Katri Vihavainen

Managing Data Integrity as Part of Master Data Management

Helsinki Metropolia University of Applied Sciences

Master's Degree in Business Administration

Business Informatics

Thesis

30.3.2014

Helsinki Metropolia University of Applied Sciences

Author	Katri Vihavainen
Title	Managing Data Integrity as Part of Master Data Management
Number of Pages	56 pages + 36 appendix pages
Date	30 March 2014
Degree	Master's Degree in Business Administration
Degree Programme	Business Informatics
Instructors	Pia Hellman, Senior Lecturer, Metropolia James Collins, Lecturer, Metropolia

The present thesis concentrates on data cleaning actions due to a Master Data Management (MDM) Program in a case company. The goal of the MDM project is to streamline the data in order to offer high quality master data which will support business processes.

The purpose of this research was to provide guidelines and best practices on how to achieve and maintain high level integrity of customer data. In this research, first the theories around MDM, data quality management and data cleaning were studied, followed by carrying out a current state analysis, and then providing practical guidelines based on the existing literature and experiences with data cleaning. This research answers to the question how to ensure data integrity in the future. In addition, it provides understanding on how data cleaning can be executed manually. The research results can be used for future data cleaning projects inside and outside the case company. This research was conducted with using qualitative approach as the prime emphasis was on gaining understanding, and the data was collected with action research.

According to the research, good data quality has many benefits. It makes the data more trustworthy and decreases the time that the user has to spend searching, checking and correcting the data. Quality problems in the core system are most likely also transmitted to the business target systems. Good data integrity results from valid, accurate and consistent data. When data integrity is high, data follows business rules and it is timely, and satisfies the business needs. Common business rules and data quality rules ensure that the data is entered similarly throughout the organization in an accurate way. Both business rules and data quality rules were studied and reported for the case company during the project and they are introduced in this Thesis. Defining and communicating the rules is usually not enough, and data controlling is necessary. In this research means for data controlling are introduced.

Finally, guidelines for data cleaning project are provided in the thesis. With this research it is proven that the data cleaning project can be executed manually without external cleaning services, but the most beneficial combination of manual and system cleaning would be an interesting topic for future studies.

Keywords	Data quality, data integrity, Master Data Management, data cleaning, business rules, data quality rules, data controlling, data enrichment, data harmonization.

Contents

1 Introduction

1.1 Background for the study

Companies have been measuring their business along three dimensions for many years; People, Process and Technology. There is a fourth dimension that needs to be included: Data. Few businesses have put in place the appropriate processes to ensure high data quality on a consistent basis (Strout, Eisenhauer, 2011). Goal of the Master Data Management (MDM) project described in this thesis is to streamline the data in order to offer high quality master data which will support business processes. Data cleaning is one part of this.

Master Data can be defined as being the core information to the running of business and typically it refers to people, places and items. Master data may include customer and vendor lists and addresses, product information and part numbers. As processes and organizations grow larger and more complex, there is an inevitable tendency for master data; critical pieces of information that must be correct. When duplicates and inaccurate data occur, there is a master data problem, which will have a negative impact on the business (J. Shah et al., 2012). In today's economic uncertainty, the company must be able to trust its data when making decisions (Strout, Eisenhauer, 2011).

This Master's thesis was written as part of the Master's Degree Programme in Business Informatics. It is a Degree Program for people already working in the field, who want to deepen their strategic and managerial expertise and it is organized by Helsinki Metropolia University of Applied Sciences. Studies can be conducted as business or engineering studies, depending on the student's previous studies. Master's Thesis is a work-related, cooperative development process typically implemented as action research. As I work in Financial Services, I was searching for the thesis topic inside that function. A MDM Program was launched in 2010 that has a direct influence to my daily work in Customer Master Data (CMD) Europe team, so I contacted the Customer Domain Manager in late 2012 and asked for an opportunity to prepare my Master's Thesis for the project. The request was accepted and I started to work 20% of my working time for the project to get more involved. While working with the MDM Program mostly concentrating on reporting and analyzing business rules and data quality rules, and

simultaneously performing data cleaning actions with the Customer Master Data team, I noticed that the actions should be reported for future as they might be of assistance in ensuring the data integrity of the case company.

Customer Master Data maintenance and development in the case company is part of Financial Services function providing transactional services for all case company's business and functions. CMD Europe team is maintaining all the general data of case company's business areas, miscellaneous invoicing customers, and the company code and sales area data of units registered in Europe. Other regional teams are located in USA and China, where the company code and sales area data is maintained for customers registered for their region (NA and APAC). Other related tasks for CMD Europe team are for example bank master maintenance, data transfer monitoring, mass update runs, assistance in business projects and end-user training. Master data is maintained in 16 different systems companywide. There are approximately 40 systems where this data is utilized in the processes. Approximately 90 end users have customer master data maintenance rights to the core system. The data flows from the core system partly automatically to the accounting system and business target systems. There is also one business area that has two additional teams assisting in maintenance of some of the customer types. There are approximately 15 300 basic customer requests handled yearly, of which CMD Europe team takes care of a bit over 90%.

This research concentrates on the data cleaning actions due to a Master Data Management program in the organization I am working in. Research has two purposes. Firstly, it aims at investigating the aspect of data quality in the customer master data. Secondly, this study aims at providing the case company with practical guides on how to maintain high level data integrity in the core system. Research question is: How to ensure data integrity in the future? Also, since the data cleaning was made without outside assistance in a form of a data cleaning program or a tool, another research question is: How to clean the company's data manually? This research is done by first conducting a review on the existing literature about Master Data Management and especially on the data integrity and data cleaning perspective and then introducing the current data quality situation of the case company and a plan how it can be cleaned and what might be the actions in the future to maintain data integrity.

Customer master data refers to the master data information of the customers that the case company is selling to. These can be invoicing customers, delivery customers, internal customers, employees, consumers, payers or bill-to customers. Full list and definitions of the customer subtypes can be seen in Appendix 1.

1.2 Research limitations

As I am working solely with customer domain, this thesis is limited to concern only this domain, and not involving the other domains in case company's MDM project, which are vendor, item, HR, and finance.

This research is prepared for the customer master data team using the SAP ERP system, which is the core system for customer master data in the case company. The data used by the team is global data, meaning that it is the core data of the customer. Because of that, the research results might not suit the data cleaning of a business target system.

In this thesis, the possibility of using an external party to clean the data was not taken into consideration.

1.3 Research process

This thesis is organized in research design, literature review, current state analysis, research results, and finally conclusions and summary.

Figure 1 Research process

In research design chapter the method, design and data collection are described. In the literature review the existing theories on Master Data Management, data integrity and data cleaning are introduced. Chapter 3 is providing a Current state analysis indi-

cating the company's current state including a data quality assessment implicating it with the desired state. Research results will be introduced in chapters 4 and 5 where I will give my suggestions on how to manage the data integrity in the future, followed by conclusions where the main findings are indicated, and a short summary of this Master's thesis.

