amongst different stakeholders, and alters it to the goal of achieving desirable behavior of using data. Wende (2007) attempts to create a data governance model stating that other approaches to data quality management are mainly targeted to one specific role. An example is the single role of information product manager in Total Data Quality Management, or TDQM, which was one of the methodologies discussed by Batini et al. (2009) and developed by Wang (1998). A point that Wende (2007) would like to address is that data quality management should be considered a daily process of making decisions; data governance is merely a framework for establishing what decision rights are given to management.

Five roles have been established for the data governance model of Wende (2007). The executive sponsor is the embodiment of oversight; this could be a CEO or CIO position. The data quality board develops the strategy, standards and rules applied to the improvement plan of data quality company-wide; alignment between the data quality improvement strategy and the objectives of the company has to be achieved. The chief steward is partly responsible for creating such alignment by aiding with the adoption of standards. Business and technical data stewards are then busy with processes which could be related to data semantics and overall presentation for decision-makers: business data stewards are occupied with developing metrics for data quality, data vocabularies and business rules for using and managing data, while technical data stewards define the data elements and the correct formatting.

Several decision areas that are to be included in the data governance model are standards and policies, the data quality strategy, its management processes and the architecture of data, for example the development of a data dictionary. The decision areas are broken down in three areas: strategy, organization, and information systems. For the organization area, besides choosing for data metrics and performance indicators that the majority of the actors involved in the data quality organization will understand, it is important to focus on the needs of data consumers, both internal and external. As for information systems, system support should not be neglected as data dimensions such as integrity and security still need to be up to par. Other suggestions are data cleaning practices and prevention of loss of data, e.g. disaster management.



Roles Decision Areas	Executive Sponsor	Data Governance Council	Chief Steward	Business Data Steward	Technical Data Steward	...
Plan data quality initiatives	A	R	C	I	I	
Establish a data quality review process	I	A	R	C	C	
Define data producing processes		A	R	C	C	
Define roles and responsibilities	A	R	C	I	I	
Establish policies, procedures and standards for data quality	A	R	R	C	C	
Create a business data dictionary		A	C	C	R	
Define information systems support		I	A	C	R	
...						

R – Responsible; A – Accountable; C – Consulted; I – Informed

Figure 13 Data Governance Model Draft (Wende, 2007)

A point of criticism that is acknowledged by Wende (2007) is that the method of achieving alignment between IT governance and the organization's business goals could depend on the context of the organization, such as its size and its type of corporate structure, influencing the potential performance that IT governance could deliver. Follow-up studies have examined a contingency approach to data governance, and contingency factors identified were the size of the firm, its structure, the competitive strategy applied, the corporate governance and the style of decision-making (Wende & Otto, 2007; Weber et al., 2009). The first four are to be placed within the category of organizational placement related to the authority structure of decision-making, while the decision-making contingency factor is paired with the design parameter of the data governance model of the coordination of the decision-making authority. This way, each contingency factor contributes to the final set of guidelines and standards to be applied in data quality management of the firm. Interesting for such approach is that the contingency factors can be placed into a two-by-two matrix: the contingency factors hint for a centralized or decentralized organizational structure, and for a hierarchical or cooperative approach for management. This allows for a better design and configuration of the data quality improvement approach within a company. However, Wende & Otto (2007) do mention that validation of said contingency factors is yet to be proved. Contingency factors have been modeled by Weber et al. (2009) as a moderator variable between the design characteristics of the company-specific data governance model and its effectiveness in data quality management. Also, these contingency factors have slightly been altered, as the *degree of process harmonization* and the *degree of market regulation* have been added, making a total of seven. Through future quantitative analysis of the effect of these contingency factors on the success of data quality management, this particular governance model of data quality could be tested.

