TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MARKUS HEISKANEN

DATA QUALITY IN A HYBRID MDM HUB

Master of Science Thesis

Examiner: prof. Samuli Pekkola
Examiner and topic approved by the Faculty Council of the Faculty of Business and Built Environment on 9.3.2016

TIIVISTELMÄ

MARKUS HEISKANEN:
Tampereen teknillinen yliopisto
Diplomityö, 72 sivua, 1 liitesivu
Toukokuu 2016
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Pääaine: Tiedonhallinta
Tarkastaja: professori Samuli Pekkola

Avainsanat: master data, datan laatu, MDM hub

Data ja sen laatu ovat merkittäviä menestystekijöitä nykyaikaisessa organisaatiossa. Ydintieto edustaa organisaation merkittävimpiä dataobjekteja. Sen heikko laatu johtaa ongelmiin liiketoimintaprosesseissa jotka johtavat kustannusten kasvuun ja liiketoimintamahdollisuuksien menettämiseen. Keskeinen ongelma on monimutkaisten liiketoimintaprosessien yhteensovittaminen monimuotoisen järjestelmäympäristön kanssa.

Tämän tutkimuksen tavoite on määrittää tärkeimmät tekijät jotka vaikuttavat ydintiedon laatuun MDM hub – kontekstissa. Tutkimus tehtiin jotta saavutettaisiin ymmärrys mainituista tekijöistä yleisellä tasolla. Täten tutkimusta ei ole rajattu yhteen tiettyyn organisaatioon. Tavoite oli löytää lista kriittisimmistä datan laatuun vaikuttavista tekijöistä MDM hybridi hubin kanssa toimittaessa.

Tutkimus on kaksiosainen ja muodostuu teoreettisesta kirjallisuuskatsauksesta sekä empiirisestä haastattelusta. Kirjallisuuskatsauksessa alan keskeisestä tutkimuksesta muodostettiin tiivis kokonaisuus jonka tehtävä oli tukea tutkimuksen empiiristä osuutta. Empiirisessä osuudessa haastateltiin ydintiedon hallinnan ammattilaisia. Tulokset analysoitiin ja niitä verrattiin teoriaan.

Merkittävimmät löydetyt tekijät olivat ihmiset ja heidän vastuunsa ja roolinsa, datan laadun hallinnointi korkealla tasolla liiketoimintaprosesseja tukevien prosessien muodostamiseksi, datan laadun hallinnan virtaviivaistaminen ja datan laadun arviointi ja parantaminen sopivien työkalujen ja automaation avulla.

Tulokset esittävät että MDM hybridi hubi tukee datan korkeaa laatua tarjoamalla työkalut keskeisempien tekijöiden huomioimiseen. Nämä työkalut helpottavat roolien ja vastuiden määrittämiseen ja mahdollistavat työnkulut jotka tukevat datan laadun hallinnallisia prosesseja. Se tarjoaa myös työkalut metadatan ja data sanakirjan hallintaan sekä datan laadun arviointiin sekä sen hallinnan automatisointiin.

ABSTRACT

MARKUS HEISKANEN:
Tampere University of Technology
Master of Science Thesis, 72 pages, 1 Appendix page
May 2016
Master's Degree Programme in Information and Knowledge Management
Major: Business Information Management
Examiner: Professor Samuli Pekkola

Keywords: master data, data quality, MDM hub

Data and its quality play a large role in the success of a modern organization. Master data represents the most important data objects of an organisation. Its poor quality leads to problems in the business processes which lead to overhead and loss of business. The core problem is the alignment of complex business processes to information processes in complex system environments.

The goal of this study is to determine the most important factors affecting the master data quality in a specified context that is the context of MDM hybrid hub. The research was done to reach the understanding of the mentioned subjects in the general level. It was not restricted to one specific organization. The aim was to find a list of most critical factors to data quality that need to be assessed when working with MDM hybrid hub.

The research had two parts which were the theoretical literature review and an empirical assessment in the form of an interview. In the literature review the relevant research was assessed and summarized to support the empirical part of the research. In the empirical part a multitude of professionals of the area of MDM were interviewed and the results were analyzed and reflected with the theory.

The most important factors found were the people in the form of responsibilities and roles, the data quality governance which helps forming processes to support the business processes, the streamlining the data quality management and assessment with data quality tools and automation.

The result also shows how MDM hybrid hub supports the high quality of data by addressing the factors with relevant tools. These tools help in the assignation of roles and responsibilities. It enables the related workflows which support the data quality process. It also gives tools for metadata and data dictionary management and offers tools for assessing data quality and automating its management.

PREFACE

I began the thesis process in summer 2015. After nine months it is finally finished. This thesis is the final product of a long study journey and also embodies the transformation from a student to a business professional.

I would like to thank my supervising professor Samuli Pekkola for the advice and support on this research process. I would also like to thank my colleagues and my employer for the insights and resources that made it possible to conduct this research. Most of all I would like to thank my family and friends who have supported me through the whole journey in both academic and personal sense.

Tampere, 25th of March 2016

Markus Heiskanen

TABLE OF CONTENTS

APPENDIX A: THE INTERVIEW THEMES AND QUESTIONS

LIST OF SYMBOLS AND ABBREVIATIONS

CRM	Customer Relationship Management system
CRUD	Create, Read, Update, Delete
DQM	Data Quality Management
DQMS	Data Quality Management Services
ESB	Enterprise Services Bus
ERP	Enterprise Resource Planning system
ETL	Extract, Transform and Load
ICT	Information and Communications Technology
LOB	Line of Business
MDEMS	Master Data Event Management Services

MDM	Master Data Management
MDS	SQL Server Master Data Services
OLAP	Online Analytical Processing
SOA	Service Oriented Architecture
SQL	Structured Query Language
UI	User Interface
XML	Extensible Markup Language

1. INTRODUCTION

Data is a vital resource for companies and those who invest in it do stand a stronger chance to success than those who neglect it (Eckerson 2002, p.3). Companies tend to be more and more information intensive nowadays and use more and more data in their everyday operations. Companies have more data in their databases than they know what to do with (Scarisbrick-Hauser 2007, p.161). They have increasingly invested in technology to collect, store and process vast quantities of data but still often find themselves unsuccessful in the efforts to translate this data to meaningful insights. (Madnick et al 2009, p.3) Thus data quality and problems related to it are more and more relevant (Redman 1998, p.80; Wand and Wang 1996, p.86-87).

Most companies experience data quality problems at some level (Huang et al. 1998, p.92). Even though data quality is crucial to company's success, it is still often left without proper attention (Eckerson 2002, p.3; Marsh 2005, p.105; Xu et al. 2002, p.47). Data quality problems cost 600 billion dollars a year in U.S alone (Eckerson 2002, p.3).

Master data describes the most important business entities of a company, such as customers and products (Loshin 2009 p.6; Haug et al. 2011, p.288). In many companies this master data is kept in many overlapping systems and its quality is often unknown. This situations leads to a dilemma where it is difficult for the organizations to implement change. Architectural approaches such as Service-Oriented Architecture (SOA) are difficult to implement when an organization lacks common definition and management of its core information (Dreibelbis et al. 2008, p.1). One of the most common reasons for implementing a MDM hub is to provide clean and consistent data to support a SOA implementation (Wolter 2007). That is why it is safe to say that master data quality is one of the most important context of data quality.

The way MDM is modeled and implemented has a great effect how well MDM efforts succeed (Dreibelbis et al. 2008; Allen and Cervo 2015). The robustness and customizability of the model are very important, and the solution on which the modeling is done greatly effects how well the model can server its purpose. (Dreibelbis et al. 2008; Allen and Cervo 2015).

1.1 Research objectives and scope

The central objective of this research is to determine the key factors in order to maintain the data quality in MDM hybrid hub –based architecture. First, theoretical foundation of all the elements of the research question are introduced. These form the relevant supporting research questions that help to answer the main question from all the relevant angles. These include defining the concepts of “master data” and “data quality”. Also the definition of “MDM hub” is crucial to understand the viewpoint of this study.

When shaped as research questions the supporting research goals are to determine:

- Which data of organization is really master data and how is it managed?
- What is data quality from master data perspective?
- What are the key concepts of hybrid MDM hub –architecture?
- What are the roles of master data quality management?

The main research question is:

- What are the key factors in supporting data quality in hybrid MDM hub?

The first four questions define the terminology and viewpoint used in this research so the reader can fluently understand the main research objectives and concepts behind that. The latter two questions intersect the theory behind the research and the real world context where many stakeholders take part in the master data process.

Information system environments are complex with multiple operative and legacy systems. Architectures are very vast and have various technologies and modeling philosophies utilized in them. They are born with time and are expanded as more needs arise. That’s why the motivation is to focus to the center of the enterprise information architecture, the master data management hub.

The more concrete and situational motivation comes from everyday needs of working with evolving information architecture. There are more and more needs for the MDM system to support the various other systems, applications and processes, and ultimately the business. As the count and variety of such systems grow, the effects of data quality become more and more critical. At the same time the ability to assign resources to the manual improvement is limited and the manual work with data assets consumes resources from other important project work. That is why there needs to be focus on the data quality improvement. It all culminates into the optimization of the usage of resources.

In summary, the scope of the research is in describing the concepts that define data quality and its management and also define MDM hub. After that the current reality of the data quality management in the MDM hub is described from view point of this study. Finally the reflection is made from the reality to the theory in the hopes of finding ways to utilize the theory to determine the most important factors for mastering the data quality in the reality of this case.

1.2 Research methodology

This thesis being a scientific study, it is foremost crucial to introduce and depict the methodologies behind it. Methodology states the theory *how* research is undertaken (Saunders et al. 2011, p.3). In contrast “method” refers to the techniques and procedures to obtain and analyze data (Saunders et al. 2011, p.3). The decision of methodologies is not trivial, Hirsjärvi et al. (2004) note that the possibilities on the choices behind the research are endless.

A good way to represent the hierarchic model under which the research is defined is the “research onion” introduced by Saunders et al. (2011, p.108).

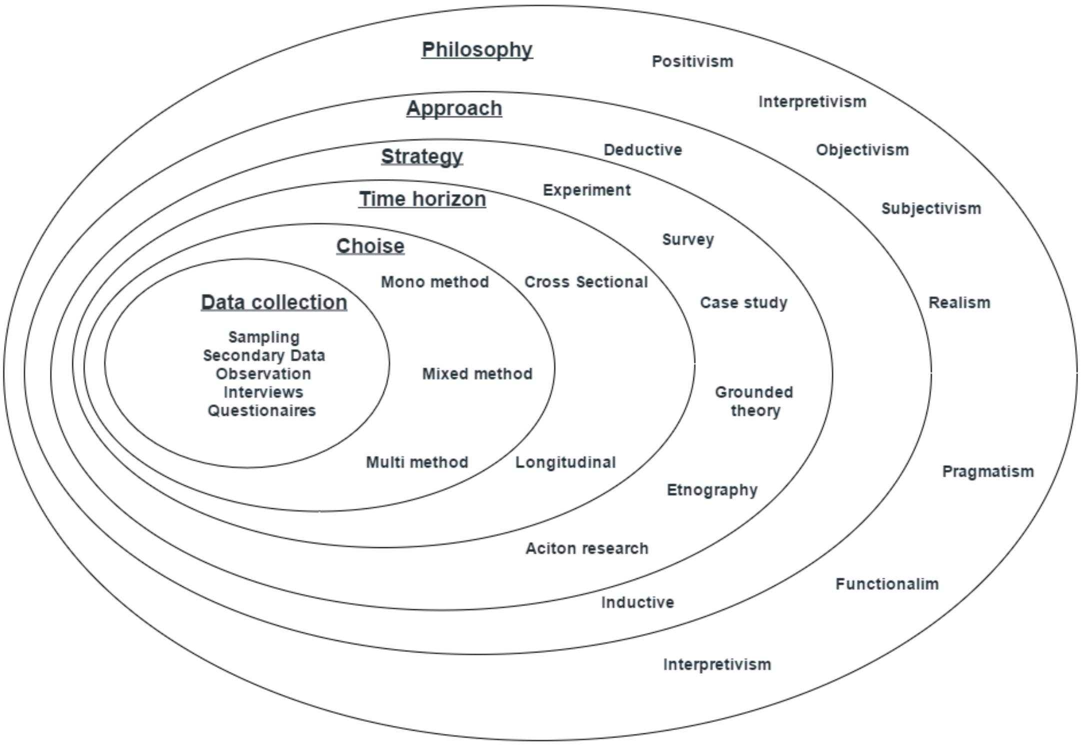


Figure 1. *The Research onion adapted from Saunders et al. (2011, p.108).*

Peeling through the onion helps to do a well formed overview of the research methodology and understand the motivation behind chosen methods. The focus is on the chosen methods and explaining why they were chosen, not why others weren’t.

In this subchapter the research onion is peeled moving from the philosophical choices to the more concrete choices. First the research philosophy is defined, then the research approaches, strategies, choices and lastly the data collection techniques. In this subchapter the data collection mostly refers to the theoretical part of the study. The empirical data collection is discussed in later chapter of the case study.

1.2.1 Research philosophy

The research philosophy contains the assumptions of the way in which the researcher views the world in the research (Saunders et al. 2011, p.108). In the area of business and management the philosophy is crucial in understanding what is concerned in the research. Same study can be performed by concerning facts as well as it can be performed by concerning the feelings of the stakeholders involved (Saunders et al. 2011, pp.108-109). The basis of this study is on facts since it is more aligned with the background of the researcher and the end-product of the study, a factual representation of actions to resolve the underlying issues.

From the ontological point of view the research must determine the view of nature of the reality being observed. Positivism is the view that the reality is external, objective and independent of social actors. In the other hand the reality can be viewed as a socially constructed, subjective, changing and not same for all. Realism takes the reality to be same for every observer but states that the interpretation may change through social conditioning. Lastly the pragmatism philosophy takes the focus on answering the research question at hand and so accepting that the reality is external and may also be multiple. (Saunders et al. 2011, p.119)

Taking consideration that the purpose of this study is to offer background for finding ways to improve complex business processes pragmatically, the natural step is to work with pragmatism.

From the viewpoint of epistemology defining the researchers view on what constitutes as acceptable knowledge, the pragmatism fits well. As it focuses on practical applied research and integrating different perspectives to interpret the data it offers a wide range of tools and freedom to work towards answering the research questions. (Saunders et al. 2011, p.119)

Axiology defines the role of values in the research. In pragmatism the values play large role in defining the results since the researcher adopts both objective and subjective points of view. (Saunders et al. 2011, p.119)

As the researcher in this case is a subjective actor working with the everyday challenges in the field of master data management in the organization the values tend to be subjective even when working towards maximal objectivity.

From the point of view of data collection techniques, pragmatism offers the possibility to use mixed or multiple methods that can be either quantitative or qualitative (Saunders et al. 2011, p.119). This fits the goal of the study well, since qualitative data is the central focus in a complex environment where the phenomena’s are intertwined and multifaceted. In the other hand quantitative data offers something very tangible which can be effective in communicating the findings of the study.

1.2.2 Research approach, purpose and strategy

The purpose of research approach is to help determining the design for the research project. There are two main research approaches; deduction and induction. Simply put, deduction can be viewed as testing a theory and induction building one. (Saunders et al. 2011, p.124)

As both of the approaches can be clearly defined to differ from another, it is not crucial to be able to pick only one of these. As using both is possible, it is still important to underline which of these is used and when. As deductive research has emphasis on already existing theories, it will be quicker and more straightforward to implement in a research. The focus is in explaining causal relationships between variables using a coherent collection of quantitative data and a highly structured approach. Deduction is stricter in that sense. (Saunders et al. 2011, p.125)

Induction has focus on gaining the understanding of the meanings human attach to events. It concerns more on collecting the qualitative data and offers more flexible structure of research and gives the possibility to do changes as the research progresses. It can also be viewed more pragmatic in the sense that there is less concern with the need of generalizing the results. (Saunders et al. 2011, p.125)

As there are intertwined and complex phenomena behind the research questions there is clear need for flexibility. And as the study reflects the collected theory to a real life case which is observed by the researcher, the focus will be more on collecting qualitative data. Since there still are quantitative elements involved it is hard to underline this research to be just induction. So it can be stated that the study has inductive emphasis with deductive elements.

Before introducing the research strategy, it is important to underline the purpose of the research. Saunders et al. (2011, p.138) introduces three different strategies to execute a research; exploratory study, descriptive study and explanatory study. Exploratory study is about finding out “what is happening” and so to shed new light to the phenomena (Robson 2002, p. 59). Descriptive study portrays an accurate profile of persons, events or situations (Robson 2002, p. 59). It is important to have a piece of explanatory research as a forerunner to this kind of research (Saunders et al. 2011, p.140). The emphasis of explanatory research is to study a situation in order to explain the relationships between variables (Saunders et al. 2011, p.125).

From the point of view of this research it is important to explain the foundation on which the empirical part of the study is based on. From the empirical point of view, the main concern is to portray an accurate profile of the situation or the environment where operations take place. From that point of view the study can most confidently be determined as descriptive.

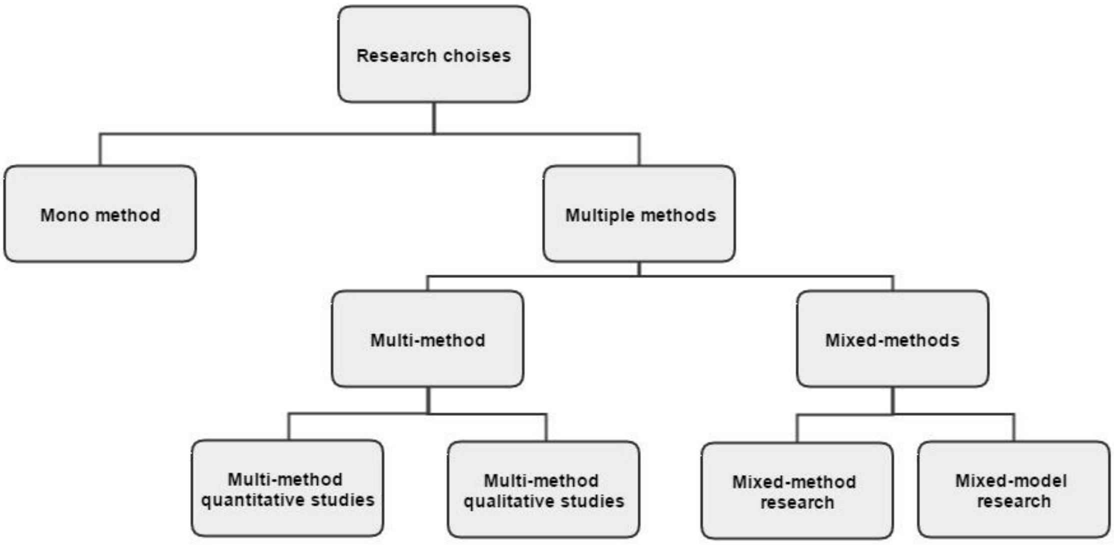


Figure 2. *Research choices adapted from Saunders et al. (2011, p.152).*

There are a multitude of strategies how to perform a research. Case study is a strategy for doing research which involves an empirical investigation of a contemporary phenomenon in real life context while taking advantage of multiple sources of evidence (Robson 2002, p. 178).