1.4 Research method

Research process and research method used are influenced by the researcher's background when it comes to research orientation. A particular research orientation prescribes the relationship between methods, data, theories and values of the researcher. As I am involved in the project that this study is based on, I use induction and deduction from my daily work. I observe and record what is seen without any prejudice. Induction is based on empirical evidence, while deduction is based on logic (P. Ghauri, K. Gronhaug, 2002).

It is often assumed that theory should precede data. Often, however, this is not the case. The researcher might observe something he or she does not understand. This is often the case in qualitative research with the prime emphasis on gaining understanding. Interactions between theory and data take place when doing a research (P. Ghauri, K. Gronhaug, 2002). This thesis was conducted using a qualitative approach.

For this thesis, the topic was raised from MDM project to which I was taking part. There was a need for cleaning and harmonizing the data before any MDM technology will be introduced. There had been cleaning actions done before in the CMD team but not in same extent as it will be done now. I realized a need for recording the current cleaning project and studying its influence on the master data of the company. By doing this, the data of this project would be available for anyone starting a customer master data cleaning project later on in the company.

In descriptive research the problem is structured and well understood. In this case the problem was that customer data is not as accurate and complete as wanted. There is a need for Master Data Management project. In order to have the MDM technology, the data needs to be as clean as possible. I was cleaning the data while preparing the the-

sis. The research was conducted with the help of the existing literature on MDM and data integrity, and the findings made in the project team while cleaning the data.

1.5 Data collection

This thesis was conducted as action research. Action research is known by many other names, including participatory research, collaborative inquiry, emancipatory research, action learning, and contextual action research. Action research is used in real situations, rather than in contrived, experimental studies, since its primary focus is on solving real problems (O'Brien, 1998). There are some ethical issues to be considered when conducting action research; Need to make sure that the relevant persons have been consulted, and that the principles guiding the work are accepted in advance by all. Permission must be obtained before making observations or examining documents produced for other purposes. Descriptions of others' work and points of view must be negotiated with those concerned before being published. In addition, the researcher must accept responsibility for maintaining confidentiality (Winter, 1996).

2 Conceptual framework

In this chapter I will explain the most important concepts of data integrity and data cleaning. As every case of data managing is different, it is still advantageous to understand the basic concepts in order to have a fuller understanding on the matter. This might help in seeing the data from a wider perspective. In chapters 4 and 5 I will give guidelines on how to manage the data by using the existing theories on MDM, data integrity, and data cleaning.

2.1 Master Data Management

> Data is a very valuable asset of any business. It's intellectual property owned by the business that must be treated as such. That means it must be protected, guarded, managed, and governed in such a way that it retains or increases in value (Strout, Eisenhauer, 2011, p. 4).

Importance of Master Data Management has been noted especially after mergers and acquisitions when the company's customer base might be scattered in several different systems. Master data elements appear over and over again in the many different information systems in the organization. As an example, a customer name in the transaction record should be equivalent in the other systems. If the customer name is different in each database, there can be errors, waste and other business problems. Businesses need to manage master data to keep it consistent and clean across multiple databases and systems (J. Shah et al., 2012). One goal of MDM project is to clean and harmonize the data so that the elements are the same across all systems that need them.

Master data management brings many benefits to the company but it requires resources as in time, money and staff. I will now start with giving an overview what is master data management as a concept, what is master data management as a project, and then describe the benefits of it. This should substantiate the reader that a great part of MDM is managing the data integrity.

Walker and Ganapathy (2009) define master data as being the core data of a company; customers, vendors, products and employees. Whereas MDM is a combination of

data governance, business processes, data quality, data enrichment, and a technical solution. It is a companywide program that requires co-operation of a company's business units and business areas. MDM also must ensure the availability of accurate and up-to-date master data that is of good and trusted quality (A. Walker, J. Ganapathy, 2009).

Master data management (MDM) as a formal discipline has been around for over a decade. According to Shah, Manathara and Hoeppe it is becoming increasingly critical as the business world and its attendant information systems grow more dynamic. MDM improves the business' processes, helps decision making, and furthers the liability of the data. MDM is also of a vital help in case of mergers of acquisitions.

Maintaining accurate master data is an action happening several times every day. Each time a customer or a vendor is created or changed in business application it needs to be done correct. With MDM, the company standards are followed and completed with full and accurate data record details. This means that if in a day-to-day work the data quality rules are not followed, with MDM all the existing data records will be checked and corrected. It might take years to change the users' attitudes concerning how they treat and respect your master data records (A. Walker, J. Ganapathy, 2009).

2.1.1 MDM program

In the first parts of MDM project, it is necessary to get everyone who needs to be involved, aware of the project and to be "on the same page". Gartner, Inc has created the "Seven building blocks of MDM" which can be used in this stage. It can also be used later on in the project to assess the current status and future needs and initiatives.

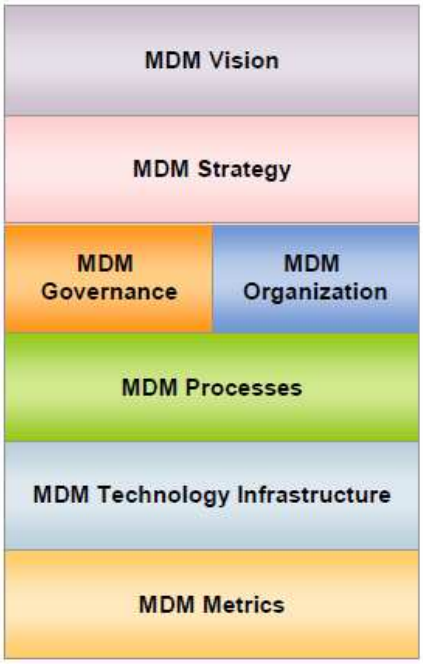

Figure 2 Seven building blocks of MDM, Source: Gartner 2007

Vision: Here the scope of the MDM can be described, for example the domains which it is affecting. It needs to be clearly stated how the MDM vision supports the company's business vision. The business justification for the project has to be clear and enduring. Similar to business vision, there can only be one MDM vision. As sustainability and responsibility are part of the case company's business vision, the vision of the MDM project described in this research is aligned to that vision, as it is to ensure high level in quality, accuracy, availability and completeness of master data.

Strategy: How the MDM vision will be realized and how to manage the master data assets within the organization? Who is the data owner and in what manner will the data be enriched? Where the data will be stored and where it will be published? Who needs to consume it and who will have the rights to modify it? These requirements need to be analyzed to determine the priorities. The result will be a prioritized road map; an implementation plan. The case company's strategy is to reach the MDM objectives by analyzing the current situation in detail for five domains and based on that the future governance, processes, data models and business rules for managing the master data will be designed.

Governance: Without effective governance, an MDM initiative will probably fail. Gartner defines MDM governance as the specification of decision rights and an accountability

framework to encourage desirable behavior in the ongoing authoring, storage, enrichment, publishing, consumption and maintenance of master data. It specifies the processes, roles, standards and metrics that ensure the effective and efficient use of master data in enabling an organization to achieve its goals. When choosing the members to MDM governance group, it is good to choose employees from different parts of the organization to ensure a wider view on the matter.