B. Advanced CFDs for Data Quality Improvement

Fan (2008) continues with a second type of conditional dependencies regarding inclusion. Again, a short example is given by the author: imagine that there is a relational schema between 'source' and 'target', with 'target' being the product that the customer wants to order. However, there are two types of products available, being books and CDs. A simplified schema is the following:

```
Source:   order (title, type, price)

Target:   book (isbn, title, price)

          CD (id, album, price)
```

One could use an inclusion dependency where the title and price as registered in the order should be included, and therefore match up, with the title and price of either a book or a CD. For this to succeed though, the 'type' of the product, as mentioned in the source schema, should be specified as follows:

Example 1: (order(title, type = 'book', price) ⊆ book(title, price)

Example 2: (order(title, type = 'CD', price) ⊆ CD(album, price)

The symbol ⊆ means that the left hand side should be a subset or equal to the right hand side of the function. Thus, in this example, for when the type has been specified as either a book or a CD, there should be a tuple where the same title and price for either a book or a CD has been found; the attributes should be correct. It might be even more complex when exceptions to the rule have to be considered, such as when a city may hold multiple area codes rather than one due to its size. Such exceptions have to be considered as extensions to a CFD, possibly in the following way as demonstrated by Fan (2008):

CFD$_1$: city ∉ [New York City] → [area code]

CFD$_2$: city ∈ [New York City] → [area code] ∈ {001, 002, 003}

In this example, it would mean that whenever the city name does not correspond with 'New York City', the value of the area code is limited to one 'true' value. In the second CFD, it is acknowledged that multiple area codes are acceptable for when the city is recognized as 'New York City'; the area code must be one of the three in the set membership (∈).

The main problem that remains is that New York City may often be referred to as 'NYC' or simply 'New York', which may be confused with the state of New York. Challenges in identifying objects or entities in the 'real world' exist when multiple data sources use different formats for representing the same thing. Setting up matching rules as a form of data dependency rule would allow for deduplication of data, better data integration and better performance when running queries. For instance, when it comes to verifying the owner of a credit card for having purchased particular items, the information on both the card and the billing should be identical:

1. card[telephone] = billing[phone] → card[address] ⇋ billing[post address]

2. card[e-mail] ⇋ billing[e-mail] → card[first name, last name] ⇋ billing[first name, last name]

3. card[last name] ⇋ billing[surname] ∧ card[address] ⇋ billing[post address] ∧ card[first name] ⇋ billing[first name] → card[owner] ⇋ billing[purchaser]

However, such matching principle will not cover the problem that values may differ in notation although it is intended they are referencing the same thing. Conclusions have to be drawn by the operator whether different notated values are closely similar or are referring to the real-world object. In the first rule given above, notice the use of both the '=' and the '⇋' symbols. It is not concluded that whether phone numbers on both the credit card and the billing are the same, the addresses are also exactly the same. Instead, it is suggested that both addresses should match despite possible formatting differences. Similar suggestion could be said for the notation of the first and last name on both credit card and billing, e.g. 'John Doe' and 'J. Doe'. Once a number of similarities have been deduced by the operator, the final conclusion of determining whether or not the owner of the card is responsible for the purchases should be made.

A challenge of such matching dependencies is, of course, the level to which the operator may deduce a similarity of data values. Fan (2008) speaks of an *edit distance* which should be determined by domain-specific operators. Thus, one of the challenges of introducing and pairing conditional and matching dependencies is the level of consistency and reasoning behind them.

In a follow-up study by Fan et al. (2009), the matching principle rules, as exemplified on the previous page, have been named relative candidate keys (RCKs) for matching tuples in a database. A new addition to this paper is the introduction of a prototype system called *SEMANDAQ*, and the explanation of *blocking* and *windowing* of tuples for matching practices. Blocking involves that certain matching keys are used to partition certain relationships between attributes in a data tuple; a comparison is then made between the tuples in the same block. With windowing, a fixed window (or vision) is given and the tuples within the window are compared. The window can then slowly slide over to a new set of tuples where consistency and matching is continuously being performed over the entire database. Of course, as Fan et al. (2009) state, the performance is dependent on what matching keys are used during such practices. The proposed system addresses the following challenges regarding improvement of data quality: first and foremost, it allows for automatic discovery of a set of CFDs which are non-redundant, therefore ensuring the reasoning behind the dependency and avoiding conflict of dependencies. Errors are detected using generated queries that should only return tuples violating the given set of CFDs of the system and are repaired with minimal difference to the original database, similar to the technique of consistent query answering proposed by Fan (2008). Finally, *SEMANDAQ* allows for non-domain specific analysis of data semantics with promises to create accurate matches compared to domain experts performing manual work. A challenge for the system would be to work on unstructured data from Web sources or when incomplete information occurs throughout the database.