The case study strategy can be portrayed as single or multiple case with and with holistic or embedded viewpoint. Single case is most common for students who work in an organization for which the case can be based on. Holistic and embedded refer to the unit of analysis determine the level on which the organization is concerned. Holistic refers to an organization as a single entity whereas embedded views the organization as a number of logical sub-units. (Yin 1994, p.38-39)

This study perceives a company as and single entity it can be perceived as a holistic case study. It is notable that the company is not addressed in identifying fashion, but the goal is to supply answers that are relevant in the general level.

1.2.3 Data collection and analysis

In order to get the desired information for the research, the data collection method needs to be defined. Before being able to conduct research in specific field, it is necessary to understand the previous research in that field. Saunders et al. (2009, p 98) state that the best way to achieve this is to conduct a literature review where the previous research is critically referenced and the most important findings are pointed out in readable and logical way.

The first part of this research is a literature review where the most relevant and trustworthy sources in the field of data quality and master data management area discussed. The architectural point of view of MDM is also discussed. As the amount of scientific research in the areas of MDM and its architectures is scarce, much of the literature is based on the practitioners' views conducted from the most respective books in the field of study.

Berg (2004, pp.4-5) notes that the when multiple lines of research are referred to a more substantive view of reality and the concepts related to them are achieved.

The data collection and other choices of the empirical part of the study are discussed more closely in the chapter five.

1.3 Research structure

The thesis is structured in the way that the reader builds his background knowledge on the theory of all the parts of the study.

In the first chapter, the thesis goals are presented, which are represented by the research questions, are introduced. In the three following chapters the theoretical backbone of the study is formed by giving the reader a deeper understanding on the matters behind the main research question. That happens by answering the supporting research questions.

The second and third chapters are about master data. They answer the questions on; "which data of organization is really master data and how is it managed?", "What is data quality from master data perspective?" and "How is master data quality maintained?" These two chapters are based on separate lines of research and so they are their individual entities in this research, but they intertwine around the master data and by that answer these questions together.

The fourth chapter is about master data management architectures. It briefly describes the ways master data management architectures can be categorized and which are the elements that help distinguish one architecture from another. The focus of the chapter is in describing the architecture that the case study is based on, which is called hybrid MDM architecture. The research question answered is as follows "What are the key concepts of hybrid MDM hub –architecture?” In addition to the theory of the MDM architecture, the technology of the case environment is introduced.

Fifth chapter explains how the empirical study is conducted and describes the background of the case.

Sixth chapter discusses the reality of master data quality maintenance, its cost and the tasks made in practice to make the quality better. It also discusses which adequate data quality is and how to achieve it.

Seventh and final chapter concludes the research. It summarizes the results and guides for possible further research.

2. MASTER DATA

Data itself can be perceived as an end product by itself, what company uses and consumes (Wang 1998). Data should not be perceived as a by-product, because that leads to focusing in the systems and not the real end-product, the information (Lee et al. 2006, p. 125).

Data can be divided in many types and one of them is master data. The classification of the rest depends on the source. The classifications can be transaction and inventory data (Otto & Schmidt 2010, p3) or metadata, reference data, transactional data and historical data (Dreibelbis et al. 2008, p.35). Watson and Schneider (1999, p.18) found four data types in their research which are master data, transactional data, configuration data and control data. Ramaswamy (2007, pp. 1-2) finds also four data types that are similarly master data, transactional data and configuration data which can be divided to control data and less prominent version of master data, sub-master data. All of the definitions agree that master data and transactional data are the most consistently noticed data types.

Table 1. *Key data characteristics (Adapted from Dreibelbis 2008, p. 35 and Ramaswamy 2007, p.1)*

	What Kind of Information?	Examples	How Is It Used?	How Is It Managed?
Metadata	Descriptive information	XML schemas, database catalogs, Data lineage information Impact analysis Data Quality	Wide variety of uses in tooling and runtimes	Metadata repositories, by tools, within runtimes
Reference Data	Commonly used values	State codes, country codes accounting codes	Consistent domain of values for common objects	Multiple strategies
Master Data	Key business objects used across an organization	Customer data Product definitions	Collaborative, Operational, and Analytical usages	Master Data Management System
Transactional Data	Detailed information about individual business transactions	Sales receipts, invoices, inventory data	Operational transactions in applications such as ERP or Point of Sales	Managed by application systems
Historical Data	Historical information about both business transactions and master data	Data warehouses, Data Marts, OLAP systems	Used for analysis, planning, and decision making	Managed by information integration and analytical tools
Configuration Data	Describes the busienss processes in an operational system such as ERP	Process from invoice to delivery	Determines the procss flow in applications such as ERP	In ERP system based on business needs

Master data represents unified set of business objects and data-attributes that are agreed on and shared across the organization (White et al. 2006, p.2; Dreibelbis et al. 2008, p.35).These are commonly recognized concepts that are the focus of business processes, such as customers, vendors, suppliers and products (Loshin 2010, p.6). This data can be seen as one of key assets of a company and it’s not unusual that company is acquired primarily to access its master data. (Wolter & Haselden 2006, p.2). Knolmayer & Röthlin (2006, p. 363) describe master data being, once created, largely used and rarely changed.

Transaction or Transactional data is detailed information of individual business transactions like invoices used in operational applications such as ERP. (Dreibelbis et al. 2008, p.36). It is gathered and used in daily operations in organization (Davenport et al. 2001, p.3). It is highly dynamic and the most common examples are invoices and billing documents which are related to sales and purchase orders (Meszaros & Aston 2007, p.3). Davenport et al. (2001, p.3) also note that transactional data can be enriched and turned into knowledge which would lead to business results.

Historical data is transaction data enriched with master data to form a view of historical events used to analyzing, planning and decision making. This can be basic reporting or dashboards showing customized view of the company's state for the user. This data is stored in data warehouses and published via data marts and OLAP (Online Analytical Processing) systems for business intelligence purposes. Historical data is also required from legislation point of view allowing the company to meet regulations and standards. (Dreibelbis et al. 2008, pp. 35-36). Davenport et al (2001, p.3)

Reference data is commonly used data in a specific domain such as US state codes or accounting codes in a particular company (Dreibelbis et al. 2008, p.36). Reference data is often stored close to master data since many master data entities rely on reference data.

Metadata is data about data, descriptive information of data itself. For example metadata can be information of data quality or data lineage. Metadata is managed within metadata repositories and by metadata tools. (Dreibelbis et al. 2008, p.36)

2.1 Master data management

Master data management (MDM) is a collection of best data management practices that support the use of high quality data (Loshin 2010, pp.9). Berson & Dubov (2009) expand the concept of master data management and state that it's a framework of processes and technologies and its goal is to create and maintain a suitable data environment. White (2006) notes that MDM is a workflow-driven process in which business and IT work together to cleanse, harmonize, publish and protect the information assets that need to be shared across the organization.

MDM incorporates business applications, information management methods and data management tools in order to implement procedures, policies and infrastructures that support capture integration and use of timely, consistent and complete master data. (Loshin 2010, pp.8-9). The goal is to end the debate about whose data is right and whose data should be used in decision making.

MDM and its establishment into organization can be seen as a stepwise process. Many authors and researchers have discussed steps to take. Joshi (2007) has had the widely cited approach on which Vilminko-Heikkinen and Pekkola (2013) add from other sources.

Vilminko-Heikkinen and Pekkola suggest eight steps that should be followed in order to establish MDM successfully.

Step 1: Identifying the need for MDM

Step 2: Identifying the organization's core data and processes that use it

Step 3: Defining the governance

Step 4: Defining the needed maintenance processes

Step 5: Defining data standards

Step 6: Defining metrics for MDM

Step 7: Planning an architecture model for MDM

Step 8: Planning training and communication

Step 9: Forming a road-map for MDM development

Step 10: Defining MDM applications characteristics

This list is very comprehensive. It has the same elements listed as Loshin (2010, p.9) but has an even wider organizational perspective. In this thesis almost all of these are steps or aspects are noted and some discussed in deeper level. The motivation behind MDM is introduced, means to identify the core data are discussed and governance is defined in general level. Maintenance processes are referred to, but not discussed in detail. Data standards are seen as an important factor and examples of the metrics are introduced. Architecture is also covered from the MDM hub point of view. The training, road map are left out of scope whereas MDM applications especially relating to data quality are discussed.

The benefit of establishing MDM is to enable core strategic and operational processes succeed better. MDM itself is not an end objective but it offers means for systems like CRM or ERP to succeed in what they are planned to do. It helps breaking the operational silos. This supporting role leads to the fact that it is hard for senior management to give MDM the needed embrace in order to succeed. Even though it enables significant benefits in traditional business developing such as productivity improvement, risk management and cost reduction. (Loshin 2010, pp.8-11; White 2006, p.5).

Loshin (2010, pp.11-14) lists tangible benefits of MDM of which Smith & Keen (2008, p. 68-69) agree on. Comprehensive customer knowledge is when all customer records are consolidated in same repository enabling a full 360 degree view of the customer. This enables improved customer service via meeting customer expectations better in terms of availability, accuracy and responsiveness to their orders. (Loshin 2010, pp.11)

Unified and harmonized data enables a consistent and unified view to the state of the company which is important when making business decisions based on reporting. (Loshin 2010, p.11, Fisher 2007). Reports are highly dependent on master data which underlines its significance. Aside from reports, the consistency provided by MDM adds to the trustworthiness of data which enables faster decision making. (Loshin 2010, p.11; Smith & Keen 2008, p. 68) Unified data achieved by MDM adds to better competitiveness via offering a better basis for growth by simplification of integration to new systems. This straightforwardly improves the agility via reducing the complexity of data integration. (Loshin 2010, pp.10-12)

Trustworthiness of financial data is crucial for managing enterprise risks. This is most important when there are lot of data with low degree of granularity which leads to greater potential for duplication, inconsistencies and missing information (Loshin 2010, pp.10-12). Trust in the data is also crucial for the user acceptance of any initiative based on such data (Friedman et al 2006). Unified view also enables the organization to reduce operating costs by minimizing the replication of data which logically means replication of same routines which cost and also by simplifying the underlining processes (Loshin 2010, pp.10-12, Smith & Keen 2008, p. 68). From the point of view of spend analysis and planning, can product, vendor and supplier data help predict future spend and improve vendor and supplier management.

From legislative point of view MDM tends to be more and more important as regulations concerning MDM entities tend to increase, for example the privacy laws or personal data acts in Finland and European Union. From compliance point of view MDM plays big role with regulations such as Sarbanes-Oxley and Basel II to offer improved transparency to mitigate the risks involved in big and complex financial actors. (Cervo & Allen 2011, pp.144-145)

Metadata plays important role in representing the metrics on which information quality is relied on. Standardized models, value domains and business rules help to monitor and manage the conformity of information which reduces scrap and rework. Standardized view of the information assets also reduce the delays associated with data extraction and transformation which speeds up application migration and modernization projects as well as data warehouse and data mart construction. (Loshin 2010, pp.10-12)

Master data helps organizations to get understanding how the same data objects are represented, manipulated, or exchanged across applications within the enterprise and how they relate to business process workflows. The standardization must go beyond syntax to common understanding of the underlying semantics and context. This understanding gives enterprise architects a vision of how effective organization is in exploiting information assets to automate and streamline its processes. From Service Oriented Architecture (SOA) point of view, consolidated master data repository can offer a single functional service for data entry. For example, instead of creating same products in different systems, it is possible to create them to the MDM system which allows other system to subscribe to that data which simplifies application development. (Loshin 2010, pp.10-12; White 2006, p.4).

As MDM offers clear advantages and improves the organizations ability to benefit of business prospects, it does not come without challenges. Numerous technologies have tried to address the same problems MDM is concerned with. They have not succeeded so it is no surprise that MDM is under the same criticism. These technologies have been traditionally adopted with IT-driven approach while presuming them to be usable from out of the box. In addition, the lack of enterprise integration and limited business acceptance have lead such implementations to fail. (Loshin 2010, p.15).

Resolving the pointed issues in implementing a successful MDM program, it needs to start from the organizational preparedness and commitment. There needs to be technical infrastructure for collaboration around the MDM and the enterprise acceptance and integration should reach all ends of the enterprise. This means that the organization should be committed to an enterprise information architecture initiative. In addition, the data quality needs to be high and it needs to be able to be measured in order for the benefits to be clear. All of these are wrapped under overseeing these processes via data governance procedures and policies. (Loshin 2010, p.15; White 2006, p.2-4).

2.2 Data Governance

Fisher (2007) and Wailgum (2006) state that ultimately the MDM is a political and consensus building effort for the stakeholders to agree on common definitions and key data items they use.

Term “Data governance” can be perceived in a multitude ways. Khatri & Brown (2010, p.148) distinguishes governance as referring to the decisions which need to be made to ensure effective management and use of IT and who makes the decisions. In contrast, management involves making and implementing these decisions. For example, governance establishes the information who holds the rights in determining data quality standards whereas management involves determining the actual data quality metrics.

Loshin (2010, p. 68) sums data governance as being a collection of information policies that reflect business needs and expectations, and at the same time the process of monitoring conformance to those policies. Whether the discussion is about data sensitivity or financial reporting, each aspect of business can be seen from the viewpoint of meeting specific business policy requirements. These policies rely on enterprise data and so each of them define a set of information usage policies. Information policies represent a multitude of data rules and constraints associated with the defining, formatting and usage of underlying data elements. Qualitative guidelines on the quality and consistency of the data values and records represents the very basic level of data governance. This creates the basis for business metadata represent the factors needed to meet the conformity of business policies. (Loshin 2010, p. 68)

In order to be able to create foundation to effective data governance, there are requirements that need to be met. The information architecture needs to be clear, information functions need to be mapped to business objectives and there needs to be a process framework based on information policies (Loshin 2010, p. 70). Khatri & Brown (2010, p.148) agree with this stating that there needs to be a clear view of IT and data architecture, effective linking of data principles to the business and processes that ensure consistent governance implementation in the whole enterprise.

2.3 Identifying master data and metadata

Before determining how to manage the master data, more fundamental questions needs to be answered regarding master data itself. Loshin (2010, p.130) offers few questions to support the identification:

- Which business process objects can be considered as master data?
- Which data elements are associated with each of the master data objects?
- Which data sets would contribute to the master data?
- How to locate and isolate master data objects?
- Hot to standardize the different representations of data?
- Hot to asses differences between representations?
- How to consolidate standardized representations to a single view?

The company may have multitude of application architectures. That’s why master data objects may be represented very differently. One system may store customer first, middle and last names distinctly whereas other may have them in the same field. In order for data to be potential master data, it needs to have the means for consolidation and integration. (Loshin 2010, pp. 131-134)

This identification can be supported by using data profiling techniques such as frequency distribution and primary and foreign key evaluation. Every source needs to be evaluated independently with support from both IT and business. Data objects that are not populated at all or are very scarcely populated normally are not identified as master data. Assessing the difference between representations need to be supported by deep understanding of the business processes. (Allen & Cervo 2015)

Loshin (2010, pp. 131-134) states that master data can be identified bottom up or top down. When identifying master data bottom up, the key is to determine data structures, entities and objects that are already in use in the organization and can be identified as master data and can be resolved to fit the proposed master data environment. The top down approach seeks to identify the master data from business process perspective. The key here, is to find business concepts which are shared across business processes in the organization as well as are aligned with the strategic imperatives of the organization.

Allen & Cervo (2015) state that it is relatively easy to start identifying master data by recognizing clear domains in data and processes that the organization use to operate. Master data management domain refers to a data domain where master data initiatives focus. Customer, product and employee are some of the most universally targeted domains and thus are a logical starting point for MDM. These domains can vary greatly between organizations and for example domain for an educational organization would include students and faculty whereas for manufacturing organizations it would include items, products and materials. These domains can also be influenced by system architecture if it includes applications with predefined data domains.