Organization: A unified MDM program will create change. Therefore, communication, training and change management will become challenges and they must be planned and resourced with appropriately skilled people. The MDM program will require a matrix organizational approach, with leadership coming from the business side and the IT organization being a major contributor. At the core of the MDM program there should be a small number of dedicated resources in a central team, the so called steering group. The appointment of data stewards, who have responsibility for the quality of master data, will be the key to success.

Technology Infrastructure: In this step it needs to be determined what technologies are needed to enable the MDM vision and strategy, and where and how this technology should be sourced.

Metrics: In the Gartner research it is said that without measuring the quality of master data and its effects on business performance, there is no objective basis for reporting improvements. The top level of an organization's performance management metrics are the corporate goals, which are typically revenue, profit and market share related. On the next layer are strategic metrics related to the operational effectiveness, customer intimacy, and product or service leadership goals. On the next layer there are process-level metrics. MDM improvements should have a positive, direct effect on the majority of these metrics.

There are issues that need to be taken into consideration when implementing the MDM project. Company Hewlett-Packard (2007) has defined a paper on issues to be avoided when conducting MDM in a company and they will be shortly introduced here. According to the paper, managing master data is more about business and people than about technology. Too often companies are looking at MDM purely in technical terms. The

truth is that the master data management must come from the business side. Similarly to the Seven Building Blocks of MDM research, also Hewlett-Packard states that data governance organization must include members from the business who have enough influence and dedication to be part of the project. Referring to data stewards, the paper continues by saying that it is important to give them tools enabling them to see the business rules around the master data so that they have the competences to validate and maintain the rules in their work.

A great amount of MDM skill and knowledge exists inside the various businesses within the company, but the equal practices, technologies and governance or stewardship is missing. It might be impossible to get the data management completely unified inside the whole company, and for this reason the project should take into consideration the different needs of the different departments and businesses. Therefore, MDM cannot be conducted as a "big bang" inside the company at once. It must be divided and company must work with only a manageable number of business units at one time (Hewlett-Packard, 2007).

In the case company it was discovered that poorly handled master data had led to high efforts in analyzing and verification of actual data, and also longer reaction time due to bad data. This can affect revenue losses due to delays and the associated costs of maintaining the data are high. Therefore, the case company defined the objectives of MDM to implement processes and tools to ensure high quality of data, to provide a solid, scalable information technology platform that supports business needs, and to provide transparent reporting of master data and subsequently improve business reporting.

2.1.2 Benefits of MDM

Outside of the company, customers view the company as a single entity. For external customers, it might be difficult to understand that a customer entity might for example have different customer numbers inside a company due to several business areas and more than one system used in the company.

> The internal structure of divisions and departments is purely for the company's convenience, not the customer's (S. Tuck, 2008).

According to S. Truck, MDM enables a company to start thinking and behaving as a joined-up organization because of the complete and accurate view of the customers and the products they hold. In addition, the time and effort (and therefore money) is wasted in organizations today because data are fragmented, incomplete and inaccurate.

Master Data Management project develops a common vocabulary, the appropriate policies and rules in order for the business to keep data organized and consistent. Data needs to be consistent from one person to another, consistent over time, consistent from system to system, and from one department to another (Strout, Eisenhauer, 2011). In the case company, there has not been common processes for data management with the consequence that multiple systems and redundant processes exist that can lead to inconsistencies.

One of the benefits of MDM project is providing the means to make organization more agile. "Business Agility" includes the ability of an organization to enter new markets or current markets with new products as quickly as necessary (Strout, Eisenhauer, 2011).

MDM also helps in the creation of a single version of the "truth", and managing the quality of the data that makes up that "truth" is a significant issue for almost any company. The volume and variety of data quality issues and cross-system inconsistencies accumulate over long periods of time making it challenging to list and prioritize the issues across the organization (Strout, Eisenhauer, 2011).

Better collaboration is also one benefit of MDM. By having access to common, consistent, and accurate data, employees can utilize and better rely on the technology to collaborate and bring new value to the business (Strout, Eisenhauer, 2011).

One objective of the case company's MDM project was to establish a master data template. The existing tools used for master data management would be assessed and modified, or if needed, completely new tools would be implemented along with new processes. When implementing new tool, it is crucial that the data quality rules are

clearly defined and existing data has to be cleaned to match those rules. In case new tool will not be implemented during the MDM program for the case company, it is still beneficial for the company to go ahead with the data cleaning project to have as high quality data as possible. In the next chapter I will describe the benefits of data integrity more deeply.

2.2 Data integrity

Data integrity means the accuracy and consistency of stored data. Since so many important decisions are based on the information generated from the information systems, data integrity is of vital importance for the companies (Morley, Parker, 2010). Hewlett-Packard (2007) defines master data quality as the relative comparability, consistency, and confidence business users ascribe to the master data record. It is also said that quality problems in master data are likely to cascade to other systems and that data quality is an ongoing program instead of a one-time event. Meta data, on the other hand, enables the stewards to see the business rules, hierarchies, and definitions for the master data.

In today's wired world, the costs and consequences of inaccurate information are rising exponentially. With the complications arising from merging different data sets, as in the aftermath of a merger or acquisition, the difficulties of data cleaning multiply (M. Wheatley, 2004). Protecting and improving the integrity of the data on which the business makes key decisions will propel the organization to new levels of trust and agile decision making (Strout, Eisenhauer, 2011). Data quality also means that the data follows business rules and it is timely, and satisfies the needs of the business (Bischoff, Alexander, 1997).

Shah, Manathara and Hoeppe have stated that data quality involves factors such as accuracy and completeness. For example, an address list full of misspellings and missing postal codes is of low quality.

> In the master data context, data quality extends the basic definition of data quality and adds the issue of data consistency across multiple systems (J. Shah et al., 2012).

According to Shah, Manathara and Hoeppe data quality dimensions with master data fall into the following major categories:

- Duplicate data
- Conflicting data
- Incomplete data
- Invalid data
- Inaccurate data

Duplicate data means that there are for example two records for the same customer in the same system. According to M. Wheatley (2004), the manager of a data cleaning project in his company stated that the first step was to identify homonyms and synonyms. Homonyms are two or more different items with the same identifier. Synonyms are the same item with more than one identifier. Synonyms can be here considered as duplicates. Users of the data must keep in mind that the data integrity of the unique key needs to be projected and therefore duplicate records need to be avoided at all costs (A. Walker, J. Ganapathy, 2009).

Conflicting data appears when for example, a record for the same customer exists in two systems, but the addresses do not match. Unfortunately, not everyone enters the data the same way. Data is spread all over the organization in different systems, using different languages and standards, and sometimes replicated in different forms (Strout, Eisenhauer, 2011).

If any data on any variable from any participant is not present, the researcher is dealing with missing or incomplete data, for example when a customer address is missing a region when it should be mandatory information. Large data sets are full of legitimately missing data. Need to be noted too, that in the case of legitimate missingness, missingness is meaningful (Osborne, 2013). This means that in some cases an information field is left empty for a reason.