Earlier work on matching rules regarding data semantics has been done by Madnick & Zhu (2006) discussing into further detail the improvement of data quality by interpreting data quality problems caused by inconsistent data semantics as misinterpretation problems. The example given of the inconsistent format of a name (John Doe or J. Doe) is one of many. The authors give an example of how a financial ratio of a company may differ per source possibly due to hidden misinterpretations of the seeker; one site may give an annual ratio while the other gives a ratio per quarter, an inconsistent error that could cause financial losses for a trader. Madnick & Zhu (2006) explain that the fault should not be in the data, since it is present, but that the data is not meeting the expectations of the receiver, since every receiver is likely to perceive real-life objects or entities in a different way. A typology of three problems with data semantics is given under the name "*Corporate Householding Problems*" by the authors. The first one is *identical entity instance identification*, and is related to the inconsistent appearance of an entity throughout the data, such as with names. Although the authors acknowledge that using a social security number as an identifier for a person would solve this issue, they may not always be present.

The second problem is *entity aggregation*: given an entity, what other entities should be included and perceived as an extension or aggregation of the original entity? To give an example, the school of Catholic Theology at Tilburg University has two locations: one in Tilburg and the other in Utrecht. Should the building located in Utrecht be recognized as an aggregation of the definition 'Tilburg University' or should it be its own separate entity? It could depend on the context; if one is curious to know which buildings are at the campus located in Tilburg, the building in Utrecht should be considered excluded from the definition of that campus. However, it should be included when one needs to know all buildings of all faculties of Tilburg University.

Finally, the third problem is described as *transparency of inter-entity relationships*: for some queries, it should be established exactly what the suggested relationship between entities is being referred to. For instance, a transaction between a buyer and a seller could involve a broker in between; therefore, the answer will differ depending on the context. For knowing how many students are enrolled in Tilburg University, answers may differ depending on the scope: are international exchange students being taken into account or is such entity relationship neglected? This issue may get even more complex when entities are broken down to faculties, divisions, departments, branches, or may be transformed due to mergers and acquisitions or joint ventures.

Madnick & Zhu (2006) stress the importance of letting the data be understood in the right context. The technology that is proposed is *COntext INterchange*, or *COIN* in short. In short, the COIN approach allows for queries performed by the user to be solved in terms of context and data semantics and scans for any semantic conflicts. It needs (a) a *domain model* which describes all the semantic types for an entity to allow for data integration, (b) an *elevation axiom* which scans all the semantic types present in a given data source and allows for any constraints to be identified given the semantics of present data, and (c) the *context definition* so that one type of interpretation is used for data semantics, e.g. ‘length’ given in ‘centimeters’ or ‘feet’. While not an easy feat, the authors conclude that focusing on data semantics to achieve improved data quality is necessary since part of the quality is derived from what the end-user can or cannot do effectively with such data. The authors propose that for future research, the examination of metadata containing expectations of end-users combined with what type of semantics have been used for the end-user queries may be useful to further develop context mediation of data.