Before the single master record can be materialized, there needs to be a way to manage the key data entity instances distributed across the application environment. This boils down to managing the master metadata. In order to do that, the metadata must be identified. Loshin (2010, p.136) offers six steps to determine the elements which help determine the master data. First discover data resources containing entity information. Then determine which of those is the authoritative source for each attribute. Third, understand which of the entity's attributes have identifying information. Then extract identifying information from the data resource and transform the identifying information to a standardized form. Lastly establish similarity to other standardized records

This is a process of cataloging data sets, their attribute, formats, data domains, contexts, definitions and semantics. The goal is to determine the boundaries and rules which help automating the master data consolidation and governing the application interactions with MDM system. The metadata should resolute the syntax or the format of the element, structure of the instance of elements, and semantics of the whole entity. Loshin (2010, p.136)

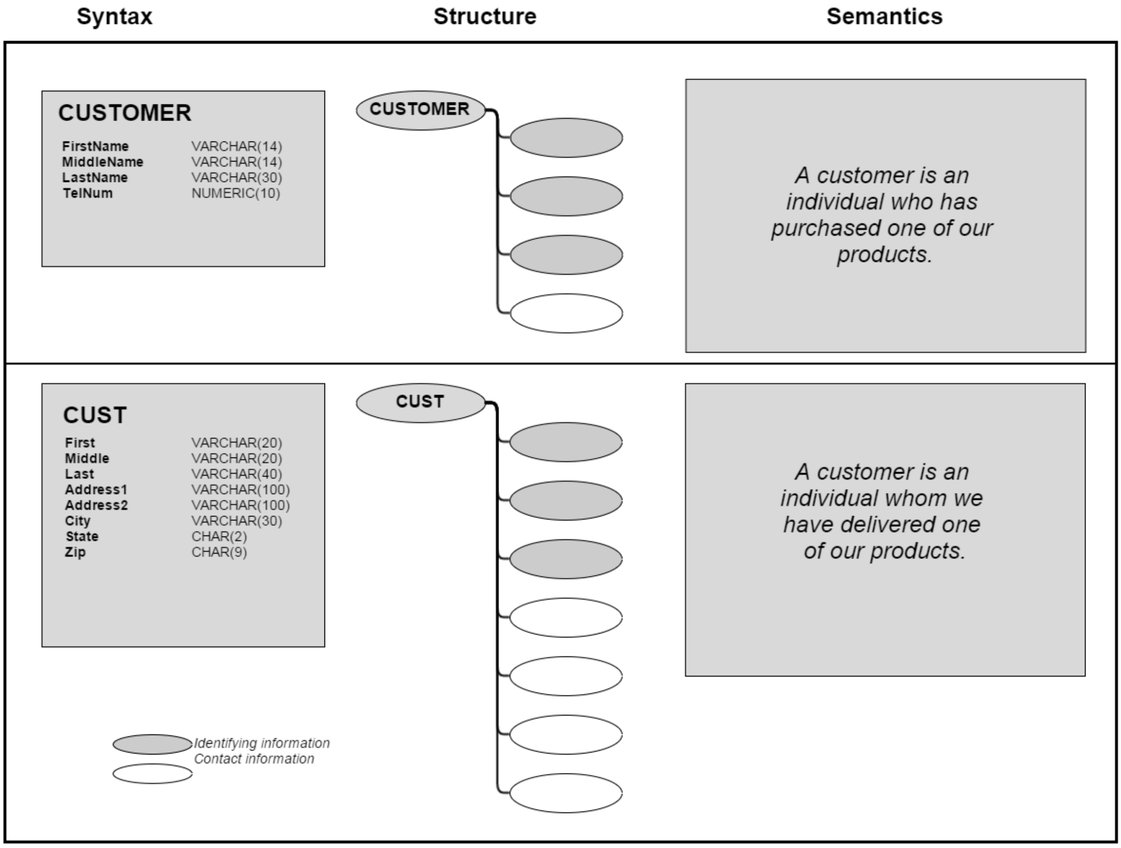


Figure 3. *Syntax, structure and semantics (Adapted from Loshin 2010, p. 136)*

For example, the way customer name is represented is the format level, the attributes of the customer make the structure and the business definition describes the semantics. Understanding the semantic difference prevents errors. Loshin (2010, p.136)

Allen & Cervo (2015) present steps to catalogue the metadata of every specific domain. First data models need to be documented in logical, physical and conceptual level. This documents also the business concepts, data entities and elements and their relationships. Second, a data dictionary listing the data elements, definitions and other metadata needs to be associated with the data model. A functional architecture needs also to be documented depicting how systems and processes interact. For specific data elements, source to target mapping needs to be made between source and target system. Documenting data life cycle helps to depict the flow of data across application and process areas from the creating to retirement. CRUD (Create, Read, Update, Delete) analysis indicates the assignation of permissions to form various groups and types of data.

2.4 Data responsibilities, ownership and accountability

Processes creating data are very similar to processes creating physical products (Wang, 1998, p.59; Lee et al. 2006 p.125). These processes have similar phases, such as collection, warehousing and usage. (Huang et al., 1998, p.91; Strong et al. 1997, p.104). There are similarly roles in data processes. Wang (1998, p.60) presents four roles related to data process, the supplier, the consumer, the manufacturer and the manager or owner. Lee & Strong (2003) introduce three roles that are data collector who gathers the data, data custodian who stores and maintains the data and the data consumer who accesses data, uses data and consolidates the data.

As master data objects are an enterprise resource and the processes to relate to them are similar to any process the role of the ownership becomes crucial (Loshin 2010, p. 75). Owner is responsible for the whole data process and his/hers responsibility is the usability and the quality of the data (Wang 1998, p.60).

A key challenge is to identify a primary business owner for each data object (Smith & Keen 2008, p. 69) The problem in such endeavor is that individuals may feel threatened when stripped the responsibilities and control to a data objects close to them (Loshin (2010, p. 75). Berson & Dubov (2007) and Ballou & Tayi (1989, s.320) both find the owner as a person who has enough authority in the organization to create, access and manage the data. That gives them a natural incentive to take care of the quality of the data. Berson & Dubov (2007) also state that the owner should rather be from business than IT. Hodkiewicz et al. (2006, p.10) note that the term data owner can be problematic if it leads to other stakeholders neglecting the quality and putting all the responsibility to the owner.

Perception of ownership can also be very different in different parts of a large organization. This perception may be based on their own information architecture consisting of applications that are not attached to the enterprise as whole. When centralizing master data, these kinds of traditions must be broken and different lines of businesses must conform to the centralization of data and so accompanying the data governance policies. (Loshin 2010, p. 75).

Data collector (Lee & Strong, 2003) or data producer (Wang 1998; Xu et al. 2002) is a person who is responsible of collecting, creating and producing data. Their purpose is to collect data for the consumer to use (Lee & Strong, 2006, p.17). Lee & Strong also note that the data needs to be accurate, complete and timely to serve the purposes of the data consumer. In this thesis these dimensions of data quality are deemed as intrinsic, thus the focus is in the work of data collector who is the person responsible for the intrinsic dimensions of data quality.

The data consumers use the data in their daily tasks for example in reporting (Xu et al. 2002, p.49). They use data by consolidating, interpreting and presenting it (Lee & Strong, 2003, p.16). These are all related to the contextual dimensions of the data quality which are presented in this thesis.

The data custodian (Lee & Strong, 2003), data manufacturer (Wang 1998) or data steward (Friedman 2007; Wende 2007) is a person responsible for data maintenance, data warehousing and data processing (Lee & Strong (2003), p.17). Wang (1998, p.60) notes that the data manufacturer develop, design and maintain the data and the systems for information products. In literature, the steward role is divided in technical and business steward (Wende 2007 p.429).

The business steward works in close contact to business representatives. They document the requirements of the business and evaluate the effect of those requirements to data quality and the data quality effects to business. Commonly this kind of steward is designated by business unit, business process or a data domain. They are responsible of the data quality standards and policies which are demanded by the data quality council Wende (2007) also referred in the literature as data governance council (Dyche & Levy 2006). They are able to communicate with the data quality council to create these standards and policies based on business requirements. (Wende 2007, pp.420-421)

The technical data steward complements its business counterpart. They focus to the technical representation of the data in the information systems. They can be assigned by a business unit, information system or a data domain. Their job is to offer standardized data and make sure data is well integrated in the whole system architecture. (Wende 2007, pp.420-421)

Data quality council defines the data governance model for a company. It sets strategic goals and assures that they align with the business goals of the company. It is responsible for companywide standards, rules, policies and processes to assure the constant improvement of data quality. Data council assigns the responsibilities to the data stewards and owners. Data council is led by a chief data steward whose role is to make the councils decisions take effect. He has strong business and ICT background and understands deeply the data quality effects and challenges in a company. (Wende 2007, pp.420-421)

Data owner is a person responsible of the whole data process (Wang 1998). The data owner has a role in the organization which enables him to create, access and manage data (Berson & Dubov 2007). The owners should come from the business side rather than ICT side in the company (Berson & Dubov 2007). Ballou & Tayi (1989, p.320) see the data owner to be the person whose everyday responsibilities include the data collection, maintenance and usage. This gives them an incentive to take care of the data quality. Pekkola (2012) agrees to that and also notes that every critical data domain should have its own owner and that optimally the ownership should last the whole lifecycle of the data. The data ownership can also be seen problematic since when assigned a specific owner to data, the other users of the data may neglect their responsibilities and trust on the owner to take care of the data by himself (Hodkiewicz et al. 2006).

3. DATA QUALITY

Classically data is defined high quality when it satisfies the requirements for its intended use. This is referred as “fitness for use”. (English 1999; Redman 2001; Orr 1998; Wang 1998). It is important to notice that data is an important end-product by itself (Lee et al. 2006, p.125). When understanding that, the measurements of data quality become more business related instead of technically oriented (Lee et al. 2006 p.134).

Data quality is a critical factor in many information systems and implementation processes such as implementing an ERP (Xu et al. 2002, p.47). Most organizations experience data quality problems in some level (Huang et al. 1998, p.92). In many cases organizations believe that implementing a new system resolves all problems related to poor data quality. This often leads to the problems getting more complicated as the system architecture becomes more complex (Lee et al. 2006 p.3).

3.1 Data quality costs

In the research of master data, it is often claimed that the effects of poor master data quality are tremendous. For example, The Data Warehousing Institute estimated in 2002 the data quality problems to cost 600 billion dollars a year in U.S only (Eckerson 2002, p.3). Classically it is also agreed upon that data quality is responsible for significant costs in the range of 8-12 percent of revenue (Redman 1998, p.80). Olson (2003) conducted a survey for 599 companies and found that poor data quality management cost over 1.4 billion dollars to these companies every year.

Although the costs have great monetary impact, Eppler & Helfert (2004, p.312) state that there are very few studies that demonstrate how to identify, categorize and measure such costs. Eppler & Helfert also note that this is not only an academic research problem but a pressing practitioner issue.

In order to develop a systematic classification for data quality costs Eppler & Helfert (2004, p.313) conducted a meta-analysis that researched the literature for different cost categories. The major finding was that data quality consists of mainly two types, improvement costs and costs due to low data quality.

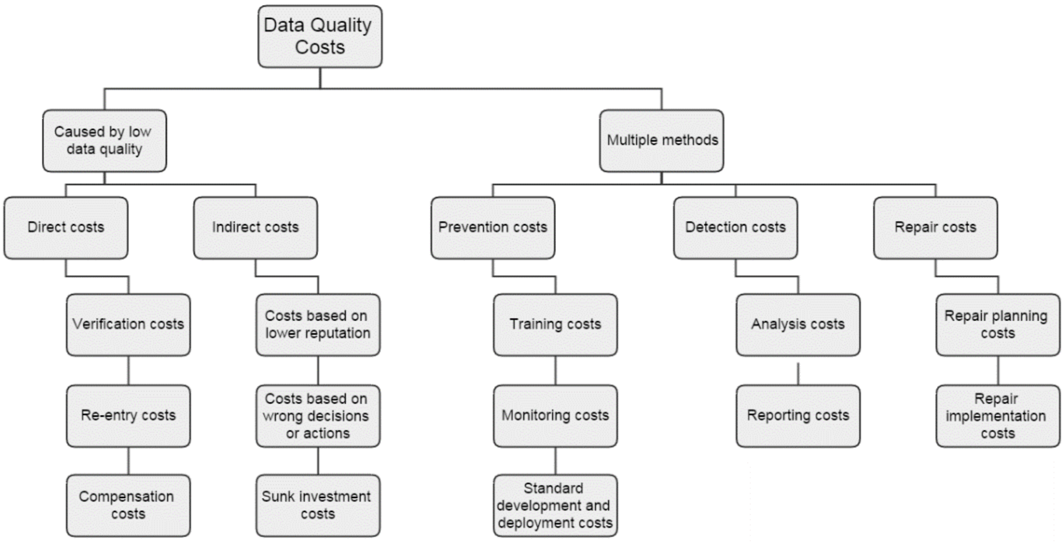


Figure 4. *Data quality cost taxonomy adapted from Eppler & Helfert (2004, p.316)*

The clear distinction in the low data quality section are the indirect and direct costs. Direct costs are those that have negative monetary effects that arise straightforwardly from low data quality. These costs would include the costs of re-entering data because it’s wrong, costs of verifying data for it to be right and the costs of compensation for the damage that results from bad quality data. Indirect costs are those effects that rise intermediately from low quality data. Those effects would include costs of deteriorating reputation to the premium of products or the effects of sub-optimal decisions based on bad data. The costs that arise in order to improve the quality data and so to diminish the previously introduced costs, are distinguished among prevention, detection and repair costs. (Eppler & Helfert 2004, p.317)

3.2 Data quality dimensions

Data quality has multiple dimensions that can be divided in different ways. Originally defined by Ballou and Pazer (1985) and most commonly mentioned in literature are accuracy, timeliness, completeness and consistency. (Xu et al 2002, p.47).

Wang & Wand (1996) conducted a meta-analytical literature review which summarized most often cited data quality dimensions. The most notable dimensions where accuracy, reliability, timeliness, relevance, completeness and currency. This is in line with Loshin (2010) who lists accuracy, consistency, completeness, timeliness and currency as most relevant dimensions for master data quality. Notably uniqueness steps out in Loshins definition in the master data point of view.

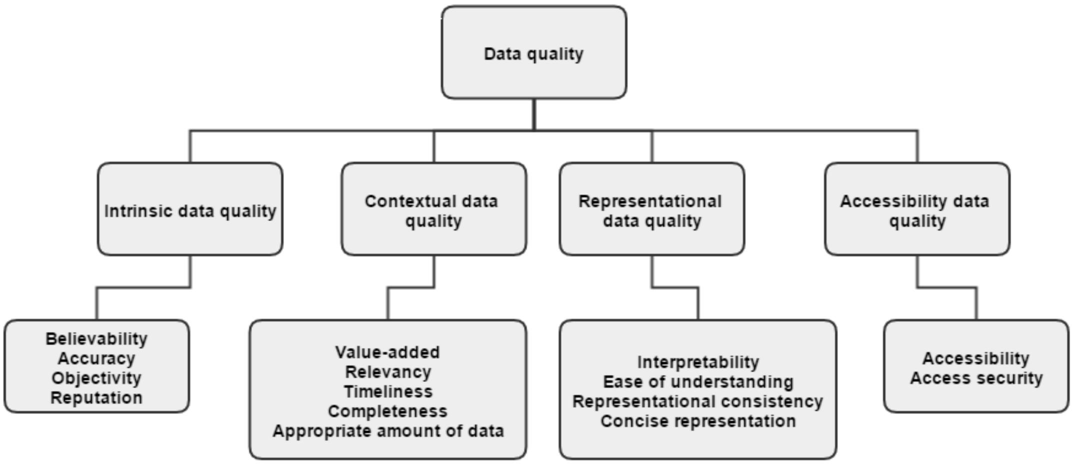


Figure 5. *Data Quality Framework adopted from Wang & Strong (1996)*

Loshin (2010) divides data quality to three different types which are also found in the data quality frame work by Wang and Strong (1996) depicted in Figure 4. Intrinsic data quality dimensions mean that data itself is valid and the syntax matches the requirements demanded from it. For example phone numbers in a specific area should follow a specified form. This relates to the accuracy dimension.

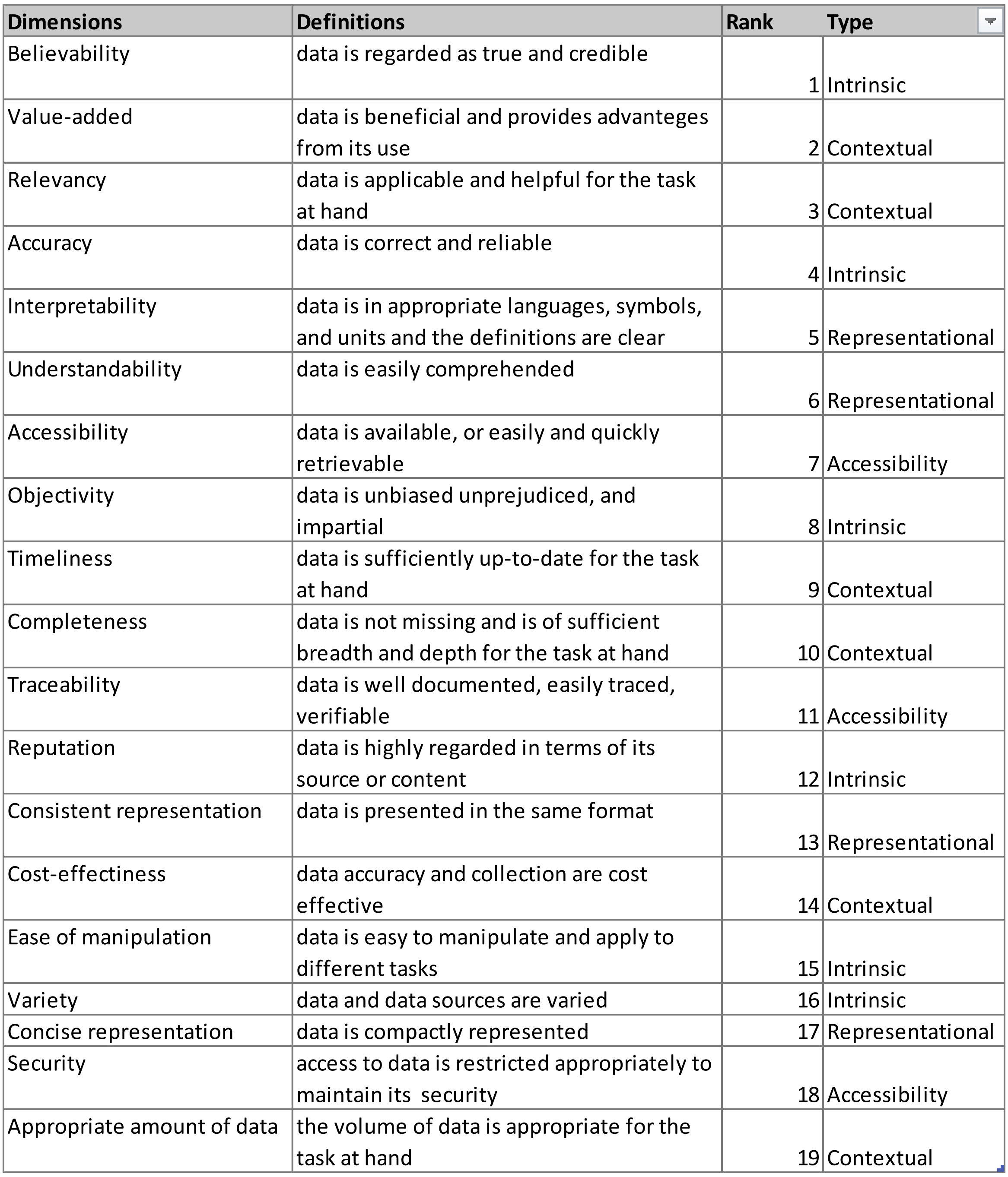
Wand and Wang (1996, s.93) notice internal or intrinsic and external or contextual. The intrinsic dimensions as how the information systems objects relate to real world. Their point of view to data quality how much it has errors. Perfect data would be data which has no errors in describing the reality. In other words, it would map the data objects perfectly to real world objects. For example, perfect employee data would have all the employees with all their predetermined attributes and nothing else.