Concerning invalid data, M. Wheatley (2004) describes in the example case that there was also the relatively easy but time consuming task of examining the validity of the identifiers. After that they had to correct the wrong information in the other data fields. These errors could cause major problems in the supply chain process. To avoid this

kind of errors, the data cleaning project team turned to a data-profiling tool, which highlighted errors and inconsistencies.

Inaccurate data means that the data is inserted in a wrong way, which can be caused by a typo in keyboard entry or having incorrect information. As an example, when choosing a value from a reference list, the user might accidentally select the value before or after the intended value without noticing.

2.2.1 Business rules

In the example case of M. Wheatley it was pointed out, that the data cleaning project succeeded because team members worked closely with the users of data. In addition to the good communication with the business users of the data, clear rules should be defined what values are acceptable in master data.

Entity rules can be divided into three groups: *Uniqueness, cardinality,* and *optionality* (Adelman, et al., 2005). Uniqueness means that each entity must have a unique identifier which cannot be null. In case there are more than one attribute identifying the uniqueness of an entity, the amount of attributes must be kept minimal. Cardinality refers to the degree of a relationship, that is, the number of times one entity can be related to another, for example a customer entity can only have one sales manager at a time, but a sales manager can be linked to several customers. Optionality identifies the maximum and minimum number that entities can be related. There are only two options; either there is a mandatory relationship which means that two entities must be related at least once, or an optional relationship, meaning they do not have to be related (Adelman, et al., 2005).

Adelman et al. (2005) have also defined two attribute rules: *Data inheritance* and *data domain* rules. The data inheritance applies to supertypes and subtypes, where a supertype means for example a customer number in ERP system and subtype is the same customer in marketing system. All generalized attributes of the supertype should be inherited by its subtypes and the key identifier should be the same. Data domain rules refer to the list and range of values available in throughout the master data. For ex-

ample, a date should be inserted in the similar format by all users and only a set of allowed characters should be used.

Data dependency rules apply to data relationships between two or more entities as well as to attributes (Adelman, et al., 2005). *Entity-relationship dependency* means that the existence of data relationship depends on the state of another entity, for example it is impossible to create an order for customer who is blocked, or when creating an order it is mandatory to insert incoterms. *Attribute dependency* means that two or more attributes are dependent on each other, for example, for customer in business area "X" there should not be a possibility to choose a value that only applies for business area "Y".

2.2.2 Data quality rules

Data quality is determined with various sets of rules. According to Adelman et al. (2005) data validity rules can be divided into six groups: *Data completeness, data correctness, data accuracy, data precision, data uniqueness,* and *data consistency.*

Data completeness rules can be divided into concerning entity, relationship, attribute, or domain. Entity completeness requires that all instanced need to exist for an entity. Relationship completeness means that referencing exists between all the entities that need to be referenced. Attribute completeness means that all fields need to be present, that are determined for that business entity. Domain completeness refers to the values in attributes, meaning that values need to be acceptable and there is a distinction between a null and a missing value (Adelman, et al., 2005).

Data correctness rule means that all data values for an attribute must be correct and representative. For example the list of available values needs to be correct and up-to-date, whereas data accuracy rules states that the value inserted into an attribute needs to be valid and true. Data precision rule specifies that the value filled into the attribute has to meet the business requirements, intended meaning, and usage (Adelman, et al., 2005).

There are several aspects to data uniqueness rule. Every entity needs to exist only once, so there must be a rule against creating duplicate records e.g. duplicate payment terms. Also the identification key cannot be created more than once, otherwise a several different customers could have the same customer number as an example. There should not be homonyms or synonyms, in other words each attribute should have only one unique definition and name. Also, every attribute should only be used for one purpose (Adelman, et al., 2005). As an example if there was a field called Customer terms, which would include values for both payment terms and incoterms, it would mean that the field is overloaded and causing data quality problems.

Data consistency means that the values in the attributes must be consistent through time, users, and systems (Adelman, et al., 2005).

2.3 Data cleaning

Data cleaning, also called data cleansing, refers to the correction of erroneous data. This means exploring the data for possible problems and correct the errors when possible. Data cleaning project should be clearly defined before starting the actual cleansing of the data.

When choosing the approach for extracting the data from the business system in order to clean the data, it should be done in such a way that the process is repeatable and can be reused for different systems. The required attributes need to be determined and when possible only those attributes should be extracted that are necessary for the MDM design (A. Walker, J. Ganapathy, 2009).

2.3.1 Data cleaning phases

According to Van den Broeck et al. (2005), data cleaning can be divided into three phases: Screening *phase, diagnostic phase,* and *treatment phase.*

In the screening phase the types of oddities are distinguished, such as strange patterns, inconsistencies, and incomplete or excess data. The screening methods can be

statistical, but in addition the inconsistencies might be already discovered earlier by the investigator, or they might be detected during a pilot studies, evidence in the literature, or common sense.

In the diagnostic phase the deviations are divided into groups of whether they need correction or not. There can be for example erroneous values, values that were thought to be false but are in fact correct, or values that are suspected to be false but need checking.

Final phase is treatment phase where the options are to correct, delete, or leave unchanged those values that were found or suspected to be erroneous. According to Van den Broeck et al., impossible values should never be left unchanged, but either corrected if a correct value can be found, or deleted. The values that are, or are suspected to be erroneous, can be further examined individually or as a group and analysed before doing any corrective actions or the decision of leaving the values unchanged. Here, it should be considered what the impact of the wrong data is if it is left unchanged.

Data management and data cleaning require transparency and proper documentation. Proper documentation should exist for including differential flagging of suspected errors, diagnostic information, and information on the editing, dates and personnel involved. (J. Van den Broeck et al., 2005).

2.3.2 Data cleansing actions

According to Kelly (1997), there is a selection of actions on how to deal with integrity violations once encountered. The cleansing actions can be categorized to three options: *audit data, filter data,* and *correct data.*

In all three actions, data is first tested against integrity rules. When using audit data option, the occurrence of integrity violations are reported without taking any steps to restore integrity to the data. This helps to know the scope of the problem, but does not fix it. Filter data action means that after the data violations are detected the offending data is removed from the set used to populate the given data. Filtering may be

applied at any of several levels including removal of individual data elements or entire records or rows. Correcting data refers to the repairing of data using logic defined by the business to restore its integrity.

3 Current state analysis

I will now use gap analysis to define the current state analysis. Gap analysis is used to determine where you are and where you want to be. This shows you the gap between how your business wishes to perform and how it actually performs (Franklin, 2005). First step of the analysis is to identify the objectives that need to be achieved, in other words what is the desired future state. Second step of the current state analysis using gap analysis is to analyse the current situation. Final step of the analysis is to identify how to bridge the gap between the desired state and the current situation.

After the Current state analysis I will go through the steps in the data cleaning project, and give my suggestions on how to manage the data integrity in the future.

3.1 Desired future state

Desired future state is to ensure that all necessary master data can be found in the core system without having any duplicate data as well as to ensure that the data is complete, of good quality and correct. The vision is to have high quality customer master data which will support business processes and enable new business opportunities (Source: Case company internal material).