C. Assessment & Cost Curves of Data Quality

Caballero et al. (2007) identify the risk that multiple interpretations of monitoring data quality can exist, but attempt to create a standardized one based on a ISO standard, ISO 15939 (ISO/IEC, 2000). It included a so-called *measurement information model* (MIM), which is used as the blueprint for a standardized data quality MIM. Their motivation for encouraging the introduction of such standard is to eliminate the subjective aspect of analyzing data quality; it could be different dependent on the type of stakeholder due to its role and level of usage with the data. Questions that need to be answered by the standard are the *why*, *what*, *where*, *who*, *whose*, *how* (*much/many*), and *when* questions. However, for simplicity purposes, only the *why*, *what*, *who,* and *how* questions will be given further attention. One of the challenges of converting this particular ISO standard to the realm of data quality is to match the terminology from both sides. As is shown in the table below, matching the terms by ISO 15939 with the terms used in data quality literature is important to create continuous understanding of how the data quality MIM needs to be implemented.

Table 12 Comparison of ISO & Data Quality Literature Terms for MIM (Caballero et al., 2007)

Type of Question	**ISO/IEC 15939 Terms**	**Data Quality Literature Terms**
Why	Information Need	Information Quality Assessment Objectives, data quality 'problem'
What	Measurable Attribute	Data Quality dimensions
Who	Stakeholder	Data Customer / Producer / Managers
How	Measurement Method	Quality Criteria Assessment, Assessment Scores

ISO 15939 describes the word *'measure*' as a set of operations that are conducted in order to quantitatively or categorically represent the determined value of a given attribute, in this case it could be a selected data quality dimension. However, the data quality dimensions in question should in its turn be translated to a '*measurable attributes*', and as discussed earlier, multiple researchers prefer their own set of data dimensions. Interestingly, Caballero et al. (2007) seem to agree with Wang & Strong's (1996) selection of data dimensions for general business information needs. Stakeholders need to be defined carefully too. A *data* owner is to which the data belongs to, but it is likely to assume that the data owner is the only stakeholder authorized to perform measurements on the dataset. Finally, an *´entity*' in the ISO library means in this scenario the data model or value that will be measured as a form of object.

Another item of their work that adds value of the thesis is their two-by-two matrix explaining how data can be perceived in terms of their values and the judgment of the data owner or customer. Metadata (data about data) can help to determine the data quality of the original dataset. Based from their literature review, Caballero et al. (2007) state that researchers in the data quality field are in mutual agreement that metadata should be perceived subjectively, while data pulled from the actual data stores are to be observed objectively. Several key terms are placed in the matrix: timeliness, reputation, added-value and believability. Working from top left to bottom right, timeliness is one of the dimensions given by Fan & Geerts (2012), Fisher & Kingma (2001), Liebchen & Shepperd (2008), and Wand & Wang (1996). It describes when data is pulled from a data store in order to identify whether it is up-to-date, thus creating an objective value created when transferring the data from the data store, and allows for objective judgment of the data since the timestamp is automatically generated rather than created through human input. An objective value can still be judged subjectively when another data attribute is focused on primarily. Caballero et al. (2007) give the example of a person looking for a new digital camera, with particular interest on the optical zoom characteristic. Manufacturers of digital cameras may include data about price, power, connectivity ports, etc. Data about these attributes may deliver *added-value* to the object that is the digital camera, but the observer of the data has to decide on its own whether these extra data attributes provide usable information about the digital camera itself, allowing for subjective judgment.

Objective judgment with subjective values may occur when a data customer consults the review of another person (or relevant stakeholder). This outsider may be a critic with such a recognized status that other people trust the reviews (s)he has produced as truthful. This means that the judgment of the critic is perceived as objective by the data customer. An example of a subjective value in a more understandable setting would be a movie; not everyone shares the same experience from watching a movie, meaning that there is no one particular ‘true value’ of how people should feel, think, or describe the movie, making it a subjective value. Taking this into the concept of data quality, the reputation that one may have about the data quality of a given data store may not be same for the other. *Believability* is when the critic is missing from the example given because a new product, element, or object is being introduced to the world. Data customers have no reference (e.g. a movie review) to steer their judgment towards the objective side of the matrix, thus having to deal with both subjective values paired with subjective judgment.