Second type is contextual dimensions which are noticed by Loshin (2010), Wand and Wang (1996), Wang and Strong (1996) and Haug et al (2009, p.1058) who refer them as the usability dimensions. That means that data is concise between two records. The conciseness depends often on the agreements inside an organization such as how a specific business objects should be referred to.

The previous two types, contextual and intrinsic are common in literature. Other types vary more. Third type by Loshin (2010) is representational quality which relates to more subjective dimensions such as interpretability and ease of understanding. Wand and Strong (1996) notice also this dimension, but they found an additional dimension which relates to accessibility. Accessibility or availability is also noticed by Haug et al (2009).

As a summary, there are four types commonly noticed dimensions, intrinsic, contextual, representational and accessibility. These types include many dimensions.

Table 2. *Data quality dimensions ranked by importance adopted from (Pipino et al. 2002 ; Wang & Strong 1996)*



Dimensions	Definitions	Rank	Type
Believability	data is regarded as true and credible	1	Intrinsic
Value-added	data is beneficial and provides advanteges from its use	2	Contextual
Relevancy	data is applicable and helpful for the task at hand	3	Contextual
Accuracy	data is correct and reliable	4	Intrinsic
Interpretability	data is in appropriate languages, symbols, and units and the definitions are clear	5	Representational
Understandability	data is easily comprehended	6	Representational
Accessibility	data is available, or easily and quickly retrievable	7	Accessibility
Objectivity	data is unbiased unprejudiced, and impartial	8	Intrinsic
Timeliness	data is sufficiently up-to-date for the task at hand	9	Contextual
Completeness	data is not missing and is of sufficient breadth and depth for the task at hand	10	Contextual
Traceability	data is well documented, easily traced, verifiable	11	Accessibility
Reputation	data is highly regarded in terms of its source or content	12	Intrinsic
Consistent representation	data is presented in the same format	13	Representational
Cost-effectiness	data accuracy and collection are cost effective	14	Contextual
Ease of manipulation	data is easy to manipulate and apply to different tasks	15	Intrinsic
Variety	data and data sources are varied	16	Intrinsic
Concise representation	data is compactly represented	17	Representational
Security	access to data is restricted appropriately to maintain its security	18	Accessibility
Appropriate amount of data	the volume of data is appropriate for the task at hand	19	Contextual

As seen in figure 5 the type of the dimension does not correlate on the importance of the dimension. Intrinsic and contextual dimensions were the most important in the research by Wang & Strong (1996). Representational were felt also important by the interviewees in their study. The least important dimensions were those related to accessibility or availability but in the study of 355 people it was found the differences were not large between the four types of dimensions.

3.3 Master data quality and its barriers

Almost all organizational functions use data and it is the basis for operational, tactical and strategic decisions (Haug & Arlbjorn 2011, p.290). That's why in order to improve its effectivity it is critical for organization to have high enough quality data (Madnick et al., 2004, p.43). Many studies reveal that data quality is often left without the attention it needs (Marsh 2005, p.105).

Master data is once created, largely used and rarely changing (Knolmayer & Röthlin, 2006, p.363). Knolmayer and Röthlin (2006) notice that despite that, the data quality maintenance must be ongoing.

In this sub-chapter the barriers of master data quality are discussed. These are the factors that prevent achieving higher master data quality in the organization. Overcoming the barriers, the organization can more easily allocate resources to right causes achieving higher quality data.

In their study, Haug & Arlbjorn (2011) set out to determine the biggest barriers to master data quality. They sent the questionnaire to over 1000 companies in Denmark. They also conducted a literature review regarding the most important challenges to the master data quality. This subchapter follows their review.

In according to Umar et al. (1999, p. 299) describes six barriers of data quality in his case-study conducted in ICT industry. They are the lack of roles and responsibilities, data quality owners, reward systems, organizational procedures and the lack of scheduling of the data movements in multiple system architecture.

English (1999, s.422) defines the critical success factors to data quality, and reasons why these factors don't realize. The reasons are the lack of training, inducements and the lack of managerial understanding and participation.

In according to Xu et al. (2002, pp. 54-55) in their review of literature publisher before year 2000 they list the factors that affect data quality. They include the support from management, the organizational structure, change management, employee relationships, data quality training and data quality controlling such as input controls and segregation of duties. They state that the most important of these are the managerial support and the education of employees. In this particular study the education was defined as how well the end users were able to use the end-system which had direct effect to the quality of the data they put in.

In their study Lee et al. (2006, p. 31) researches information quality assessment. They also focus on the challenges to data quality. They find the lack of accountability of information quality, tools, fitting technologies and right procedures to be the main reason for the low quality of data.

Data siloes is the point of view that Smith & Keen (2008, pp. 68-69) take in their study. The data siloes in this case mean how the data is managed locally in local companies or distinct LOB (Line of Business)'s which leads to the data to be stored in diverse places, siloes, which are not harmonized with one and other. This also adds to the problem of indistinguishable ownership of data. They state that the problems of data siloes has gone worse since data storing techniques have improved, data is stored more but at the same time the ability to manage, use and analyze has not improved nearly as fast (Davenport 2007, p.154). New and more integrated systems such as ERP's make the data management even complicated (Fisher, 2007). Companies tend to try to solve the data problems with half-measured and ineffective solutions that can be even counterproductive. As companies work towards global management of data, the means are often supporting only the short term goals which leads to problems in the longer term.

As a summary, Haug & Arlbjorn (2011) found out from the literary reviews that the main barrier for master data quality was the roles and accountabilities regarding master data and its maintenance. The questionnaire-study which they then performed would support this conclusion and would suggest that more specifically, the lack of delegation of said accountabilities regarding master data maintenance, would be the core problem. Other reasons which were found in the questionnaire as well as the literary review were the lack of control routines and lack of employee competencies. The rewards and incentives around data quality which was also introduced in literary did not find support in the questionnaire-study.

3.4 Data quality assessment

In order to answer the question about "how good is company's data quality" or to assure if data is "fit for use" we need to assess the current data quality (Pipino et al. 2002, p.211, Woodall et al. 2013, p.369). As introduced in previous chapters, data quality can be viewed from many different directions. In that sense, data quality can be measured with a myriad of ways. The subjective quality can be determined by making business users answer a questionnaire about the data quality. In the other hand the intrinsic quality can be measured with technical measures such as fitting to regular expressions. In order to determine how well data quality meets the business user requirements, it is necessary to assess data along the business process (Cappiello et al. 2004 p.68)

Pipino et al. (2002) and Lee et al. (2006) introduce three functional forms for developing data quality metrics. The approach combines the subjective and objective assessment of data quality and illustrates how it can be used in practice. They note that in practice, the terms data and information are used interchangeably by business personnel, but here the focus is only on data. Loshin (2011) represents same kind of forms of data quality assertion but calls them control charts. These forms are well noticed in different studies.

The first from is the simple ratio of the data. It measures the ratio of desired outcomes compared to total outcomes. In most cases the undesired outcomes are more crucial so this can be turned to be the ratio of undesired outcomes compared to total outcomes. Many data quality dimensions can be measured with this form of measurement such as consistency, accuracy and completeness. In practice this can be a matching to a regular expression or counting missing values of a specific data attributes Loshin (2011) refers to the simple ratio as percentage nonconforming. (Pipino et al. 2002)

Min or max operations can be applied to handle data dimensions which need the aggregation of multiple data quality indicators. The individual indicators can for example be the simple ratio values. If the quality of a data set would be evaluated with the simple ratio which would state the percentage of undesired outcomes regarding different attributes, the max operation would state that the largest of these values would be the one to be considered. This would be applicable for example to the believability of the data, so if for example, four metrics out of five would be in required level and one would have too much of undesired values, the largest and worst one would be the one that determines the believability. (Pipino et al. 2002)

For some dimensions the weighted average would server better than min or max to determine the quality rating for the data. For the believability this would mean that every dimension would be determined with some importance value and the weighted average would state the overall quality of data set with multiplying the ratio with the importance and summing it all up. (Pipino et al. 2002)

These are all very simple methods for evaluating data, but in many cases the data quality is not measured at all (Batini et al. 2009 p.2). Regarding master data, it is important to understand that the simplest dimensions would need to be the ones that would most easily be improved thus giving most cost benefits with least effort. This also follows the Pareto principle. Loshin (2011) notes that regarding the data quality, the 80% of effects result from 20% of the causes. This leads to the notion that it is crucial to be able to focus to the most relevant variables that cause most of the problems. In this case the min or max analysis could be used to determine the most deteriorated dimension of data and it could be focused for maximum results for the minimum effort.

In practice, Pipino et al. (2002, p.214) suggest that there are three steps need to be taken in order to increase organizations data quality. First a data quality assessment needs to be made in objective and subjective manner. Then the results need to be compared, discrepancies identified from them and root causes determined of those discrepancies. Last the necessary actions for improvement needs to be determined and taken.

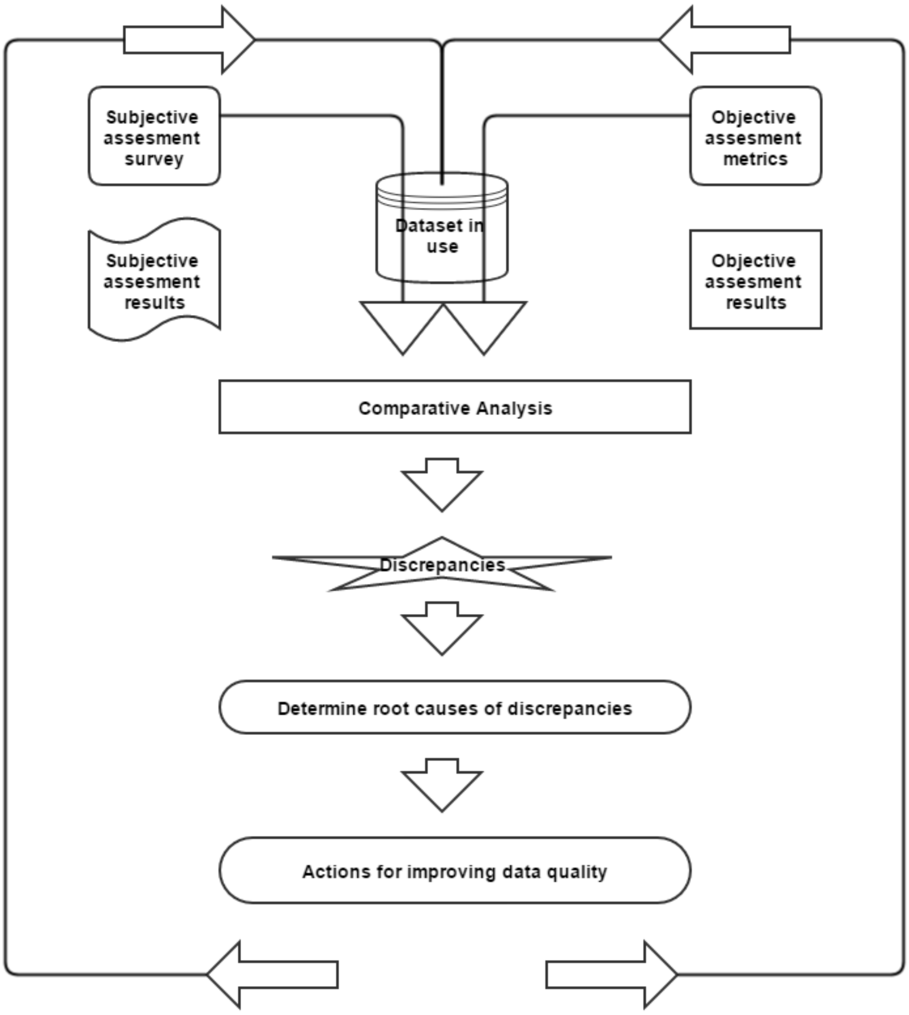

Subjective assesment survey
Objective assesment metrics
Dataset in use
Subjective assesment results
Objective assesment results
Comparative Analysis
Discrepancies
Determine root causes of discrepancies
Actions for improving data quality

Figure 6. *Data quality assessment in practice (adopted from Pipino et al. 2002, p 216)*

Following the data quality improvement process introduced in the figure X, the goal is to achieve data quality that is high in the subjective as in the objective point of view. If either the subjective or objective data quality seems low, the company needs to iterate the process determining the discrepancies, root causes and the actions needed.

4. MDM MODELING AND ARCHITECTURES

The core issue for MDM is to be able to capture, consolidate and deliver the master representation for each uniquely identifiable entity (Loshin, 2010, p.143). As Loshin (2010, p.143) states, a skilled data modeler needs significant amount of training and experience and for that reason it is relevant to only introduce the most significant issues to be considered in MDM model developing.

4.1 MDM models

There are a multitude of ways to logically and physically building a master data repository (Loshin, 2010, p.144). The core issue in every one of these ways is the need to model the identifying attributes for each master entity. There can be lots of identifying attributes, some more identifying than others. Loshin (2010, p.144) states that in order to create a defining key for an entity, using the identifying attributes, one can follow a simple heuristic:

1. Determine truly descriptive data elements for the entity
2. Seek out the data element whose values are most distinct and add it to the key
3. If the key is not yet unique, return to step 2
4. Key is complete

This process can be supported by data profiling tools introducing frequency analysis, null value analysis and uniqueness assessment.

Dreibelbis et al. (2008) observers that there are three cornerstones in starting MDM modeling. The data model should be robust which provides a solid foundation for low-risk implementation that leads to an early win. Robust data model usually requires only few data model extensions. If extensions are required, the software selected should easily support the extension of data models ensuring the consistency. Allen and Cervo (2015) agree on this and note that the robustness and customizability play a large factor on MDM modeling and the solution on which the modeling is done.

4.2 MDM hub architectures

Loshin (2010), Dreibelbis et al. (2008), Allen and Cervo (2015) and White (2006) all agree that MDM logical architecture can be distinguished to three different styles. This subchapter follows their definitions.

Transactional implementation style is one where every transaction goes through the master repository. All master data is persisted in the hub. It is the strictest of the architecture styles and consists of highly coupled architecture which means that it takes massive work on integrating the existing applications to it and can be referred as the “thick” style. The most minimal architecture style is the (virtual) registry. Virtual registry style only references or links to master data entities stored elsewhere. They are distributed across different systems and can be referred as the “thin”. None of the data is persisted in the hub. The medium of these is the “centralized registry” (Loshin 2010), “coexistence hub” (Dreibelbis et al. 2008) or “hybrid hub” (Allen and Cervo 2015; Loshin 2010). From now on this style is referred as hybrid hub as it is hybrid between the most well defined registry and transaction styles. The hybrid hub is the repository for master data and is the source for core data objects which are published out to the application systems.

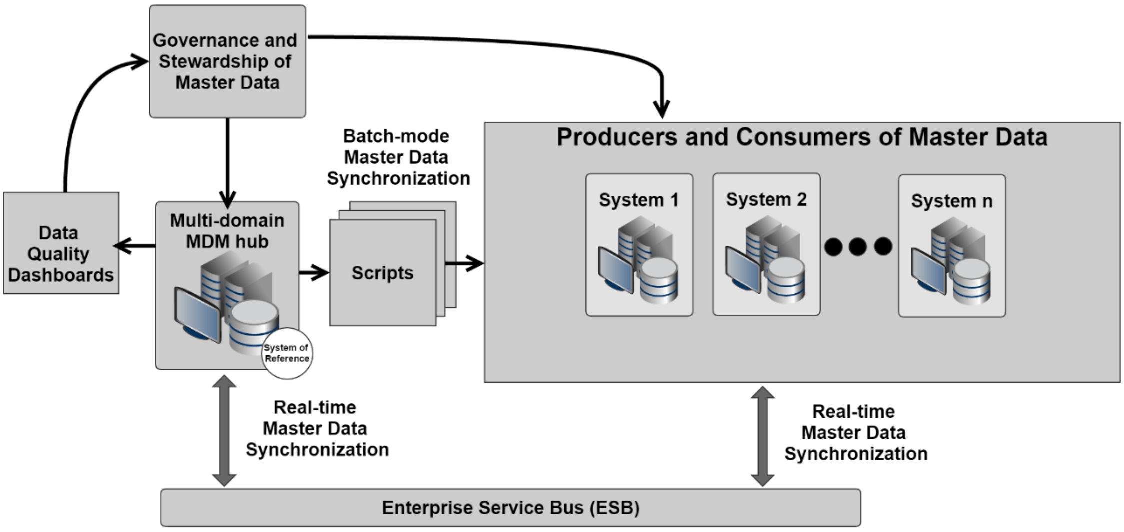

Governance and Stewardship of Master Data
Batch-mode Master Data Synchronization
Producers and Consumers of Master Data
System 1
System 2
System n
Data Quality Dashboards
Multi-domain MDM hub
System of Reference
Scripts
Real-time Master Data Synchronization
Real-time Master Data Synchronization
Enterprise Service Bus (ESB)

Figure 7. *MDM hybrid architecture styles (Adapted from Loshin 2010; Dreibelbis et al. 2008; Allen ja Cervo 2015)*

In this thesis only the hybrid version of MDM architecture, which is called here “hybrid MDM hub” and is specified more closely in the following subchapter

Hybrid MDM hub offers a single model to manage the identifying attributes as well as common data attributes consolidated from applications. It provides a single version of the truth with high quality master data physically stored in a centralized repository. It also enables business agility, since a new application can subscribe to MDM system to retrieve the changes in the master data. Thus it simplifies integration architecture and reduces the amount of point-to-point connections needed. (Loshin 2010 pp.169-170, Dreibelbis et al. 2008)

The integrated view of common attributes in hybrid MDM repository provides the “source or truth” or the “golden record” with data absorbed into the master from the distributed business applications. Hybrid MDM hub is not the system of record itself, and the maintenance happens in the original sources. The hybrid hub is synchronized to the source systems and the replicated data is kept refreshed. Since in many cases it is logical that systems keep continuing on their own without a constant connection to the MDM system, the hybrid approach is well fitting. It means that the system offers harmonized view of unique master data objects but does not require a high degree of integration and synchronization across the applications. (Allen and Cervo 2015, Loshin 2010)

Loshin (2010) also introduces the basic characteristics of hybrid master data architecture:

- Common attributes are managed in a central repository
- Applications maintain their full local copies of master data
- Centralized master data is standardized representation and it is published out to the client applications
- Unique identifier maps the master instance objects to the client side instances
- Application-specific attributes are managed in the client application side
- Consolidation for master data is performed periodically
- Integrations may have flags that notify the central MDM system of the changes in the other applications

Dreibelbis et al. (2008) describes a theoretical reference architecture for the hybrid (or coexistence hub). The reference architecture consists of eight architectural building blocks which together form the Master Data Management Services reference architecture.