3.2 Current situation

Most of case company's business processes have evolved over time independently inside the business areas and functions. The main emphasis has been on optimizing the own responsibility area and getting results quickly. This has led to a situation in which nobody has been managing the overall process. The tangible end result is that master data has become redundant, inconsistent and inaccurate (Source: Case company internal material).

Customer Master Data Europe team was established in 2010 to maintain the core data in a centralized and unified way. Need for data cleaning was already discovered in

2011, but due to limited resources only some actions were conducted that time. When MDM program was started, it was decided that the data cleaning should be conducted again, in order for the data to be as accurate and complete as possible. Total customer number was approximately 219 700 global customers of which 154 500 belonged to the scope of data cleaning activities.

The statistics of the data gathered in late 2012 indicated that almost 60 000 invoicing customers were missing a business area code of total amount of 113 600 invoicing customers. This means that for over half of the invoicing customers we are not able to identify to which business area the customer belongs to.

Account groups	
A	38404
B	189
C	4208
D	30
E	523
F	6501
G	13
H	27
Blanks	**59828**

Figure 3 Invoicing customers sorted by business area

Another alarming issue was that there were thousands of customer entities that had not been used for years. Large amount of active customers in the system makes reporting slow. Also, validity of old customer information, for example tax codes and addresses is uncertain. Queries taken from the core system and accounting system showed that there were 11 896 dormant customers in the system that were not marked as inactive but not used during recent years.

What was also noted was that there were several company codes and sales areas still active in the system that are not in use anymore. Company code is an organizational unit within financial accounting. Sales organization is an organizational unit responsible for the sale of certain products or services (definitions from SAP).

It was discovered that in the core system there were almost 40 000 active invoicing customers in EU area without a VAT registration number. There are several reasons

why this is a data integrity violation. Customer identification is difficult when a VAT number is missing which may lead to creation of a duplicate customer, or a wrong customer entity is used. There are also legal reasons why all EU customers should have VAT number in place when having invoicing transactions with that customer entity.

Customers from EU countries	
VAT's entered total	43375
Active accounts	42893
Inactive accounts	485
Missing VAT's total	45101
Active accounts	**39778**
Inactive accounts	5323

Figure 4 Invoicing customers inside EU sorted by VAT registration number

Address attributes were also audited, and there were many inconsistencies and missing data in the customer records. There were a lot of missing postal codes for example, and the region information was written in a Street attribute instead of the Region attribute. It was decided, that Name, Street, PO Box addresses, City, and Country attributes should also be checked during the data cleaning project.

The team also identified several customer entities suspected to be duplicates.

Account group	Duplicates
Bill-to customer	37
Delivery customer	36
Payer	4
Internal customer	49
Ship-to customer	782
Sold-to customer	1375

Figure 5 Duplicates

Some business specific data violations were identified too, for example missing sales managers or inconsistencies in payment terms.

It had been also noted before the project that there were some duplicate and inaccurate values in the drop-down lists of attributes. For example in Place code list, there were values for "Dublin Great Britain" and "Dublin Ireland", and in Payment term list there were duplicate payment terms with different codes. It was decided that although maintenance of the information inside a drop-down list is not a responsibility of data custodians, the team would audit these data quality issues and report them forward for correction.

This data quality assessment shows that there are data quality violations with all quality dimensions described in the existing literature by Shah, Manathara and Hoeppe.

3.3 Next steps

Data cleaning and enrichment plan was drawn according to the current situation:

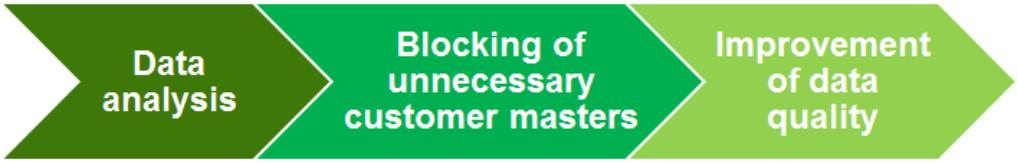

Figure 6 Data cleaning and enrichment plan

Plan included following points:

- Correcting and analysing the data
- Defining the critical attributes
- Identification and blocking the out of business customers
- Identification and blocking the inactive customers
- Duplicate data identification and blocking
- Data enrichment actions

Detailed plan is available in Appendix 2.

There were also a short term and a long term plan for the CMD Europe team and CMD processes: Short term plan was targeted to year 2013 and the plan was to enforce the current process and practicalities by continuing with the high quality level of work and also add the data cleaning actions to daily tasks. During 2014 further data cleaning actions will continue in customer master area as well as data standardization and authorization issues will be taken into consideration. The long term plan was applying to 2015 onwards and it includes the implementation of new process and possibly a new customer master tool determined by the MDM project. Here also the business rules, responsibilities and controls, and the team structure will be reviewed. (Source: Case company internal material)

4 Data cleaning process

In the following chapter, I will explain how most of the data cleaning actions were executed in the case company. These actions were done by using as tools SAP's own queries and Microsoft Excel. First phase of the project took place during 2013, but since there were a limited amount of resources to execute all the wanted actions, the second phase was started in 2014. The first phase was mainly conducted by all CMD Europe team members and supervised by team leader. This was done in addition to daily tasks in the team. The second phase lasting approximately 6 months will be conducted with only 2 team members, but working full time with the cleaning project. To view the detailed description of the tasks during the 2013 cleaning project, see appendix 3.

A spreadsheet for monitoring the situation was created for critical data, duplicate data, inactive customers, data enrichment, and data harmonization after which the tasks were divided to team members. This report needs to be kept as accurate and up-to-date as possible, as it also serves as a report on what was done. Contact persons were decided, who would then receive lists of customer records with erroneous data. They should review the changes team wants to perform on the customer records and give their approval or rejection. Contact persons would represent a business area, geological area or knowledge on a certain customer subtype, for example internal customer records. Cleaning tasks were started simultaneously but later on in the cleaning project it was discovered that the data enrichment should be have been done only after duplicate and inactive customer cleaning, since there are less active customer records and the scope therefore is smaller. I will now describe the process of these actions.

4.1 Data cleaning

When trying to decrease the amount of errors in the master data, it is referred as data cleaning. There can be several customer records as active in the system, which should not be present. These are duplicate and inactive customers. In the existing data cleaning theory there was no mention about inactive customers causing data quality problems, but in the team's experience with maintaining the master data, we have discovered they actually do. I will now explain how we started the data cleaning project, by

first scoping the problems and then trying to decrease the number of active customer records in the system.

4.1.1 Critical data

In the beginning of the project it was of great importance to scope the problems in the master data. This is the screening phase as mentioned in chapter 2.3.1.

The whole master data was screened for oddities in address attributes, tax codes, business area codes, and some business specific data attributes such as segment, sales manager and head office information.

With the help of these statistics the scope was determined and team could start thinking about possible solutions to solve the erroneous data. As CMD Europe team is responsible mainly on the customers used by company codes registered in Europe, it was decided that CMD Europe team will screen and diagnose the data globally, also for APAC and US team, but the corrections would be handled locally in those teams.