Table 13 Objective & Subjective Dimensions of Data Quality Assessment (Caballero et al., 2007)

	Objective Judgment	**Subjective Judgment**
Objective Values	Timeliness	Added-value
Subjective Values	Reputation	Believability

For the description of the case study approach, the term ‘baseline’ will be used, which can be compared with *base measure* given in the ISO standard library, meaning that the initial measurement is taken place in order to create a starting reference point for further improvement. *Derived measures* are done using a function using a base and possibly with other derived measures. An *indicator* uses a model and is paired with some decision criteria, such as *‘low*’, *‘acceptable*’, and *‘high*’, given the values determined from the functions included in the derived measures. For simplicity purposes, a derived measure including a function for the reliability and completeness data dimensions can be the following:

```
Completeness = 1 - #not-complete-values/#-values
Reliability = 1 - #not-reliable-values/#-values
```

Table 14 Decision Criteria Model (Caballero et al., 2007)

		Completeness	
		[0, 0.8)	[0.8, 1]
Reliability	[0, 0.6)	*Low*	*Acceptable*
	[0.6 – 1]	*Acceptable*	*High*

Of course, an organization may decide to change the minimum score requirements for both dimensions to be satisfactory (0.6 for reliability; 0.8 for completeness) and may introduce more increments in decision criteria than the three given in the previous table. Of course, Caballero et al. (2007) acknowledge that assessing a score for each data quality dimension taken into consideration for all data stored is infeasible; sampling techniques are discussed to represent the population of the data records, and are discussed using the work of Zseby (2008). Although the *when*-type of question has not been discussed, the authors do mention that the ISO standard does not specify how often measurement practices should be performed. Given the possible scenario that data is used by multiple types of stakeholders at different times, determining the frequency of data measurement should be discussed internally at the organization. This is why measurement reports about data quality should be attached to its relevant data model to allow the opportunity to recheck whether the data quality figures are still up to date.

Finally, a SWOT-analysis is made of their proposed data quality MIM, and their given weakness is that the economic impact of data quality, or the costs of measurements are neglected from their MIM, meaning that it is focused on improving measurement practices of data quality only for the sake of improving data quality rather than to inspect whether costs could be cut or economic impact of poor data quality could be mitigated by improving said data quality. Earlier work by Eppler & Helfert (2004) discusses the economic impact of poor data quality on multiple cost types, and the researchers have attempted to create a total cost graph of the investments made for data quality to improve towards the 100% mark.

Table 15 SWOT Analysis for a proposed data quality MIM (Caballero et al., 2007)

Strengths	**Weaknesses**	**Opportunities**	**Threats**
Unification of data quality measurement terms with ISO standard Can be used with multiple data models	Neglects economic impact of data quality measurement	A standardized methodology for data quality measurement practices is needed Applying case studies to this methodology would strengthen its purpose	No proven concept yet – acceptance of the terms and applicability have not been achieved yet.

Batini et al. (2007), whose future work regarding the comparison of data quality methodologies has been discussed earlier in the chapter (Batini et al., 2009), have developed their own framework based on Basel II operational risk practices. Their methodology, called ORME-DQ (based from the Italian word for tracks, *´orme*') covers four areas of risks related with data quality: *prioritization*, *identification*, *measurement*, and *monitoring*. The latter two are strongly relevant to the research topic and cover the assessment of data quality, either qualitative (through categories) or quantitative (through exact values or figures), and monitoring of chosen data quality dimensions similar to how the decision criteria model of Caballero et al. (2007) is applied, using minimum targets values to be achieved for each dimension. A risk is then identified when data quality scores below the given minimum for one or multiple dimensions. For clarification, Batini et al. (2007) assume that the data quality dimensions most frequently chosen for quantitative analysis are accuracy, completeness, currency, and consistency; all are present in the list of data quality dimensions by Fan & Geerts (2012).

An area that has not been elaborately discussed yet is a breakdown of which costs can occur with data quality monitoring and improvement practices. The second step of their ORME-DQ methodology discusses profiling of possible economic losses where values can be either exact or descriptive. When risk profiling has been completed, a selection can be made on which processes can be labeled 'critical' based by looking at the frequency of an *event loss* happening due to data quality paired with the gravity of the economic losses due to that event.