Interface Services of such hub can support multiple technologies for the interfaces. The interface service logic is same for single transaction via interface or large batches in order to maintain a consistent business logic.

Lifecycle Management Services in the hybrid hub supports the authoring of the master data, including CRUD operations. Business logic can also be authored. The Life Cycle Management Services use Data Quality Management services to enforce data quality rules and perform harmonization tasks for the data.

Data Quality Management Services is responsible for the data quality functions in the reference architecture. This building block is not only important in the build phase but also in the operational phase. DQMS verifies every new master data record in terms of duplication, completeness, accurateness and other data quality dimensions selected.

Master Data Event Management Services register events taking place in the MDM system. For example when a batch import of data exceeds the time window set for it, the MDEMS may trigger a risk notification.

Hierarchy and Relationship Management services determine the hierarchies and relationships inside and between the data entities.

Authoring Services take care of the authorization of MDM system. In the case of hybrid MDM hub, the authorization can be done in MDM layer, but in most cases it is done in other levels such as database.

Base Services represent the basic security, privacy, search and audit logging of the MDM system. Workflow capabilities may also be available in the hybrid hub implementation. Usually this is integrated to Enterprise Common Services as Microsoft Active Directory.

The Master Data Repository is the building block supporting the actual architecture. Master data model is fully instantiated to this repository and master data is materialized in the MDM System.

4.3 Choosing the MDM solution

Gartner (Radcliffe, 2007) states that there are seven building blocks to work towards selecting a MDM solution. They are the vision, strategy, governance, organization, processes, technology infrastructure and the metrics of MDM.

First there needs to be a business vision that requires an MDM vision to enable it. The business vision needs to be clear and there needs to be a clear vision of the scope of the MDM solution. It must be clearly stated how MDM vision supports the business vision and there needs to be clear and enduring business benefit justification for the MDM in order it to make itself useful in long term.

Allen & Cervo (2015) note that the difficulty in finding an existing MDM solution, that fits the organization’s needs, originates from the impossibility of universal data model that would reflect every company’s business requirements. That makes it clear that the MDM solution should be very customizable.

4.4 Microsoft SQL Server Master Data Services

Microsoft SQL Server Master Data Services (later MDS) is a MDM product from the multinational software company Microsoft. The product was originally introduced by company called Stratature which Microsoft acquired in 2007. The product is shipped with Microsoft SQL Server and is compatible with other Microsoft products including an Excel add-in. (Microsoft 2015)

MDS is a MDM product which aims to create a centralized data source and keep it synchronized reducing redundancies across all the applications which process the data. MDS uses Microsoft SQL Server database as the physical store and is part of the Master Data Hub Architecture introduced by Microsoft. The hub extracts data from source systems, validates and harmonizes the data, removes duplicates and updates hub repositories and synchronizes the external sources with it. Entity schemas, attributes, entity hierarchies as well as validation rules and access control are specified in MDS metadata. MDS allows custom business rules, which are user made rules for data validation. Business rules can be made via the web interface and so is available for the business users. The web-based UI (User Interface) can be in general used to view and manage data. All changes made in the MDS are validated against the rules and a persistent log of the transactions is stored. MDS supports versioning of the data entities as well as the option to notify user from all business rule violations. (Microsoft 2015)

MDS allows the hierarchical categorization of entities. For example a company's specific operational site is a subtype of the legal company, which represents the legal entity. Hierarchies are generated by relations between data attributes. MDS stores data entities to the database in such a way that it can be subscribed to using SQL Server views, which are dynamically generated having the latest data available. The access control used by MDS is role-based where the specific roles can be appointed to a user or a user group. (Microsoft 2015)

Other features available in the MDS product is the Web service interface for exposing the data and an API which can be used to manipulate the data programmatically. This gives the possibility to create for example web UI for a specific business use on top of it. (Microsoft 2015)

MDS has its own terminology which is in line with the terminology in general of the industry and the most crucial terms are introduced in the following subchapters. (Microsoft 2015)

4.4.1 Model

Model is the highest level of data organization in MDS. It defines the structure of data in the master data solution containing entities, attributes, hierarchies and collections.

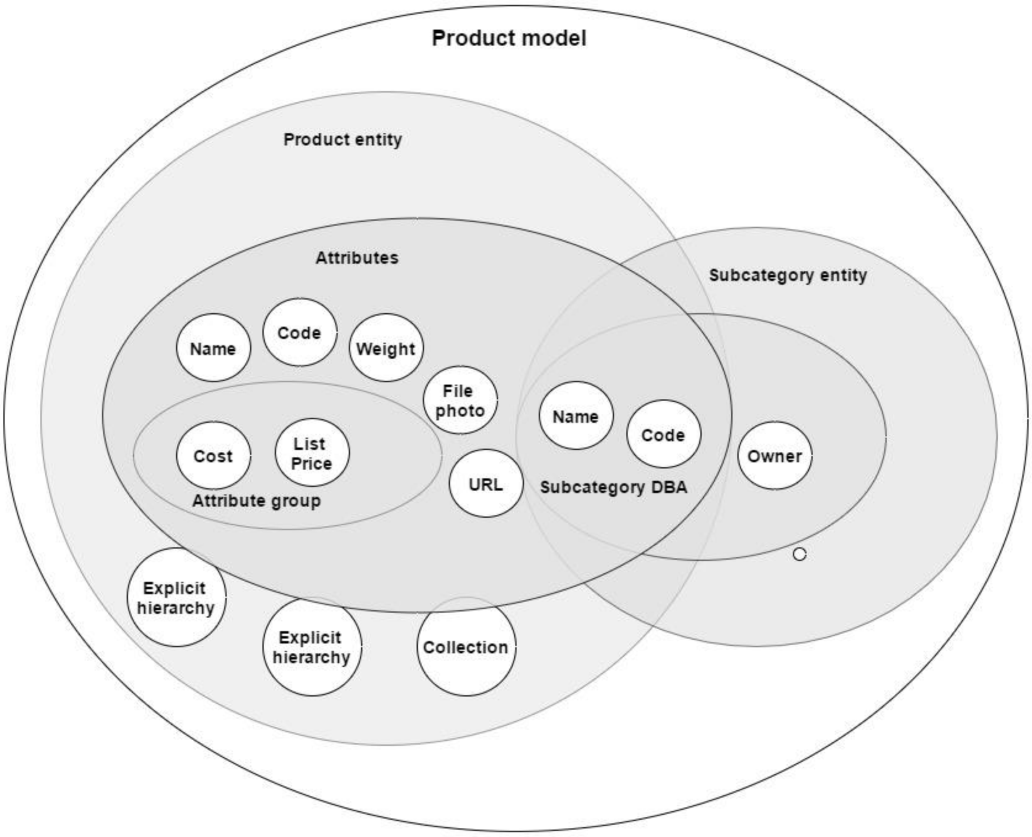


Figure 8. *MDS model example (adapter from Microsoft 2015)*

As seen in the figure 7, the model in the MDS is the highest level of data structures. MDS may have one or many models which group up similar data, for example product master data model to contain product related data or customer master data model to contain customer related data. The permissions can be assigned within a model, and for every model the permissions must be set for the user to be able to see the data. Copies can also be made constantly of the data and these copies are handled thru version management application which is called “versions”. Same model used in test environment can be deployed to production easily without the data, thus preventing the need for recreation of the specified model. (Microsoft 2015)

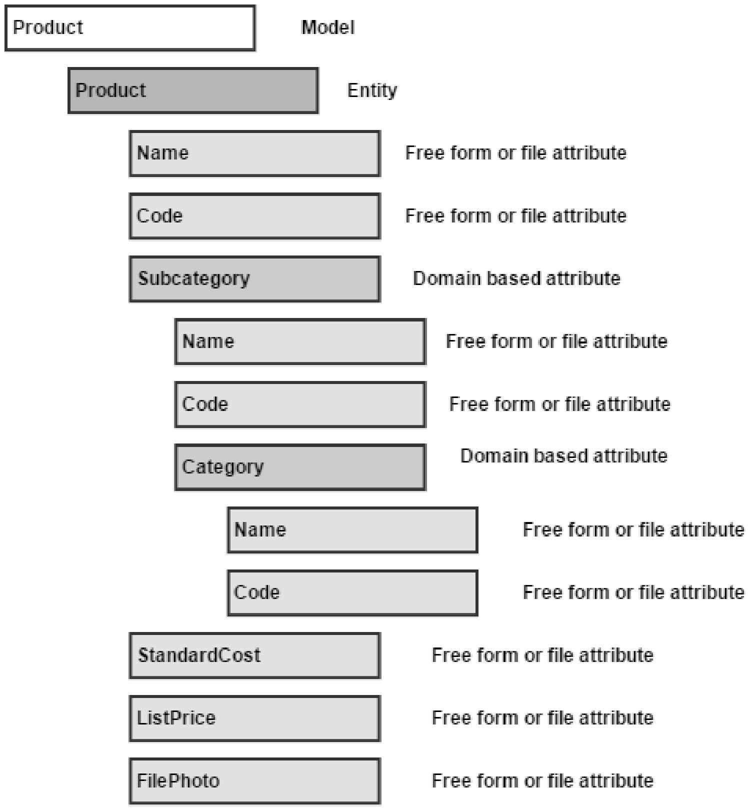

Product
Model
Product
Entity
Name
Free form or file attribute
Code
Free form or file attribute
Subcategory
Domain based attribute
Name
Free form or file attribute
Code
Free form or file attribute
Category
Domain based attribute
Name
Free form or file attribute
Code
Free form or file attribute
StandardCost
Free form or file attribute
ListPrice
Free form or file attribute
FilePhoto
Free form or file attribute

Figure 9. *Hierarchical example of a Product MDS model (adapted from Microsoft 2015)*

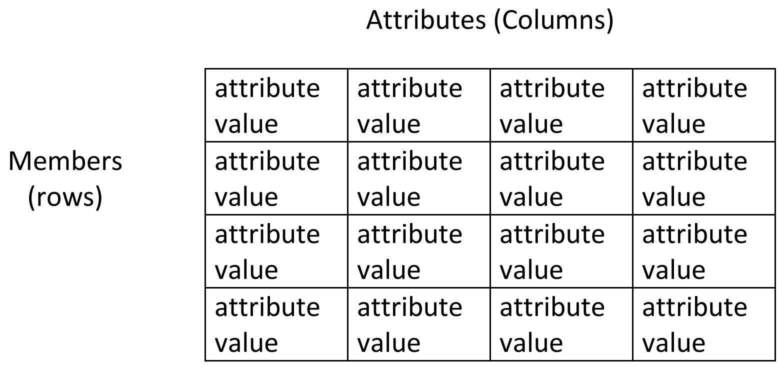
Other common models to be represented in MDS are accounts which could include balance sheet accounts, income statement accounts and so on. Usually geographical data is also stored in MDM hub and a geography model could include entities such as postal codes, cities, states etc. It is important to notice that there cannot be references between entities in different MDS models, so it could be problematic to have geographical data separate from customer and his address information. (Microsoft 2015)

4.4.2 Entity

Entities are data containers within a specified model. They contain members which are the rows of master data that are managed. In that sense entity is somewhat similar to a table in a database.

Models may have an indefinite number of entities. Entities should group similar kinds of data. For example an entity could represent the master list of products in a company or list of product categories. Most cases other entities are more relevant for business than others and to which the other entities in the model are related to. For example the product category is related to the product entity. There can be also several entities that are of equal importance, for example linking products to suppliers.

Entities can be thoughts as a table for master data where rows represent the members and columns represent the attributes.



Attributes (Columns)

Members (rows)

attribute value	attribute value	attribute value	attribute value
attribute value	attribute value	attribute value	attribute value
attribute value	attribute value	attribute value	attribute value
attribute value	attribute value	attribute value	attribute value

Figure 10. *Product entity example in MDS (adapted from Microsoft 2015)*

The example of a product entity in Figure 9 shows how the entity defines a real life business object in the MDS database. Entities can build derived hierarchies where for example product entity could reference a product group entity which references a product segment and so on. These entities are level-based and contain multiple entities and are so called derived. Explicit hierarchies in the other hand are hierarchies where an entity refers to itself for example a product entity could have an attribute “component” which refers to the entity itself and to a member that is a component of a the specified product. (Microsoft 2015)

Entities may act as constrained lists having a list of values for a specific purpose. For example, a unit of measure –entity would have all the business relevant unit of measures listed to which a product entity refers to. This helps managing the possible unit of measures used in the products and so helps diminishing invalid values thus supporting data quality. Entities can be used as constrained lists in more general sense for example when a product refers to a supplier via MDS linkage. These kind of linking attributes depicting foreign key relationships are called domain based attributes in the MDS. It is also possible to set a base entity for a model which is the entity that is shown first when opening the explorer in the web UI. In a product model the base entity would most probably be the product entity to which other entities would relate to. (Microsoft 2015)

4.4.3 Members and attributes

Members are analogous to rows in a database table and attributes are analogous to columns. Related members are contained in an entity and each member is defined by its attribute values. In previous figure 9, the attributes can be seen as columns and attribute values in distinct cells. MDS has three types of members that are leaf members, consolidated members and collection members. Leaf members are the default members in an entity. “Leaf” is analogous to a leaf of a tree so they are at the lowest and most specific level in a hierarchy. Consolidated members are formed when there are explicit hierarchies for the entity. Collection members belong to a collection, which is analogous to a view in data base.

Attributes can be domain-based resembling a foreign key. Attributes can also be included in an attribute group, so the user can pick which attributes he wants to see about the customer, not having to see every one of them.

5. CASE STUDY

The empirical part of this research was conducted as interviews. The interviews where theme-based around the research questions and objectives of this research. In this chapter it is described how the study was conducted, how the data was collected and how it was analyzed. Before that, the decisions behind choosing the methods are described.

As this study is of a qualitative nature, the aspects of qualitative data is more closely looked upon.

5.1 Methods

In order to select the most suitable methods for the data collection and analysis, it is necessary to go through the vastness of the research methods. The method should always be in line with the type of problem under research as well as the resources available. In this chapter the selection process of methods is described.

5.1.1 Data collection

In qualitative research, the methods are normally of a qualitative nature. The qualitative methods in practice are usually interviews where the subjective visions and thoughts of the interviewees help to describe the study problem.

There are many types of interviews. One commonly used is the classification to structured, semi-structured and unstructured interviews. Structured interviews have a predetermined structure and use standardized set of questions on every interviewee. In most cases there are multiple choice answers from which to select. The social interaction plays a very small role, since the goal is to keep the interview similar for everybody and the questions and answers represented just as they are written. Structured interviews are often designed for a large number of people to answer. As the results are very quantifiable data, the method can be described as quantified research interview. (Saunders et al. 2009, p. 320)

The other types of interviews are non-standardized from which the semi-structured interview is still the more standardized one. In semi-structured interview the researcher has a list of themes and questions to be covered. These can vary between interviews, but the themes are usually the similar despite the context. Also compared to the structured interview, the semi-structured are less restricted on how the questions are represented and how the social aspect of the conversation flows. This may also lead to variance on the questions. In some cases, additional questions need to be asked to make the same question well understood to every interviewee. (Saunders et al. 2009, p. 320)

The other type of these non-standardized qualitative research interviews is the unstructured interview. They are informal and suited for exploring in depth a general area of interest. There is no list of questions and the interviewee is given an opportunity to talk freely about any event, belief or behavior in the topic area. This is sometimes called as a non-directive type of interaction. It is also different from the described interview types so that the in many sense interviewee is the one that directs the interview instead of the interviewer. (Saunders et al. 2009, p. 320)

The differentiation between interviews can be also determined how the nature of the interaction is. The interview can be face-to-face, or via a medium. The interview can be also conducted one-to-one or many-to-one situations. One-to-many offers also the possibility for different dynamics as the group can be selected in multitudes of ways. (Saunders et al. 2009, p. 320)

5.1.2 Data preparation and analysis

Qualitative and quantitative data are very different in terms of data analysis. Usually, as the concept gets more ambiguous and elastic, the harder it is to quantify the data in a meaningful way.

Quantitative data is based on meanings derived from numbers. The quantitative data collection results in standardized and numerical data. This makes it possible to conduct the analysis through the use of diagram and statistics and related tools.

Qualitative data is based on meanings that are expressed through words. Collection of qualitative data results in non-standardized data which requires classification into categories. It also often needs to be summarized and restructured as a narrative to enable meaningful analysis. It may be possible to take advantage of diagrams and statistics in qualitative analysis. One of the ways could be to study the frequency of certain groups or categories in the data and from that derive some estimation of its role or importance. (Saunders et al. 2009, pp. 480-484)

Qualitative research and qualitative data analysis need conceptual framework to support it. This can be formulated before, during or after the data collection. The analysis of qualitative data can be seen as a very demanding process and so it cannot be perceived as an easy option. The basis for analyzing of qualitative data is usually structured in the earlier stages of the research process. The formulation and clarification of research topic, the reviewing of applicable literature, the decision of the research design and structure and the consideration of access issues as well as the conduction of the interview, all are needed in order to be able to start conducting the data analysis. (Saunders et al. 2009, pp. 480-484)

When starting to prepare qualitative data for the analysis, it can be in many forms. These forms can be in written form, such as reports or emails and non-written form such as audio- and video-recordings. In the latter cases it is extremely important to transcribe the recordings to ensure no data is lost. If there are some separate notes, it is also important to include them in the transcript. One other thing to note that in the interview, it is not only important to record what the interviewee said, but also how it was said. This means that the transcribing audio-recorded interviews can be very time consuming as in addition on recording what was said and by whom, it is required to record how it was said and in which tone of speech. The non-verbal communication needs to be able to be linked to the context of the interview. (Saunders et al. 2009, pp. 480-484)

The biggest problem with transcribing are time consumption and errors. There are still ways to help out the process. Two notable ways described by Saunders et al. are the dictation of the records with voice recognition software and the transcribing of only the most important sections of the interviews. This of course presumes that there are sections that can be cut off without a loss of valuable information. (Saunders et al. 2009, pp. 480-484)

In addition of transcribing the audio to text, there can also be conversion of data already in text format to a more suitable format. This data such as emails, need to be checked for typographical errors, anonymized and appropriately stored to match the other transcriptions. (Saunders et al. 2009, pp. 480-484)

5.2 Conducting the study

After deciding the research questions and objectives and the methodologies to conduct the study, it was time to start planning the data collection. The method for data collection was need to be chosen to support the style of the problem at hand. As problem was seen as of qualitative nature, it was clear that the data collection had to be done in contact to phenomenon. As the phenomenon was very hard to documentation via observation or other such method, interviews were seen as most viable collection style.