4.1.2 Duplicate data

Duplicate customers are created when a same customer is created with more than one identifier in the same system. Usually this happens when a customer is created without checking if the customer already exists. Same customer record can be used by several business areas so there is no reason to create a duplicate customer unless in very rear exceptions, for example if a rivalling business areas should be using the same customer record there might be a need to establish two identifiers for one customer so that the users are not able to see the contract terms of each other. Another reason for duplicate customers is that the address of a customer has changed, and a customer record is changed without checking if the same customer with this address already exists in the system.

Identification of duplicates was proven to be challenging due to incomplete data. A report of suspected duplicates was taken from the core system according to name,

address and tax code. Since not all invoicing customer records contain a tax number although they should, it was difficult to know for sure if the suspected duplicates are true duplicates. Additionally, several customers can have the same name and tax number, but if the address is different they are not considered as duplicates since the location is not the same. The name of the customer should always be in the official form but that is not always the case which also made the task challenging. The same problem applies to street addresses; they can be written in a short way, such as "Lincoln St.", as compared to "Lincoln Street". That is why the duplicates need to be searched by using for example 5 first letters of the street address and the country instead of only looking for completely matching values. Here it also needs to be noted, that for example in US the house number is placed before the Street address, meaning if the house number is written in the Main street attribute instead of the House number attribute, for example "5 Lincoln Street", this will not appear in our search for five first characters being the same. In Microsoft Excel this can be fixed by automatically deleting the possible digits from Street address attribute before starting the matching. These values need to be then manually checked to identify suspected duplicates. Another challenge emerged from private person customer records. A person with the same name might exist several times in one city, but it also might be the same person with a new street address. Sometimes the names of private persons were written with first, second and last name and sometimes only with first and last name. The team has no means of knowing if for example Jane Smith with city London is the same person as Jane Katherine Smith with city Luton and which one is her valid city. When identifying these sort of suspected duplicates the team had to use common sense but there was only so much they can do. If the cities are located relatively close to each other we suspect that this might be the same customer entity that has moved to another location.

After identifying these true and suspicious duplicate customers from master data, they need to be sent for checking to the business contacts. The customers decided not to be duplicates are then left active in the system but if possible an identifier is added, for example a tax code to prevent misuse in the future. What makes the task of blocking the true duplicates challenging is that the duplicate customers are created across business areas, which means that all the businesses using the duplicate customers have to decide which number is set as inactive and which one stays as active. Of course all the

businesses want that customer they have been using to stay as active, since it is extra work for them to start using a new customer record. There is a conflict of interest between determined data rules and reality as the cleaning team focus is to remove all the duplicates. That is why both parties need to be ready for compromises.

For future duplicate actions the diagnosing phase should be done immediately after the screening phase and suspected values sent to contact persons for checking. In 2013 the problem was that the customer list was extracted from the system in the beginning of the year, diagnosed during the spring, and sent for approval in the summer. This was too long time period and actually many of the records had been modified in the meanwhile.

4.1.3 Inactive customers

Concept of inactive, or dormant, customers is quite wide. Customer that has not placed orders for recent years can be considered as inactive. There can be several reasons why a customer has not had any orders in the system for the past couple of years; it might be that the contract has ended and not been renewed, or the customer has gone bankrupt, or it places orders so seldom. These are only some explanations. There is also a possibility that a new customer record has been created for this customer. It might be by accident which means we are talking about a duplicate customer, or it might be intentional for example in case a customer has merged and changed its tax number. When a tax number is changed, the customer is seen as a new legal entity and it cannot operate anymore with the same customer identifier. In these cases, the old customer record should be marked for deletion and with billing block in order to avoid misuse. However, if the old customer record still has open items in accounting, it cannot be blocked before the open items are cleared. Nobody is systematically checking this, so the responsibility lies on the data users, or if informed to CMD team, they might for example write a post-it note and check the customer's open items on a regular basis to see when the items are finally cleared and customer can be blocked. Understandably, for this reason there are several old customer records where this blocking has not taken place. Additionally, some data users have different ways of blocking a customer. They might insert only an invoice or an order block for example, which means that the old customer record still appears as active in queries.

Since the total customer number is already high, the inactive customer records should be marked for deletion and blocked for invoice postings, so that they will not slow down reporting. The marking, or flagging, for deletion does not mean that a customer would be actually deleted from the system. Instead, a customer record still exists, but the reporting is easier when these customers can be excluded from the lists of active customers. Customers marked for deletion can be re-activated at any time by clearing the deletion flag.

As there are various reasons to inactive customers as mentioned above, the identification of those customers was done by several means. One aspect was to identify customers that have not had invoicing for the recent years. The customers were divided by customer subtypes, and investigation started. By running the reports of transactions done with the active customer records, the ones with invoicing were excluded from the list. In this phase it was essential to map all the possible systems containing transactional data. As the amount of customers was too large for the system to run reports simultaneously, the customers were divided by company code. A company code can give some hint of which business area is using that customer but since there are several businesses using the same company codes, the definite identification cannot be made based on company code. In the case company the transactional data can be found in two different systems depending on the business area. The active customer record lists where tested in both systems and the ones with sales within the agreed time period where excluded. Unfortunately, this was not enough to identify if the customer was inactive, since invoice customers can also be used as delivery customers. This means that if a customer record is created as an invoice customer but used for delivery customer role, it would appear as inactive customer because of having no invoice transactions. Therefore, as there was no way of knowing if a customer without transactions was an inactive invoice customer or an active delivery customer, the team used some ways to narrow down the number of delivery customers in the lists of suspected values. It was decided that since invoice customers who have open sales area in the two business target systems maintained by CMD team could possibly be delivery customers, all the suspected inactive records were tested against having sales area in these two business target system and the records possessing it were excluded from the list. Customers created within 6 months were also excluded from the list, since

they might be created beforehand to be ready for the upcoming sales activities and therefore not having any transactions yet.

Next step was to identify the customers per business area to the extent that was possible and send the customers to the business contacts for review. Lists where divided by company code and sent to business which was mainly operating with that company code. The responses varied. Some businesses checked their lists throughout and commented on whether they approved or declined the customer to be marked for deletion, or whether that customer was actually not belonging to their business area. After the deadline and some reminders, the team made the decision to mark the customers for deletion which did not receive an answer from the business contact or the answer was that they were allowed to be marked for deletion. Unfortunately, it was then found out that due to poor communication, part of the list was not checked properly and as an outcome several customers were marked for deletion which were actively used as delivery customers, and additionally the global customer hierarchy was build up in a way, that if the head customer was marked for deletion, the hierarchy collapses. The outcome was that those customer records belonging to a global hierarchy in any way were re-activated although being inactive as a compromise in order to support the business.

As the case company has also employees as customers, the same identification methods as above were used to determine inactive employee accounts. The lists where then sent to Human Resources for approval and marked for deletion after that.