Given the four data quality dimensions used in the methodology, Batini et al. (2007) give a short explanation on how each data quality dimension can be measured. With data accuracy, the authors refer to the correctness of the data based on syntax, but use two types of syntax controls. The first one is related to typos, such as '*Jhon'* while the name '*John*' is meant. In order for such syntax errors to be detected, both the number of common characters (four in this example) and the number of transposed characters, meaning the characters that do appear but in different positions (two in this example), need to be known. However, sometimes typos can be made because of the same phonetic sound of words, such as '*die*' and '*dye*'. The authors give an extreme example with the words '*Hilbert*' and '*Heilbpr*'. Using a soundex code for similar sounding words allows the data owner or customer to be presented with possible suggestions for what the 'right' data value could be. A soundex code is four characters long, with the first character being the first letter of the word (D and H in the examples). *Hilbert* and *Heilbpr* share the same soundex code since they sound phonetically similar (H416). When doubt is given to one data value, for instance *Heilbpr* being a very unusual name, the data owner may enter the soundex code of the data value to determine whether another name was meant, in this case *Hilbert*. Using both techniques, Batini et al. (2007) cover accuracy issues causes by typos and misinterpretations.

The completeness dimension of data quality is assessed by using the Closed World Assumption (CWA) as discussed by Abiteboul et al. (2006). This means that missing values, or NULL values, only have one reason for their appearance. Normally, the CWA approach means that all attributes for a given entity are collected and stored to the database, without any other conclusions possible to be made. Incompleteness of data can therefore only occur when the value of an attribute is missing from within the database. Improvement on data currency can be stimulated by the introduction of timestamps, but as mentioned earlier, another risk could exist when the format of the timestamp is inconsistent (Dong et al., 2009). Finally, the consistency dimension is supported by using *if-else* clauses, similar to what has been described by Lee et al. (2014) regarding network management operations later on in the thesis; national law rules can be applied, such as forbidding citizens younger than 18 years old to be married. This means that a data tuple with attributes 'age' and 'marital status' cannot have the combination where an age younger than 18 has the marital status '*married*'.

A star schema is developed by the authors where each measurement is related to six other dimension tables: *metric, time*, *data quality dimension*, *process*, *organization unit*, and *data group*. Each measure needs to be associated with a metric, as is related to the earlier discussed work of Caballero et al. (2007). The 'time' dimension table is simply related to the time for when the measurement was performed to ensure the presented results are up to date. Logically, the data quality dimension table is related to the purpose of the measurement, meaning which data quality dimensions were taken into consideration. Data measurements can be done differently, thus the type of *process* needs to be specified and related through tables. It also needs to be known which *organizational unit* is going to see the measurement results of the data or is currently using the data subject to measurement. Finally, the data that needs to be measured on its quality must come from a *data group*, representing a larger data structure with relational tables.

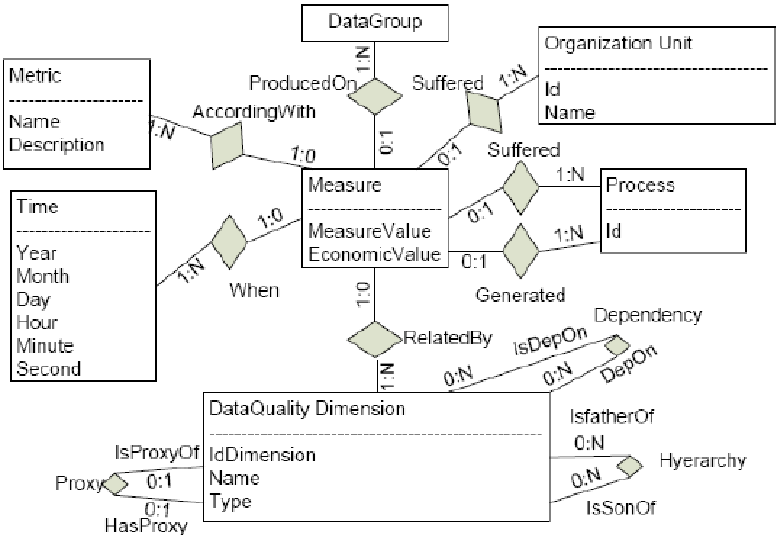


Figure 14 Star Schema of Data Quality Measurement with Relational Tables (Batini et al., 2007)