The next decision to be made was how structured the interviews would be. Structured interview seemed too restricting and liming the human interaction which would lead to the hindering the transformation of tacit knowledge to explicit. As some structure would help in classifying the data which would help analyzing it as well as help the interviewees structure their thoughts. The structure of the interview was set to start from wider subject and slowly move to questions that help answering the more specific research problems set in this study.

The interviews were designed to support answering the research questions. The interview had questions related to supporting research questions and also to the main research question. The themes of the interview questions were hierarchical, so the first question was about data quality, next was about master data quality and so on. There were six distinguished themes:

- Data Quality
- Master data and its quality
- Good enough master data quality
- Problems with master data quality
- Master data quality in MDM hub
- Improvement of data quality in MDM hub

The full list of questions can be found in Appendix A.

As discussed earlier the questions and themes move from more general to more specific. This resides on the fact that data quality is needed to be discussed before master data quality, since most things that apply on master data are defined to apply on data in general. The questions combine the three subjects of data quality, master data and its management and a specific architecture decision of MDM hub. The goal is to add these three subjects to the active memory of the interviewees when conducting the final results on how master data quality should be managed in MDM hub.

The interviewees were chosen from people who are working with master data in daily basis and have knowledge on the studied phenomenon of master data quality in an MDM hub. The goal was to include interviewees with deep enough expertise both technically and business logically. These interviewees had broad experience from different projects in different environments. They also had different backgrounds, both in studies and work history. These different points of view would help in extracting varied answers to the research questions and would help to mitigate the biases that may rise from each other’s backgrounds. The interviewees were contacted and a suitable time for interview was set. The questions were not revealed to the participants until the interview came.

Six people were interview in total. There were four consultants, one specialist and one architect. The average relevant work experience was 4 years and the all the interviewees had university education in a relevant subject. All of the interviewees could be stated to have a technical role, but everyone also had to have understanding of the business processes in order to succeed in their job. Interviews were held during the fall of 2015. The interviews were theme based and conducted face-to-face and one-to-one. The interviews were in Finnish.

At the start of the interview a brief background of the study was introduced. The reasons why this study is conducted and what its goals are were told to the interviewees. The structure of the interview was also explained. The interviews had no strict time limit, the interview was ready when all the questions were answered. The interview question and pace were still designed to last under an hour, since it was time that would be easier to reserve from the calendars of the interviewees. The interviews took from 30 to 50 minutes. All the interviews were recorded. This happened via phone recording software. Based on the recordings, the interviewees were transcribed to separate documents. After the utilization of the material, all material was destroyed to protect the anonymity of the interviewees and also to conceal possibly classified information.

After every interview, the interviewing process evolved a little bit. The first interview set a base line of what to expect and gave hints on how to present the questions so that they are clearly understood. Every question was represented on how it is found written in Appendix A. Some of the interviewees did not understand the question and some clarification on the question needed to be given. If it seemed that the interviewee answered in a fashion that would suggest he did not understand the question, it would be presented again with more clarification. The transcription was done after every interview, so it was easy to compare and reflect the answers to each other. After two interviews there were noted differences in the answers. It was also possible to group up and categorize the answers. The classification to categories was an incremental process. After interviewing all the participants, the final summarization and categorical classification of answers was done.

The analysis of the interviews was based on the answers gotten from the interviewees. In the result part of this study, the answers are guided through theme by theme. The results and answers are described and supported by the most descriptive quotes that represents holistically the answers gotten from interviewees. The interview results are synthesized to form a one view on how the interviewees think. This should give some statistical reliability to the analysis. The synthesizing is a very qualitative part of the process and the interviewers’ biases may effect on how well the answers are understood and if the most relevant notions are pointed out of the combination of the answers.

6. RESULTS

In this chapter, the results from the interviews are described. The interview had six themes and these themes are divided to three subchapters. First subchapter represents the first interview theme which data quality and the second which is master data and its quality. Second subchapter discusses the good enough master data quality and problems related to it including their root causes. Third subchapter discusses the MDM hub and sums up the most relevant proceedings to be taken. The questions and their answers are meant to support and depend on each other, so the same statements could be said under all of the themes.

6.1 Data quality and master data quality

The first question was about data quality in general. This question also served the purpose of getting in the same page with the interviewee as well as giving a quick impression of how the interviewee knows the related academic terminology. First question was formed as "What is data quality." For some interviewees the question had to be rephrased to form "What is (high) quality data" in order for them to understand where to start. After anchoring the definition of data quality, the more relevant questions could be represented.

> *"A unified set of concepts to represent the data." (1)*

> *"Data quality means that all the relevant dimensions are correct, and it is fit for the business to use." (4)*

> *"The auditability tracks who has changed data and when. It is the least used quality dimension of these, but it can be highly controlled for example via legislation." (3)*

Data quality was seen as the correctness and trueness of the data as well as how it can add value. Quality was understood to have multiple dimensions. It is safe to state that all the interviewees understood the dimensionality of the data quality but only few named multiple dimensions of data quality. This hints that the interviewees were not able to think the data quality to be as multidimensional matter as it is. This is understandable since the most easily described dimensions are those that are very clear, a phone number is in right format or it isn't. The representational dimensions of data were not mentioned at all and the contextual dimensions were summed up under the terms "fit to use" or "value adding".

> *"Data quality is valuable because the higher the quality, easier it is to derive information from it, which leads to added value" (3)*

As the meaning of data quality got clearer in the interviewees mind, the natural question arose about the importance of high data quality in general. The question was: “How important is having high quality data?”

The importance of data quality is something that is hard to measure. The discussion was aimed to get the answers from general level and to see if the interviewees would pinpoint why high quality data is important. It was stated that the importance is dependent on what the data is used for.

> *“Depends on the context data is used for. If data doesn’t matter, does the quality matter? If data is used for decision making, the data needs to be of high quality so that the decisions can be done based on truth.” (1)*

> *“Data is the water of the 21st century. The cleaner the water the healthier the business.” (6)*

So it can be seen very important or not important at all. How is it possible to know when data is important? One suggestion was the size of the organization.

> *“Depends on the organization. It can be very important. If the organization is large and the enterprise architecture is complicated, it becomes extremely crucial.” (3)*

Still, the concreteness of why it is important was very hard to pinpoint by the interviewees. One suggestion was how costly is to use it. Higher quality would lead to lower costs and improve the results.

> *“From business point of view, the higher the quality of data, and the lower the cost to use it. Data quality has a direct connection to the time costs, money costs and the comprehensiveness of the results” (2)*

> *“It’s important, but not imperative. You can manage with bad quality data, but it may become costly.” (4)*

The interviewees were not able to clearly state why it is important but they knew it really is important. Costs of low quality data and the value-adding of high quality data were the most concrete answers.

The data and its quality was seen important, but also dependent on the context. For smaller companies the importance of high data quality would not matter so much. The more complex the environment and the decisions, the more important the interviewees saw high quality data. This may origin from the fact that the interviewees worked with larger companies. Thus it cannot be stated as certain that the smaller companies don’t need high quality data. They just are not as able to hiring consultants to pay for it.

After discussing data quality in general level, the interview moved on discussing data quality from master data perspective. The first question under this theme was used to determine the background knowledge on master data and to let the interviewee and interviewer have the same view on terminology regarding to it.

Master data is a quite pragmatic concept. All the interviewees were well acquainted with master data, but it was not clear if they had the scientific definition in mind. The definitions originated from different backgrounds of the interviewees.

> *"The most critical data assets of an organization. Generally dimensional data, but it cannot be restricted to that. "(5)*

> *"Data that describes the core business entities in the real world. They are present in almost every transaction in a way or another." (3)*

Adjective "dimensional" is closely related to reporting and data warehouse modeling where the more stable data objects, which the transactional objects refer to, are seen as dimensional.

> *"The most common master data are still customer, supplier, product or item and organizational structure data. The master data of an arms dealer and hospital may be very different" (2)*

The common examples of customers and products were the most concrete way to describe master data. Often the master data is much more and it can be hard to draw the line on what is master data and what is not.

> *"Master data are core entities that are linked to the data model and which transactional data is linked to. It has effects on many things as a whole. Master data can be seen as dimensional data on which factual data refers to. That means it is stable compared to transactional data" (1)*

Interviewees had a clear view on master data and knew how it effects the organization. Still they were unable to draw a line on what is really master data and what is not. The main arguments behind master data were the stability and the fact that they are the core-entities from the business point of view. Thus it is safe to say that the interviewees understood well the most important aspects of master data. Still most interviewees did not see any difference on if master data quality differs from data quality in general.

> *"Master data is data as any data and that's why there can be no separation between the dimensions of master data quality and data quality in general." (5)*

This suggests that the master data is as any data in organization and its importance is not different from all other data.

> *“Master data does differ from any other data technically. In business sense it has larger effects and that’s why its accuracy and correctness are more crucial” (4)*

It was stated that the effects on business are more crucial. This doesn’t help in defining master data apart from other data. This also suggests that if data is important, it is master data.

> *“Errors in master data are reflected in more places and that’s why the data quality is more crucial. For example, if the address of a customer is incorrect, the deliveries or invoices do not find the receiver. In that sense the accuracy and the real life representativeness is more important to master data than data in general.” (1)*

The vast usage of data and how other data refers to it were seen the main attributes of master data.

> *“It has the same dimensions, but higher priority because of its widespread usage throughout the organization. Master data tends to be the data that has most quality problems. In general, if master data is of high quality, transactional data is it also.” (3)*

Master data then doesn’t differ in other ways that its quality is more important because it’s wide spread. It is also suggested that the master data is the data that has most problems in quality. This can also be thought in the way that master data is the most important data and that’s why it seems there are most problems in its quality since they matter more.

> *“Master data quality is managed manually more by data clerks. These data clerks need to be business users that have knowledge on the data. Data quality has usually stricter standards and its effects are larger. Master data may have different roles and if it is only used for reporting, the role may not be so large.” (2)*

Interviews suggests that master data is also the only data which quality is managed manually by the business users. It has very close relation to the business and may have a specific persons or roles attached to managing it and its quality.

Interviewees acknowledged that master data is important and it should be treated with more punctuality because of that. It has more relation to the business as all other data thus it’s more closely attached to specific roles. The master data quality then should be more important than other data but still obey the same rules when thinking of its quality. The difference arises from the specific usage and so the importance of particular data entities.

> *“Master data quality is measured by same dimensional standards as other data. In that sense it does not differ from previous answer related to data quality in general.” (1)*

So master data still is as any data but more important.

> *"It should have clear structure and universal standards throughout the organization. It is self-directing and very normalized which supports the quality of the data." (2)*

Universal standards in the organization suggest that the organization defines the rules for master data quality. Its importance is not as much generic as any data but more related to the specific organization.

It was acknowledged that master data quality should hold universal standards in the organization. It's the end product of a good data modeling and an effective enterprise architecture. In the other hand master data has its own standards set by the organization. It tends to be more intertwined with the business and in that sense it can be stated that master data always depends on the business and no generic solution leads too far.

> *"It's the end product of a business based data modeling where metadata is automated as well as possible. It has high timeliness so it's usable across the architecture when it's needed" (3)*

Interview suggests possible automation in master data lies in the metadata. Metadata is data about data and if business steers how master data is seen, and metadata of master data depicts how master data is used in organization. The metadata so seems to have a role in leading master data management more close to the business.

6.2 Acceptable quality and quality problems

Interviewees were implied by the questions that master data quality cannot be perfect, but there is a level which is good enough. The costs of master data quality improvement should be in line with the benefits.

Interviewees have experience from real life master data management and thus had seen many different issues which were acceptable and which were not. From this they should have an idea about an acceptable level of master data quality.

> *"Acceptable level is the intersection of cost and profit. The point where resources used to improve data quality cost less than the business benefits that follow it." (6)*

> *"It's the threshold value where cost meets the benefit." (4)*

This brings back the practical definition of costs and benefits described earlier. The challenge in this still stays the same, since the costs and benefits are very hard to calculate.

> *"Data quality is adequate when that business can operate normally. For some organizations the adequate quality is higher than for others. Some feel that data quality needs always be perfect." (1)*

Organization itself sets its standards in what level master data quality should be. When organization operates normally, the data quality is good enough. Organization tend to try to improve which means that the normal level of operation becomes more challenging day by day. This suggests that the data quality standards become more and more challenging.

> *"It depends on the role of master data in the company and what it is used for. It also depends on how well the processes need to be automatized."* (5)

Automatization is an aspect that would improve the performance of data flows and data management. It could also lead to higher data quality.

> *"Roughly at such level that 90% of the basic organizational processes can be fully automated." (3)*

So the high data quality enables higher level of process automation and the adequate level is where only a 10% manual work is needed. For more advanced organizations the automatization percentage may be higher a lot and for smaller and simpler organizations much lower.

The measuring of acceptable master data quality was mainly the cost and benefit view. The interviewees stated that it is impossible to measure the cost and benefits, but there can be estimates that can be based on the experiences of bad data in the current organization. Another view that can be seen interesting was how master data quality is adequate is adequate when it enables a certain level of automatization of the basic organizational processes. The matter still seemed a little vague and examples of acceptable or unacceptable issues would clarify the point.

Interviewees stated that they have had calls and emails marked urgent and thus had experience on when master data is preventing business from happening. There were lot of examples, but in general the role of critical issues was clear.

There are a broad range of issues from minor to very critical. The interviewees should state what they think are the critical issues that should be avoided and fixed immediately.

> *"It's unacceptable if the address of a customer is wrong and so the invoices are delivered to a wrong address. That would lead to problems in getting the invoice paid on time. A customer has a duplicate entry in data which differs a bit and it is impossible to know which one is the correct one" (1)*

The most straightforward example noted was easy to understand and it is true universally in companies which use addresses in invoicing. This is very obvious and does not offer any value by itself.

> *"Acceptability depends purely on the business need. The business needs determine what is acceptable or not. In retail industry, the successful sales event determines what data needs to be used and how correct it needs to be." (2)*

As previously discussed, the standards of the master data quality are set by the business needs. The success of a sales event is more relevant example, since it includes the previous example of correct address but also implicates that there are numerous other aspects that affect in the outcome.

> *"What is the core attributes used in business transactions? It cannot be said in general level that something is always unacceptable or always acceptable in most cases?" (3)*

> *"There can be low priority attributes in the data that do not need to be of high quality."(4)*

There is no one right answer. Some customers for example may not be as critical as others so it is hard to state that the correctness of the billing address of every customer is equally important.

> *"The primary keys or other key identifiers need to be correct in the data" (4)*

Technically there can also be a very simple answer on what needs to be correct. This may be true but doesn't help since the identifiers are based on business needs and the cause of those being incorrect would also be related to a business process rather than technical one.

> *"The most critical attributes and dimensions of a master data entity should be determined by the business needs. The technical dimensions of these can in most cases be automatically monitored" (1)*

Interviewees had a clear vision on what are the critical master data quality issues. The aspects that were viewed as most important were the business needs and what the critical components are in order for the business to operate. It is also important to decide what to include in the master data. If too much information is included, there may be data that is rarely used. This leads to the point that not all master data quality issues are so important. This could make it harder to evaluate the overall master data quality from the business point of view.

> *"A small variance on how something is written is not so crucial. Lack of information in hierarchical relation of data may be accepted. Few percent of wrong data is accepted in most cases for the business to still run smoothly." (3)*

Small variances in natural languages are possible to be noted and corrected with today's natural language processing tools. This includes higher level mathematics, fuzzy logic tools and pattern matching but as stated, the importance is not so big. This may relate to the fact that most postal items are manually handled by people that can interpret small infractions in written text. As things become more automated, the processes need to take account that there may not be people interpreting the data anymore.

The conclusion and idea that many interviewees stated was that there often are attributes in the entities regarded as master data that are not important and thus should not be modelled as part of master data. They underlined the importance of data modeling both in master data and also in the operative systems so that the low importance and high importance data attributes are not too intertwined in the data model.

Master data problems originate from the multiple systems. There are bigger and smaller problems and the goal was to determine what the interviewees felt to be the biggest problems that need more attention. Often the most noticeable problems are the technical problems, but they usually originate from issues that are non-technical, such as people working against policies or lack of planning of the process to support business needs.

Interviewees had very much to say about the biggest problems. Same problems were described differently from separate points of view, but the underlining problems stayed the same.

> *"There are multiple systems and data and its quality should be same everywhere. Problem becomes concrete when there; are multiple processes where data is processed, there are multiple people who process it and multiple ways it is processed. People and machines do what they are told, and that is why most of the problems' roots lie in the lack of process."(4)*

People were seen as a weak link in specifying the biggest reasons in data quality problems. People are unable to follow the processes from various reasons. As there were no clear answer on why people can't follow the processes, it is safe to state that many processes may not be aligned with the business well enough. This would prevent people from being able to follow the processes and would suggest that the processes themselves are not suitable.

> *"Some of the problems are still purely technical, for example data masses may grow so large that they are hard to process with the tools and systems selected. Technically the biggest problems are duplicates, errors and timeliness" (1)*

Technical problems were seen easily recognized and straightforwardly corrected. As most of the reasoning did not originate from the technical but from business point of view, the assumption is that the technical problems are manifested from the underlying, more business related, and problems.