In addition to customer not having sales, the team draw a conclusion that a customer lacking valid company code or sales area could be considered as an inactive customer. Queries were made and customers who were only opened for inactive company codes or sales areas with no global hierarchy were marked for deletion without business contact approval. This was also done to customers marked for deletion on a company code level, but not on general level. It was discovered, that this action should actually have been conducted before trying to identify the inactive customers by sales since this method did not require an approval from a business contact and it enabled to reduce the amount of records without business area identification. Furthermore, the amount of duplicates would have been smaller if this action had been done in the beginning of cleaning project.

Account groups	14.11.2012	11.03.2013	24.04.2013	24.05.2013	16.07.2013	19.09.2013	11.12.2013	Difference
Total	219 772	222 147	223 500	224 114	225 298	226 524	228 356	8 584
Active	145 155	137 181	138 337	130 328	115 907	116 890	111 036	-34 119
Inactive	74 617	84 966	85 163	93 786	109 391	109 634	117 320	42 703

Figure 7 The progress of customer in-activations

During the cleaning project approximately 42 700 customer records where marked for deletion. Due to limited resources actions had to be prioritized. This means that invoice customers that might act as delivery customers were not investigated further. In addition, customers belonging to global hierarchy needed to be excluded from the lists although being inactive, as the collapse of hierarchy caused too much problems and the hierarchies needed to be manually restored. These problems were rising from the conflict of interest and historical reasons. For the users of data, more knowledge about customer lifecycle management and the global data mind-set is needed in order to avoid these conflicts. In a proper MDM, a change management process plays a major role for the success of implementation of which before mentioned issues are examples.

In the next cleaning phase 2014 the inactive customer records will be investigated again. At this point more emphasis is put on the inactivation of customers even with global hierarchies. This has to be done in co-operation with one of the Business Support teams to ensure that in case a customer is belonging to a hierarchy and if it is deleted, this does not harm other customers added into same hierarchy. Additional ways of identifying inactive customers will be used in future cleaning projects, for example screening those customers with a comment indicating that the customer is no longer used but it is not blocked for posting or marked for deletion will be reviewed. As customer record cannot be marked for deletion before all open items are cleared the deletion is easily forgotten. In addition customer records which are blocked for deletion are reviewed in case a wrong deletion method is chosen by accident and the records should be in fact marked, not blocked, for deletion. The two ticks are situated close to each other and the function description is similar so a human error can occur and a wrong flag can be chosen. Additionally, one of the business areas has established a new team to maintain their delivery customers in Europe so they will be a great assistance in the identification of inactive delivery customers.

Customers possessing an invalid tax number are also considered as inactive and they should be marked for deletion and posting block. Unfortunately we do not have such system that would automatically check the validity of tax numbers of active customers, so this sort of checks have to be done manually e.g. using the tax code validation services available online.

4.2 Data enrichment

Duplicate and invalid customer records were corrected in the data cleaning tasks described in previous chapter. Next action was to diagnose and take corrective actions on incomplete and invalid critical data.

4.2.1 Business area codes

Since businesses have been previously maintaining customer records for themselves, there has not been a reason to identify customers by using e.g. business area codes for specific businesses. Originally the core CMD system was only used by one business area within the case company and later the master data of other businesses was emerged to the same system as well. So far there has been only few businesses that have needed the business area code for reporting and data distribution purposes, and it was noticed that it could be utilized by the other businesses as well to assist in customer identification. In total, there were 10 active business areas of which 3 had a business identifier in the core system, so the remaining 7 needed to be identified somehow. There was also a possibility that several business areas are using the same customer record so these records with a combination of business area code needed to be identified as well.

Team identified as many of the customer records they could themselves, for example by checking the existence of sales areas in business target systems which would indicate that a certain business was using that customer record, and by separating the customer records which only were activated to company codes used by only one business area. These customers were marked for that business area without the need for business approval. After these actions, lists of customers suspected to belong to a cer-

tain business area were sent to business contacts for checking. Some answers were received but it was visible that it was causing difficulties also for the data users to identify their customers. Approximately 24 000 business area codes were added during the cleaning project.

Business area	21.11.2012	11.03.2013	24.04.2013	24.05.2013	16.07.2013	19.09.2013	11.12.2013	Difference
A	40 940	41 548	41 777	41 943	42 415	43 352	44 300	3 360
B	188	342	383	443	472	1 186	7 242	7 054
C	3 279	4 040	4 119	4 177	4 303	4 406	4 747	1 468
D	30	55	69	86	606	619	645	615
E	533	581	594	607	628	663	687	154
F	6 555	9 258	12 516	13 356	14 884	15 163	15 473	8 918
G	14	21	22	22	115	117	122	108
H	26	388	395	398	411	453	470	444
I	60	622	621	620	620	620	619	559
J	267	399	429	443	1 234	1 267	1 597	1 330
K			1	1	1	5	5	4
L			50	50	52	54	55	5
M			73	73	81	81	82	9
O						17	20	3
P				72	72	75	75	3
							Total amount of added business area codes:	**24 034**

Figure 8 Progress of business area additions

At the moment all new customer records created by CMD team should include business area identifier. When inactive customers are marked for deletion and more active customers are having a business area code, the percentage of records with missing values is decreasing. Customer record having a valid business area code is useful for both data custodians and data users, as it makes the recognition easier. That way data custodian knows who to contact when facing questions. Furthermore, customer identification to business areas makes reporting easier. Additionally, there are some interface reasons why the business area code should be indicated.

4.2.2 VAT number and tax codes

Customers used for invoicing should have a tax code present on the customer record. Unfortunately that is not always the case. Situation is getting better since CMD team is at the moment checking that new customer creations have a tax code if it is needed. However, there are customers who operate without tax codes, such as private persons and some public associations. Tax codes are not required if a customer is not created

for invoicing purposes. Example of this would be for example a delivery customer. Due to data transferring regulations between systems, for some businesses all the customers need to be created as invoice customers to the core system, because that is the only customer subtype transferring to business target system. Since CMD team cannot know for sure which customers are used for delivery purposes or invoicing purposes in business target systems, there is no certain knowledge on which ones should have tax codes and which ones do not need them.

Acquiring the tax codes is a time consuming task since it needs to be made manually. CMD team decided to focus on the invoice customers belonging to European Union area who are missing a VAT number. When possible a missing VAT number was generated from another existing tax code, for example in Finland the VAT number is y-tunnus with a country code in the beginning and without a hyphen. In those cases when it was not possible to generate a VAT number from an existing tax code, some assistance was received from Risk assessment team in case company's Credit Risk department who was able to support on providing such information because it is part of their daily work and they have proper databases to check customer information. Therefore, lists of customers records of missing VAT numbers were sent to them for checking.

VAT registration numbers	14.12.2012	24.04.2013	24.05.2013	16.07.2013	19.09.2013	11.12.2013	Difference
All accounts	43 375	44 703	44 902	46 271	46 813	47 659	4 284
Active accounts	42 893	38 812	35 763	34 547	34 876	35 247	-7 646

Figure 9 Progress of VAT number additions for invoicing customers

After the answers received from Risk assessment team and followed by actions taken by CMD team, the situation of missing VAT numbers was a bit better. The starting point was that approximately 43 000 customers had VAT number and after the cleaning actions the figure was a bit over 47 600. The decrease in VAT numbers for active accounts in Figure 9 is due to the in-activations of dormant customer records.