Monitoring and reporting activities can be performed on a web-based platform, according to the ORME-DQ methodology of Batini et al. (2007). Caro et al. (2006) have also performed additional research on the user interfaces intended for representing the data quality monitoring reports, simply referred to as data presentation. Their concern is that the data displayed on such presentation layers should also be measured, meaning to answer the question whether the data quality of the user interfaces, possibly through a web portal, can be measured and is up to par for the data customer.

Eppler & Helfert (2004) also discussed the costs incurred due to low data quality but go beyond the work of Batini et al. (2007) by including possible graphical plots of cost curves when data quality would rise from zero to hundred percent. A classification of costs that occur or rise due to low data quality can be found in the following figure. However, besides their approach of dividing direct and indirect costs, they also break down costs in two other categories: *preventive* and *corrective*. The difference is that that with *corrective* costs, the error has already occurred, while *preventive* costs are made to avoid such errors from happening in the future. This requires for processes to be improved upon in order to detect new types of errors that could occur due to the use of new equipment in the IT infrastructure environment.

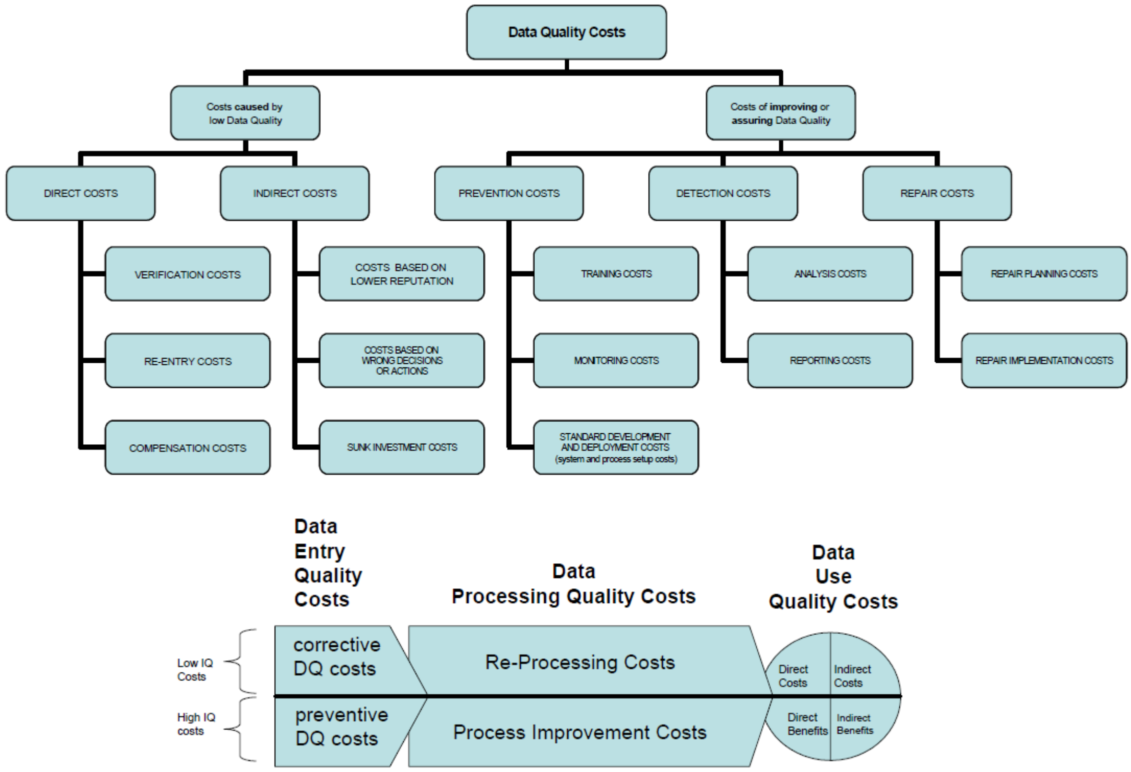


Figure 15 Simplified Taxonomy of Costs related to Data Quality (Eppler & Helfert, 2004)