> *“Diverse amount of systems linked to each other. Integration architecture is not well planned and is done “quick and dirty”. This leads to data not being equal between systems and the data and information management is done in multiple different places. (2)*

The origin from the technical problems can also be the system architecture. The technical decisions may not have been selected to be able to dynamically support the evolving business but rather to solve the problems and needs most notable at the time. This calls for more strict policy in managing the architecture and making decisions that able the constant changes needed.

The dynamic complexity of the environment and architecture were found to be the source of problems. The people added to this complexity as actors, underlined the problems. It was found hard for the processes, people and systems to evolve alongside the business and its needs.

After assessing the biggest problems, it was natural to continue assessing the reasons behind these problems. Most interviewees seemed to be able to see the root causes clearly in their minds.

> *“The root causes are bad input systems and people using them. If the input is not forced to follow process, it results in bad data. One example of these input problems are free text fields where users can write anything and it is easy not to follow process. Organizational growth leads usually to master data management challenges since business processes and systems evolve and master data management process rarely keeps up.” (3)*

It was noted that everything starts from the input of data. That is the moment when real world subjects are described in to the system. This is critical since the correctness of input of data effects on how well the system can reflect reality.

> *“Process. The enterprise architecture and the system maps are complex. Development is done gradually over the years partly in siloes so they are hard to make work well together. When people are involved, there are bound to be errors. That’s why free text fields are often a bad decision. When there is no clear process, people do whatever they feel like doing. (1)*

The complexity was stated to origin from siloes and decisions that are not meant to support the whole entity of a company. People as actors are responsible for the most errors but as the people change, the role of process becomes more and more relevant. It was also noted that the process is what sets how people should interact with the data.

> *“The people, the clearness of data governance goals and process related to them.”(5)*

Data governance was also mentioned. It was seen as the origin of processes and the starting point of making organizations master data more aligned with the business. In that sense every other process or people related matter goes back to the assignment to governance.

In the discussion, interviewees stated that people and the process were the main reasons the master data quality is not always of high level. People are behind every decision and all the systems, but the biggest flaw of people's action was the input of data and following the process. Although it was noticed that the process is not always perfect, that's why people have hard time working according to it.

After having the idea in mind of the biggest reasons and root causes of poor data, it was necessary to address how they affect business. The underlying assumption based on the previous answers was that they do affect at least in the form of costs and failed transactions.

> *"Depends if master data is used in operative systems or only in reporting. If customer information has errors, the operative system is unavailable to invoice or order. In reporting the numbers would be wrong which would lead on decision making based on false information." (1)*

The role of master data would depict how it affects business. It could prevent the operating systems from performing successful operations and it could lead to bad decision making via the errors in the reports. This is very aligned with the previous notions of master data and its quality being very dependent on its role for the organization. How business sees master data affects how business reacts to poor master data quality.

> *"Either information is not available and it cannot be supplied or the process slows to manual labor. If there is information available, it is not timely enough. An invoice can be sent to a wrong address or a shipment may be delivered too late. That leads to large overhead costs. (2)*

The manual labor and the overheads related to it would be the lead outcomes of low quality master data. This again links master data quality the automatization aspect of the business processes.

> *"In worst cases the business processes do not run which leads to unavailability to do business. Even worse scenario is that there are large costs which make the business unprofitable. Laws could also be broken which might lead to catastrophic consequences." (3)*

Business effects were well realized among the interviewees, but there was no real solution how to asses these effects more closely. One of the key effects were the decision making based on false data that did not represent reality as well as it could. Other effects were related to how bad data quality prevents the real life processes such as logistics or invoicing from succeeding. There was also view of overhead and possible legal issues originating from the data quality problems.

6.3 MDM hub and best practices in supporting data quality

MDM Hub plays an important role in information architecture. Not all interviewees were familiar with the concept of MDM hub. Before presenting the questions related to MDM hub, the concept needed to be introduced. Different possible MDM hub architectures were briefly shown to the interviewees. The hub type that these questions relate to are the coexistence hub (for example Dreibelbis et al. 2008) that is also entitled hybrid registry (Loshin 2010).

In reality there are many different types of master data management hubs. In this case the focus was on the general principle of hybrid MDM hub. In this case it would mean that it could enable the advantages of the physical hub but would not restrict the company to use it in every master data related transaction.

> *"Via MDM hub, data is distributed to connected systems. Hub should take care of communicating to other systems that their data needs to be updated. That way timeliness stays high. If data is distributed, it should be validated and correct. MDM hub would distribute the golden record" (1)*

The obvious role of the hub would be the distributor of the most correct and comprehensive depiction to the organizations master data. It would also have the role as the validator of the correctness and comprehensiveness of the data.

> *"Data would come to the MDM hub from the system that is responsible about its maintenance." (2)*

The management and maintenance of the data could be done in the source systems of data which would make it easier for the data collectors to work with it with the systems that they are most used to.

> *"MDM hub should be responsible about timeliness and linking the primary keys to corresponding values. MDM hub should help enriching the data." (3)*

The hub would have a large role in enriching the data effectively in one place. It could have access to reference data related for example to geographical entities which would help in validating address data.

For most interviewees, there were many ideas of the role of MDM hub. It should be the distributor of the golden record, which it achieves by consolidating and enriching the data from source systems. This would describe the role of the hub in more general level but in order to get more concrete answers the role from master data quality point of view was queried. The next question was straightforward and the goal was to get examples of how MDM hub could be used in improving the master data quality.

> *"It could provide data validation on entering data. It provides a central point for data input where it can restrict the form in which data can be in. Data validation and harmonization could also be done upon entering. In practice data quality would need constant monitoring and improvement" (5)*

This states that MDM hub would be the place of enrichment and validation, but also offer a place for entering the data. Controlling how the data can be entered to a system could be done in same place which would help restrict the format and so tackle the problems occurring of human error.

> *"Data maintenance would be centralized to support the data quality process. The maintenance could only be done in MDM hub which is the most logical place. In either case, the changes in data should only be needed in one place and the update would be distributed everywhere via MDM hub. MDM hub would keep track of metadata, such as the source system and the cardinality between data sources, so it is clear which source system overwrites other source systems data." (2)*

The centrality aspect of MDM hub would help in supporting the data quality process. Maintenance would be done in one place which would help distributing in everywhere else. MDM hub would offer the ability to keep track of metadata such as the source systems and cardinalities of data. This would help in achieving the golden record by helping to prioritize the data with highest cardinality.

> *"It helps clarifying the information architecture" (3)*

One notable thing that MDM would offer is making the information architecture clearer and simpler. It would be easier to manage the data one place and flow it through the central hub. In the case of hybrid hub the clarification would mean that it would be clearly defined which data goes through the MDM hub and which does not.

> *"It also supports the acceptance workflows for the most critical data assets."(4)*

The answers were optimistic in the sense that MDM hub was seen to have many possibilities in improving the data quality. Validation, enriching and manual maintenance would take place in MDM hub. Metadata would help support these tasks.

There are many practices that take place in the everyday operations of a MDM. Some of the practices are considered to be better than others. The best practices in this theme are meant to reflect the practices that every MDM project using MDM hub should exercise.

The next goal of the interview was to get list of things or actions that would help in improving the data quality and the process related to in in MDM process. Expectation was to get a list of actions how to support the data quality process in MDM hub. Interviewees had many ideas on what could be done and stated that many of these things are forgotten in the everyday project oriented way of work.

From the first interview, it was clear that the business and technical aspects should be both mentioned in the questions, since easily the interviewees locked to one of these perspectives only. For most, the technical perspectives were more concrete and they tended to mention them first. That's why business related best practices needed to be separately asked.

> *"Ownership to distinct data domains. The owners would have responsibility of the data quality and would have to make the most important decisions affecting the data. Shared responsibility seldom works in master data management."(1)*

The most noted thing that arose was the responsibility and ownership of the data. Ownership would relate to those who have the largest responsibility in a specific area of data such as customer domain.

> *"Data steward would be the responsible technician of the data and the owner should be a representative of the business. The most important thing is to have the business vision of the data. It is very rare to have both business and technological readiness."(1)*

The responsibility could also be divided in other ways such as one steward responsible for improving the data quality from the operations point of view. Another steward could be more responsible of the data quality in the technical point of view.

> *"Data quality process should be determined and there should be guidelines to assign the CRUD user rights."(4)*

Data quality processes seem to lack from existing from the interviewees point of view. The creation, read, update and delete operations in data should be a concrete way to assigning the business related responsibilities to the data.

> *"The information architecture should include MDM hub and an integration layer on it." (2)*

The architecture was seen as a key component in having high quality data. The hub would offer the data to the integration layer which would act as a service that other systems could use.

> *"MDM hub should be integrated to a data quality component that supports improving the data quality. This component could provide automatic data quality*

> *assessment and profiling. It would also provide tools to help automate the improvement process" (3)*

The interviewees implied that there should be one place to asses and profile the master data and that the place would be the MDM hub. The central role of the hub would mean that there could be advanced automatization built in the top of the system. The profiling of the data quality would offer meters that could be followed and decisions could be made on that. Gamification could be built upon the meters so different data teams could compete on who has the highest quality data from the meters point of view. This would make the task of data harmonization more interesting to the people responsible.

> *"The data MDM hub should be responsible of maintaining the integrity of a data model. It should also store the metamodel of the data model and other related metadata." (3)*

The technical integrity and the storing of metadata was a task that was seen relevant for the MDM hub to execute.

> *"A business data dictionary could be implemented to the MDM hub." (4)*

A business data dictionary was seen as one concrete tool that could be implemented to close relation to the MDM hub. It would store the entity and attribute business descriptions. These would help using the correct terms in the data modeling and also be the place from where the correct terminology could be checked to improve everyday communication

In the end of the interview the interviewees had restored as much knowledge to their minds as possible. Then it was the right time to sum up what are the really most important things that need to be done to achieve the goal of better data quality in an MDM hub.

> *"Most important would be having more and clearer roles and responsibilities."(1)*

Responsibilities and roles relating to them was seen as the most important single factor that would improve the data quality and make master data management more successful. This also relates to putting more resources in master data management as it would mean that the master data management tasks would take more time of the existing personnel and could also lead in hiring new personnel to do these tasks.

> *"More named owners of data domains. Data responsibilities to align the business responsibilities even more." (4)*

Ownership was seen as a part of the responsibilities but also as an individual aspect that would have a large effect on how master data is managed. When there is a person who is accountable for the quality and the errors that follow from lack of it, the problems tend to be solved with more motivation.

> *"The effects of master data quality should be understood better. The costs of low data quality can be great and the competitive advantage could come from high quality master data. (2)*

The effects of master data quality should be understood better from the interviewees' point of view. This relates to education of the personnel working with data and most crucially the persons with power to assign more resources to the management the quality of master data.

> *"Data dictionary, data quality component, data quality automatization" (3)*

These all relate to the technical aspects that are supported by the availability of metadata. The automatization and quality components should be more relevant as technology advances and the architecture and services available make it easier to implement them.

> *"Data governance program would be the starting point defining the goals. Every other choice should support that" (5)*

In the end, the best practices were summed up to only the most crucial mentioned with the most interviewees and highest emphasis. These were roles and responsibilities, ownership, process and data governance guidelines.

7. DISCUSSION

In this chapter the theory and the empirical part of the study meet. The empirical part is reflected to the theory and the most important findings are discussed in more detail. The structure of this chapter follows the supporting research questions, answering them first and the main question after that.

7.1 Master data and its quality

The theoretical study concluded that data quality is about data satisfying the requirements of its intended use (English 1999, Redman 2001, Orr 1998, Wang 1998). In the interviews the first question was to determine how interviewees viewed data quality and if the interviewees were aware of this definition. The interviewees did not have the definition clearly in mind in the first question, but the answers to later questions showed that they acknowledged it. In later questions it was stated that “master data quality needs to be high enough for the company to be able to operate”.

In the theory section it was also assessed that data is a crucial asset for business to operate well. As stated in chapter 3, low data quality might lead to very high costs that can represent 8-12 percent of revenue (Redman 1998, p.80).   The interviewees were very acknowledging of the high cost of poor quality data and they all found it to be a critical factor in organizations.

The interviewees did very clearly perceive that data quality is not a monolithic subject, but that it can be divided to dimensions. Accuracy and timeliness were mentioned most among the interviewees. Also the effect of added value was implied by answers telling that low quality data was harmful for the operation of the organization. In general, from top nineteen data quality dimensions found in the literature in table 2 at least ten of them were mentioned or implied upon.

Another goal for the study was to determine what master data quality is. Interviewees referred it following the same rules as data quality in general, but it also had higher emphasis in some dimensions and was critical. Interviewees also understood it had many dimensions, but the real standards for quality depends on the business. This supports the theory where Haug & Arlbjorn (2011) in their meta-analytical study conclude that the fitness for use is the main measure on data quality in vast number of publications. Another point in interviews was that master data quality hardly can, neither should be perfect. Morris (2012) supports this saying that organization does not need or does not want to pay for perfect data.

7.2 Causes of poor quality master data

Haug & Arlbjorn (2011) concluded a literary review to determine a list of barriers to master data quality. The following five major barriers were detected: Lack of delegation of responsibilities for maintenance of master data, lack of rewards for ensuring valid master data, lack of master data control routines, lack of employee competencies, lack of user-friendliness of the software that is used to manage master data. In the interviews, the biggest problems and the root causes were queried upon. The answers were mostly related to the complexity of the technology aligning with the business processes and the people working with them.

The lack of delegation had the greatest effect to master data quality in the Haug & Arlbjorn (2011) review and this interview supported this by implicating it. When asking interviewees about the problems, the roles and responsibilities were mentioned only by one out of five responses, but when asked for the resolutions five stated the roles and responsibilities being the one of the most important factors to improve. Lack of unified processes for master data was one the most mentioned problems in the interviews mentioned by three out of the six. It was stated that “People and machines do what they are told, and that is why most of the problems root from lack of process”. In the Haug & Arlbjorn (2011) review this was referred as lack of master data control routines. It was also mentioned in their study that the implementation of control routines independently supports the fulfilling of the data quality responsibilities. From the cost point of view Haug & Arlbjorn (2011) imply that even the simplest routines pay themselves back by lowering costs.

In the interviews it was stated multiple times that the input systems must force the user to input data correctly. In the Haug & Arlbjorn (2011) meta-analysis noted the same matter, but emphasized the user friendliness Smith and McKeen (2008) refer to this as complexity of it solutions. All in all, the goal is to get the users input data into order within the systems. User friendliness could lead to the same conclusions as forcing of the users.

The Haug & Arlbjorn (2011) meta-analysis also had the lack of employee competencies and the lack of rewards listed as the major barriers. In the English (1999) study to which Haug & Arlbjorn refer, the rewards are linked to the incentives of improving data quality, not actual rewards. In the interviews it was mentioned that “stricter responsibilities lead to higher motivation to ensure data quality”, which implicates that the incentives must origin from the role and the responsibilities attached to it. Employee competence was not mentioned in the interviews, although one can interpret this being hinted by the interviewees when implicating that the responsibility assignation must be done fittingly.

7.3 Role of MDM hub

MDM hub was seen as the central piece of master data management in the information system architecture. Microsoft SQL Server Master Data Services being the reference solution in this thesis, the interviewees found it offering the most important tools for master data management.

The MDM Hub is the central repository for managing master data. Dreibelbis et al. (2008) suggest that the major building blocks that MDM Hub offers are the quality and lifecycle management services. These offer the tools for the authoring the data and performing (Create, Read, Update and Delete) operations. It can also author business logic or business rules thus supporting the business even further. It also has the tools for enforcing data quality rules, assessment and harmonization of the data. In the interviews the role of the MDM hub was seen as the center of the integrations. It would offer the “Golden Record” of the data and the tools to maintain and manipulate it. It also would help to assign company-wide business rules to the data as well as acceptance workflows. The central and physical nature of hybrid MDM hub would be a necessity for these operations, since they would be impossible to perform in a virtual MDM repository. It was also noted that Microsoft MDS would offer all of the necessary tools.

The MDM hub was seen as a central piece of an information architecture in the interviews. One interviewee stated that “The information architecture should include MDM hub and an integration layer to it.” This suggest an SOA based approach to the MDM. This view is supported by Allen ja Cervo (2015), Loshin (2010) and Dreibelbis et al. (2008) as seen in figure 6.

The input of data is a central part of a Hybrid MDM hub. This does not mean that the data input is not done elsewhere. The interviews noted that “The improvement of the data quality should be done where the data is inputted.” This suggest that the other systems should have their data input controlled too in addition to the MDM hub. This means that the MDM hub should not be seen as a silver bullet to data quality issues, but as a supporting part of architecture offering multiple tools for the implementation of the business based master data management processes. As it would be easy to outsource all the data management to the hub, it would lead to the data being managed by persons with lesser knowledge of the contextual dimensions of the data.

As the MDM hub offers a lot to improve the data quality, it should also be seen as a component to help understand the effects of data quality better. One error in input to the central hub, from where the data is integrated, to the operational and analytical systems would demonstrate the importance of data quality in general. This way the MDM hub would help to start determining more comprehensive data management processes in the company as well as help assigning realistic tasks for the people responsible for the data. This would help in the creation and assignation of the roles and responsibilities for different areas or domains of data. This would suggest that the MDM hub would offer an iterative approach to master data management enabling the gradual implementation of the processes, roles and responsibilities related to the data. This seems counterintuitive to the stepwise approaches represented in the MDM literature by for example Joshi (2007) and Vilminko-Heikkinen and Pekkola (2013) who add to Joshi's original approach.

In the Vilminko-Heikkinen and Pekkola study maintenance was not included. This study did not exclude maintenance, and tries to underline what happens after the last step of "Defining MDM applications characteristics" and implementing an application such as MDM Hub. The interviews suggest that the MDM hub supports many of these steps even though they are prerequisites for its implementation. This leads to the conclusion that MDM hub plays an important role in the constantly changing information architecture offering the support for the establishment of new MDM functions. This suggests that maintenance is an ongoing process of the establishment of MDM based on the ongoing process of changing business needs. And that these steps are also valid in the maintenance phase where MDM hub could support the execution of the steps.

7.4 "Best Practices"

The goal of the study was to answer the main research question.

"What are the key factors in supporting data quality in hybrid MDM hub?"

Answering this question would lead to the discovering of a list of "practices" or factors to help maintain and improve the data quality in MDM Hub perspective. These should be based on the literature of the best practices and take account the role of a MDM hub solution. The list should also be supported by the findings on the interviews that form the empirical part of this study.

The most critical factors in the interviews was the roles and responsibilities and the ownership of the data. These are mostly factors that originate from the data governance side of data management, but they can be supported by a more technical measures such as centralized MDM solution.

After that, the most critical factor was seen to be the data quality processes. These are based on the data governance goals and are to support the alignment of business processes and information processes. The MDM hybrid hub solution would support these processes, by being flexible architecturally, to align the data flows with the business processes.

Data governance was also a central aspect. The MDM hub solution does not straightforwardly help in the governance but it helps in reaching the governance goals in practice.

The metadata management and data quality automatization was also seen important. The MDM hub solution gives a logical place to store metadata and to maintain a data dictionary. It also offers tools and services for data quality automation by using metadata.

8. CONCLUSION

This chapter concludes the research summing up the most interesting findings. The results are compared to the research questions and a concluding answer is given. After that the research and its methods and their reliability and validity are evaluated critically. This is relevant to address the weak points and limitations of the research process. Lastly, the further research opportunities are discussed briefly.

8.1 Summary

The objective of this research was to supply a list of most relevant factors that are to be taken into consideration when achieving sustainable data quality in a MDM hybrid hub based architecture. As a result a list of remarks that need to be taken in concern is produced. The research objective was presented in the form of the main research question. The supporting research questions were to support the main research question by supplying a context and a theoretical background for it. The first supporting research question was designed to answer and define what really is the master data of an organization and how it is managed. Second question was to define data quality and link it to the master data perspective. Third question was to define the key concepts of a hybrid MDM hub – architecture, which is the reference architecture of this thesis. Fourth question was to define the roles behind the master data quality. Lastly, when these supporting questions were answered, the main research question of the key factors in supporting the data quality in a hybrid MDM hub.

Which data of organization is really master data and how is it managed?

Master data is the most relevant dimensional data used by company. Generally customers and products are the main domains which can be considered as master data. The master data can still be anything depending on the company. In a municipality the citizens are the master data and in a hospital the medical equipment can be considered as master data. The notable traits of master data besides its importance to the company's business is its stability which is related to its dimensionality. After identifying master data, it can be managed. The management of master data starts from the very top of an organization where the data standards and processes are aligned with the business processes and standards. The goal is to make the data work towards achieving the business goals and vision of the company. The data governance is implemented in the everyday operations of a company by master data management which is a technology-enabled discipline of making IT work towards the uniformity and quality of the master data assets.

What is data quality from master data perspective?

Quality of master data is same as quality of any data as it has the same dimensions and can be divided in the same areas such as intrinsic and external. Master data context emphasizes some dimensions more than others. Its believability, value-addedness and relevancy are the most relevant dimensions as those are very important from the business perspective thus they are contextual. In addition, accuracy plays a large role in master data, since its accuracy has a compounding effect to the accuracy of data in the linked systems. Accuracy itself is the most relevant of the intrinsic dimensions and it can by itself prevent from succeeding if it has low quality. From the end users point of view the representability and interpretability play a big role how the data can be understood by the end user and a possible decision maker. This can lead to big business effects by itself.

What are the key concepts of hybrid MDM hub –architecture?

MDM hub is the technique master data is physically and logically stored in the systems architecture. It can be fully virtual where it is distributed between multiple systems and only a virtual register of the entities is stored. It can also be fully transactional, where every transaction goes through the MDM hub. The first option does not offer the possibilities of a centralized data management and the latter leads to very strict environment which has difficulty adapting to the constantly evolving architecture. It can also make a single point of failure which can be detrimental to the business. Thus the happy medium is the hybrid MDM hub offering the tools and possibilities of a centralized data management without the strictness of making every transaction move through the hub. The hybrid MDM hub works with the ESB (enterprise service bus) offering real time master data synchronization to the most time-intensive systems. It also offers the mass propagation of master data with batches which can be created for the receiving systems need by ETL (Extract, Transform and Load) processes. In the hybrid MDM hub the business can be comprehensively modelled to entities depicting real life business entities. The highly normalized fashion of modeling master data in the hub supports the data quality and ensures that the maintenance can be very effective. The hub also acts as a natural place to perform data quality automatization tasks and gives tools for the maintenance of metadata definitions as well as business data dictionary.

What are the roles of master data quality management?

The most common roles are the data quality council, the data owner, the data steward the data collector and the data consumer. The quality council is responsible of the quality in the higher level and assigns initiatives to improve the data quality to support the business processes. The owners are responsible for specific data and thus responsible for implementing the initiatives in practice with the help of data stewards. Stewards can be of a technical or business role. The business stewards stay in conversation with the business to serve its needs whereas the technical steward is responsible of the technical aspects of data quality and its maintenance. The collectors and consumers are the everyday creators and users of the data who also should have responsibilities in keeping and improving the data quality at high level.

What are the barriers to master data quality?

The quality of master data is the product of many factors. If one of these factors is in poor condition the overall master data quality will end up in poor condition. Everything starts with the understanding the value of data, and specifically, master data quality. The understanding leads to the motivation and the support from the high level executives who have the power to make change. As the data goals are aligned with the business goals, it is time to put these high level wishes in to action. The real life implementation of the standards and rules make the backbone on which the data quality is dependent upon, the data quality process. This may be a strict or a well-defined process as well as a process that is derived from the needs of other processes such as manufacturing and billing. In either case, following this process to reach business goals has many obstacles and barriers. It all starts with the people. People create the data, people manipulate, interpret and use the data. It is well known that people are prone to error. According to Murphy's Law, anything that can go wrong, will go wrong. This is more valid with people and data than anywhere else. That is why everything needs to be designed so that the effects of human error are diminished. This starts by making people responsible. Responsible for the data they input. Responsible for the data that is relevant to their work and which they know the best, and by requiring them to be the owner for the data, making the data theirs. The other ways to diminish the human traits in data management is to make systems that intuitively direct towards the right decisions portrayed by the data quality processes.

What are the key factors in supporting data quality in hybrid MDM hub?

The barriers that must be overcome in order to reach the goal of high data quality. It begins from the top where governance is initiated. When reached to the reality where there is a MDM hybrid hub integrated to the architecture, it can be used to reach the goals of high master data quality. The roles and responsibilities are assigned in the business functions and they are modeled in the MDM hub with user rights and acceptance workflows. The needs of specific data domains are discussed with the business owner of that domain and the required business rules and logic are applied to the MDM hub. As the business needs and entities become clearer to the data stewards, business and technical, they can be modeled to support the business. A business data dictionary can be stored and maintained by the business personnel so that everyone understands the used vocabulary in the same way.

8.2 Evaluation of the study and further research

To ensure that the study was conducted in a trustworthy fashion, it needs to be evaluated. This can be measured in the terms of reliability and validity. Reliability refers to the extent to which an experiment, tests or other measuring processes yielded the same results in repeated trials (Carmins & Zeller 1979, p.12). Validity indicates how well the metric measures what it is designed to measure. (Carmins & Zeller 1979, p.12). The reliability of this research can be perceived as good. The data collection and analysis were systematic, but in the case of interviews and analyzing them there are always some biases. The theoretical background was derived from the best available literature sources which lead to lower possibility to reference bias.

When considering the research choices for this research, they ended up being suitable. The choice of pragmatism to the research philosophy supported the qualitative nature and the goal to product findings that pragmatically would answer the research question offering tangible ways the most critical factors can be assessed. The inductiveness of the research made it possible to derive from the literature as well as the empirical part of the study to form results that are based on theory and practitioner expertise. The case study was chosen so that it would help in defining the area of interest to strict real life case which made it possible to study many factors related to it. The cross-sectionality was self-evident since this thesis had a strict time limit and there was no possibility to lengthen it more than necessary. The qualitative data collection and analysis methods done via interviews was fitting since the area is very practical and there are very few quantifiable ways to measure anything in this field.

The conduction of this study took longer than expected, almost a year. It started with the gathering of sufficient empirical knowledge from various literary sources. This would form the basis for the research. The chapter two in this study introduced master data and the most relevant concepts around it including master data management and data governance. It was also distinguished from the other data types such as transactional data. The third chapter introduced the concept of data quality and its dimensions and linked it to the concept of master data. It also highlighted the barriers that prevent high master data quality from existing in organizations. It also motivated to the subject by showing how much poor data may cost to organizations. The fourth chapter was the most practical in the theory part of the study. It linked master data management to the system architecture of the enterprise and introduced the concept of how the master data management is centrally done in a hub. The hybrid hub was the type of hub specifically chosen for the reason that it was seen as the reference architectural method in the case study. Its validity was also briefly explained by comparing it to the other types of MDM hubs. Lastly in that chapter the reference solution for hybrid MDM was introduced in the form of Microsoft product SQL Server Master Data Services or MDS in short. This is relevant as it is the reference technology behind the case study.

The gathering of the scientific literature was partly easy and partly hard. The basic concepts of master data and data quality had clear definitions in the literature but the more practical concepts of MDM modeling and architecture was harder to find. That is why many practical level books were referenced in some parts of the theory. These books were selected by the references made to them by Google Scholar. There were also vast number of books available that were relevant, but had no scientific references. No real literature concerning the reference solutions of MDM hub and Microsoft MDS was found. This is not considered as a problem since the subject is very specific. This also validates the motivation behind making this study.

The empirical part was not as hard as it was time consuming. It was clear that when conducting the interviews in a very practical area of expertise, while trying to tie all the questions to theoretical backgrounds, would be challenging. It was helpful that the theory was mostly made before the interviews since the interviews could be interpreted better and the linking to the theoretical definitions could be made more easily. The interviews were analyzed in quick succession and the results of the analyses were opened in the thesis. Some quantitative elements were used when analyzing the interviews when measuring the frequencies of specific core factors seen critical from the study’s point of view. If more validity and reliability would have been required, the interviews could have been more structured and there could have been much more of them.

When thinking about future research, the results of this study can be found as a useful reference. They could be used in making a process framework in improving data quality in the specific context of a MDM hub. It could be utilized in a company and thus its validity could be tested. This can also help in delving deeper into the master data architectures and how the different hub solutions could help tackle different master data related issues in companies. In general this can help as being a primer and supplying relevant studies in the area and showing some linkage between the master data practitioners and the scientific world of data quality and management.

9. BIBLIOGRAPHY

Allen, M. & Cervo, D. 2015. Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice. Morgan Kaufmann.

Ballou, D.P. & Tayi, G.K. 1989. Methodology for allocating resources for data quality enhancement. Communications of the ACM 32, 3, pp. 320-329.

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

Berg, B.L., Lune, H. & Lune, H. 2004. Qualitative research methods for the social sciences. Pearson Boston, MA.

Berson, A. & Dubov, L. 2007. Master data management and customer data integration for a global enterprise. McGraw-Hill, Inc.

Cappiello, C., Francalanci, C. & Pernici, B. 2004. Data quality assessment from the user's perspective. Proceedings of the 2004 international workshop on Information quality in information systems, ACM. pp. 68-73.

Carmines, E.G. & Zeller, R.A. 1979. Reliability and validity assessment. Sage publications.

Cervo, D. & Allen, M. 2011. Master Data Management in Practice: Achieving True Customer MDM. John Wiley & Sons.

Davenport, T.H., Harris, J.G., De Long, D.W. & Jacobson, A.L. 2001. Data to Knowledge to Results: BUILDING AN ANALYTIC CAPABILITY. California management review 43, 2, .

Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., van Run, P. & Wolfson, D. 2008. Enterprise master data management: an SOA approach to managing core information. Books24x7 Edition, Pearson Education.

Dyché, J. & Levy, E. 2011. Customer data integration: Reaching a single version of the truth. John Wiley & Sons.

Eckerson, W.W. 2002. Data quality and the bottom line. TDWI Report, The Data Warehouse Institute .

English, L.P. 1999. Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Books24x7 edition, John Wiley & Sons, Inc.

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

Fisher, T. 2007. Demystifying master data management. CIO Magazine 30, .

Friedman, T., Feinberg, D., Beyer, M.A., Gassman, B., Bitterer, A., Newman, D., Radcliffe, J., White, A., Paquet, R. & DiCenzo, C. 2006. Hype Cycle for Data Management, 2006. GartnerGroup Research, July 6.

Friedman, T. 2007. Best practices for data stewardship. Gartner, December 3.

Haug, A. & Stentoft Arlbjørn, J. 2011. Barriers to master data quality. Journal of Enterprise Information Management 24, 3, pp. 288-303.

Hirsjärvi, S. & Remes, P. Sajavaara Paula (2004) Tutki ja kirjoita. Kustannusosakeyhtiö Tammi.Helsinki .

Hodkiewicz, M., Kelly, P., Sikorska, J. & Gouws, L. 2006. A framework to assess data quality for reliability variables. In: Anonymous (ed.). Engineering Asset Management. Springer. pp. 137-147.

Hodkiewicz, M. & Pascual, R. 2006. Education in Engineering Asset Management–current trends and challenges. International physical asset management conference, pp. 28-31.

Huang, K., Lee, Y.W. & Wang, R.Y. 1998. Quality information and knowledge. Prentice Hall PTR.

Joshi, A. 2007. MDM governance: a unified team approach. Cutter IT Journal 20, 9, pp. 30.

Khatri, V. & Brown, C.V. 2010. Designing data governance. Communications of the ACM 53, 1, pp. 148-152.

Lee, Y.W., Pipino, L.L., Funk, J.D. & Wang, R.Y. 2009. Journey to data quality. The MIT Press.

Lee, Y., Pipino, L., Funk, J. & Wang, R. Journey to Data Quality. 2006. Cambridge, MA, USA: Massachussets Institute of Technology .

Loshin, D. 2013. Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph. Books24x7 Edition, Elsevier.

Loshin, D. 2010. Master data management. Morgan Kaufmann.

Madnick, S.E., Wang, R.Y., Lee, Y.W. & Zhu, H. 2009. Overview and framework for data and information quality research. Journal of Data and Information Quality (JDIQ) 1, 1, pp. 2.

Messerschmidt, M. & Stüben, J. 2011. Hidden Treasure-A global Study on master data management.

Microsoft Corporation. Microsoft Developer Network. [WWW, Accessed on 10.11.2015]. Available: https://msdn.microsoft.com/en-us/library/hh231022%28v=sql.110%29.aspx

Morris, J. 2012. Practical data migration. BCS, The Chartered Institute.

Olson, J.E. 2003. Data quality: the accuracy dimension. Books24x7 Edition, Morgan Kaufmann.

Orr, K. 1998. Data quality and systems theory. Communications of the ACM 41, 2, pp. 66-71.

Otto, B. 2009. Functional reference architecture for corporate master data management.

Otto, B. & Schmidt, A. 2010. Enterprise master data architecture: Design decisions and options. 15th International Conference on Information Quality (ICIQ 2010), Little Rock.

Pekkola, S. 2012. ICT STANDARD FORUM BLOG. [Accessed on 22.11.2015]. Available: http://www.tivi.fi/Arkisto/2012-10-01/Patruunan-%C3%A4%C3%A4nt%C3%A4-etsim%C3%A4ss%C3%A4-3195018.html

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

Radcliffe, J. 2007. The Seven Building Blocks of MDM: A Framework for Success.

Redman, T.C. & Blanton, A. 1997. Data quality for the information age. Artech House, Inc.

Redman, T.C. 2008. Data driven: profiting from your most important business asset. Harvard Business Press.

Redman, T.C. 2001. Data quality: the field guide. Books24x7 Edition, Digital press.

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

Robson, C. 2002. Real world research: A resource for social scientists and practitioner-researchers. Blackwell Oxford.

Saunders, M.N., Saunders, M., Lewis, P. & Thornhill, A. 2011. Research methods for business students, 5/e. Pearson Education India.

Scarisbrick-Hauser, A. & Rouse, C. 2007. The whole truth and nothing but the truth? The role of data quality today. Direct Marketing: An International Journal 1, 3, pp. 161-171.

Smith, H.A. & McKeen, J.D. 2008. Developments in practice XXX: master data management: salvation or snake oil? Communications of the Association for Information Systems 23, 1, pp. 4.

Umar, A., Karabatis, G., Ness, L., Horowitz, B. & Elmagardmid, A. 1999. Enterprise data quality: A pragmatic approach. Information Systems Frontiers 1, 3, pp. 279-301.

Vilminko-Heikkinen, R. & Pekkola, S. 2013. Establishing an Organization's Master Data Management Function: A Stepwise Approach. System Sciences (HICSS), 2013 46th Hawaii International Conference on, IEEE. pp. 4719-4728.

Vilminko-Heikkinen, R. & Pekkola, S. 2013. Establishing an organization's master data management function: a stepwise approach. System Sciences (HICSS), 2013 46th Hawaii International Conference on, IEEE. pp. 4719-4728.

Wailgum, T. 2007. Master Data Management: Truth Behind the Hype. CIO Magazine .

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

Wang, R.Y. & Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems pp. 5-33.

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

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

White, A., Newman, D., Logan, D. & Radcliffe, J. 2006. Mastering master data management. Gartner Group, Stamford .

Wolter, R. 2007. Master Data Management (MDM) Hub Architecture .

Wolter, R. & Haselden, K. 2006. The what, why, and how of master data management. Seattle: Microsoft Corporation .

Woodall, P., Borek, A. & Parlikad, A.K. 2013. Data quality assessment: the hybrid approach. Information & Management 50, 7, pp. 369-382.

Xu, H., Horn Nord, J., Brown, N. & Daryl Nord, G. 2002. Data quality issues in implementing an ERP. Industrial Management & Data Systems 102, 1, pp. 47-58.

Yin, R.K. 1994. Case study research: design and methods.

APPENDIX A: THE INTERVIEW THEMES AND QUESTIONS

Data quality

“What is data quality”?

“How important is having high quality data?”

Master data quality

“What is master data?”

“Does master data quality differ from data quality in general?”

“What is high quality master data?”

Acceptable master data quality

“What is acceptable master data quality?”

“What is an example of an unacceptable master data quality issue?”¨

“What is an example of an acceptable master data quality issue?”

Problems with master data quality

“Which are the biggest problems?”

“What are the root causes?”

“How do they affect business?”

Data quality in MDM hub

“What is the role of MDM hub in master data quality?”

“How could MDM hub be used to improve master data quality?”

Best practices

“What could be done in order to improve the data quality process in MDM hub?”

“What are the business best practices compared to the technical?”

“What could be the sum best practices in this area?”