Three scenarios are given by the authors for why classification of data quality could be useful: *risk assessment*, as included in the methodology by Batini et al. (2007), applying it in a business case where higher data quality is demanded over a given timeframe, and *benchmarking* the data quality with competitors in order to seek whether an advantage can be taken or maintained.

The main interest of the paper is the presented cost curves of corrective & preventive data quality costs and failure costs, as presented in the previous figure, and the correlation between the two. The general rule applied here is that when the level of investment in proving data quality increases, the financial consequences of the data quality underperforming at the new higher set standard will be larger too. At some point, the costs required for improving on data quality at the final percentiles are this large that investing is not recommended. However, when a higher level of data quality is established, failure costs will decrease due to the lower amount of effort required to correct the data. This is because the level of reference (and its quality) of what is presumed to be correct data is larger when the overall data quality increases. Eppler & Helfert (2004) describe how monitoring tools may perform better in detecting any data defect when the overall data quality is at a high level. With low levels of data quality, correcting the data becomes a more difficult challenge when a higher level of uncertainty is present when determining which data is considered correct, increasing corrective data quality costs.

Named the *Cost of Quality Model*, borrowed from Wasserman & Lindland (1996), the corrective and preventive data quality cost curves are accompanied with a new curve, called the *total quality costs curve*. This cost curve is the sum of the previous two curves and shows that the lowest amount of total quality costs within the 90-100% quality range does not have to be at the 90%-mark. Instead, as the level of quality improves starting from 90%, the total costs may decrease up till a certain point, for instance, 97%. Increasing data quality from the 97% mark is going to be difficult and costly as the prevention and corrective costs are going to skyrocket for these final percentages, with the sum of total quality costs increasing too. This means that there is an optimum point in the total quality cost curve where a higher level of data quality is achieved for the lowest cost possible; any backward or forward moment of the level of data quality given in percentages increases the total quality costs.

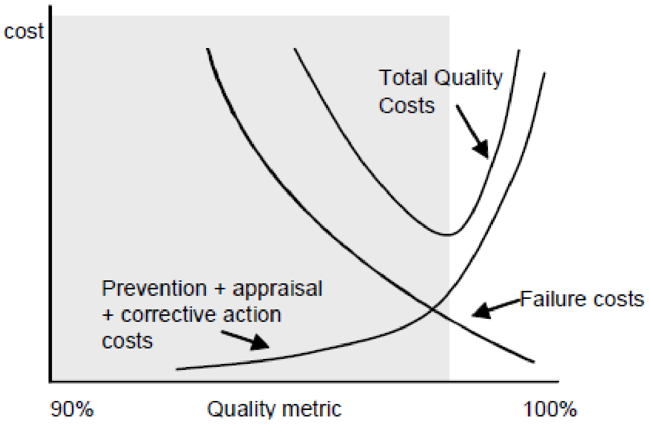


Figure 16 Cost of Quality Model (Wasserman & Lindland, 1996; Eppler & Helfert, 2004)

Eppler & Helfert (2004) ask the question though whether such cost models are actually applicable in the practice of data quality management and whether the different data quality methodologies fit with the proposed cost model. Their criticism is that quantifying the efforts made on preventing, correcting and detecting data quality errors simply on their cost curves is impractical. Of course, the optimum point for total quality costs will move when the gradient of the cost curves change, for instance when a cost saving technique is applied for preventive or corrective data quality practices. The authors state that their main contribution was the introduction of prevention measures to achieve the maximum achievable level of data quality; cutting preventive costs on higher data quality levels improves the total quality cost curve.