4.3 Data harmonization

This action concentrated on the diagnosing and correcting conflicting and inaccurate data. The purpose was to identify the most common errors due to the different ways

of entering the data in a system, and typos. Since the amount of customers was high and the check was done without an external validation help, scope needed to be defined in a way that it corresponded to available resources.

4.3.1 Address format

Name and address of the customer should be in official format so that the record can be more easily recognized. List of suspicious characters in these attributes was created, containing characters which normally do not appear in a company name or address. Suspicious characters are for example @ ^ ! \ / ? & ( ) £ [ ] > < | " # $ % { } _.

The suspicious names were checked against the company webpage or a webpage validating VAT registration numbers, or sent to business contacts for checking. It was discovered that several customer records contained numbers in Name 2 and Name 3 attributes. This is because in some countries and their customs there is a requirement to show order or tax number together with the consignee name of the goods. These records needed to be excluded from corrective actions and this needed to be taken into consideration when determining the data quality rules.

In the core system there are separate attributes for main street and house number information. Due to some business target system restrictions and differences in data input methods the house number is often inserted in the main street field. Since there are so many erroneous records, cleaning of these errors was not done systematically throughout the system, due to limited resources. This means that when for example a region of a customer record was corrected, also the street address field was checked to see if the house number was in the right field. If not, it should be transferred to the correct attribute when possible.

When screening the City attributes it was visible that some cities have various formats how they can be entered. One was Saint-Petersburg; there were 16 different ways how it was written in the core system. Together with the business contact a unified format was decided and mass updated to all records having conflicting formats. The city of Moscow was also standardized in the same context. It was proven to be challenging to

decide a unified format to a city name since the city can be entered in English or in a local language depending on which communication language is chosen for that record and for that reason decisions for name standardization need to be made together with the users of data.

It was discovered that customers in certain countries had more information in the City attribute than the actual name of city. This is mainly due to the fact that the data user does not know which ones are the correct attributes to use when they are requesting a new customer record, or on the other hand due to restrictions in business target system when there is a limited amount of attributes in use. For these reasons there was a lot of misuse of Street and City attributes when for example a region or a district was entered to those attributes instead of the attribute designated for that purpose. The countries with most misuse were screened and diagnosed by preparing a proposal to the business contacts where the region and district information were manually transferred to the attributes Region and District and removed from Street or City attributes taking into consideration the business target system restrictions. The customer countries which were targeted in 2013 cleaning project were Great Britain, Ireland, Germany, Romania, Japan, Malaysia, China, Argentina, Brazil, US, Canada, Mexico, Spain, and Italy.

Missing and conflicting postal code formats were also checked. Reviewed customer countries were Denmark, Sweden, Brazil, Azerbaijan, Japan, Estonia, Latvia, Lithuania, Portugal, Hungary, France, Germany, Switzerland, Spain, and Ireland. Since the core system is validating the postal code format for some countries, there was only a need to check missing values and dummy postal codes, for example such as 00000 or XX XXX. These dummy postal codes are widely used due to technical limitations in some business target systems where the postal code is a mandatory attribute to be entered even though in some countries such info is not available. A list of countries not using postal codes was gathered to help the CMD team in managing this in the future. Lists with erroneous records were sent to business contacts for checking and corrections were made when possible.

Additionally to conflicting data, also missing values in address attributes were screened against the existing business rules, e.g. for customers located in Brazil and US it is

mandatory to insert a valid region in the Region field in order to generate the tax jurisdiction code correctly. Another example are customer records where a country is inserted in both City and Country attributes and the actual city information is missing. The task was not easy since there are some countries where a city and a country can have the same identifier, such as Singapore. Customer records with missing mandatory data were sent to business contacts to fill in the missing values.

Country attributes were checked in a way that the customers with a country code that did not exist anymore were extracted from the data and corrected. One no longer existing country was discovered which was Yugoslavia. Data was also tested against empty values, but since the Country attribute is a mandatory attribute in the core system, it was impossible to find empty values. Invalid values were checked and corrected when possible. They were not systematically screened throughout the data due to lack of resources but most visible cases were targeted. These were for example country codes resembling each other such as SV and SE for Swedish customers.

In the core system, also an international version of address with local characters can be entered for the customer as the default version is with Latin alphabetic. Available versions are East European, Chinese, Kanji, and Cyrillic. When doing changes to master data on general level, it is good to check if the record has an international version and if so, also make the same corrections there. International versions were also screened independently in order to discover oddities in the data. This was done fairly superficially. Suspicious records in Kanji and Chinese language version were sent to APAC team to check and correct if needed.

4.3.2 Business specific data actions

All of the business areas have had their unique way of maintaining their business specific attributes, and the data is also scattered in different systems. Business specific data can refer to e.g. logistics or sales data, which is not scoped as global data within the case company. However, because the largest business area is maintaining its business specific information of sales area data in the core system, for that reason it was also scoped in the cleaning project. The data was screened and diagnosed and then the suspicious or erroneous records were sent to contact persons for approval.

Attributes that needed correction were for example partner function linkings and payment terms. In the partner functions, a sales manager among other sales and delivery information is linked to the customer. For one business area it was discovered that some invoice customers were missing a sales manager value. This means that in one of the business target systems the customer is not assigned to any sales manager. These records were gathered together and sent to business contact person for checking if either a sales manager value should be added or in case that invoice customer was in fact used only as a consumer or a delivery customer so the value can be null.

CMD team was also able to review the customer segmentation since it is maintained in the core system on a general level. As the business evolves, also the segmentations change during time. The new values are maintained to the customers both manually and with mass updates. It happens sometimes that a customer record is left without the segment correction because of the technical error, leaving it with an invalid segment value. The segments were screened and diagnosed, and after the contact person approval corrected to valid values.

The values in attribute drop-down lists were also checked in case of invalid or duplicate values. When reviewing the payment term list some values with different code but same description were found, whereas in place code list invalid values were found such as Dublin-GB when correct value is Dublin-IE. Since the content of these drop-down lists are not maintained by CMD team, the correction requests were sent to the responsible team.

5 Ensuring data integrity in the future

With the help of the theories surrounding Master Data Management, data integrity and data cleaning, and the experiences in the data cleaning project, I have prepared the following guide for case company on how to ensure data integrity in the future. This guide consists of instructions divided into three categories: Data quality rules, data control, and data cleaning project. I will begin with a short introduction of the importance of data integrity for the case company.

5.1 Importance of data integrity

Good data quality has many benefits. It makes the data more trustworthy and decreases the time user has to spend searching and checking the data. It also saves time as there are no corrections needed later, for example if a wrong customer record is used, shifting to use the correct record can be time consuming and frustrating if orders have already been placed to the incorrect record. For internal control it is also extremely important to have the data in an updated state. Using records with invalid tax numbers is very risky for the business. Good data integrity results from valid, accurate and consistent data. When data integrity is high, the data follows business rules and it is timely, and it satisfies the business needs.

Quality problems in core system are most likely also transmitted to business target systems. There can also be differences in the way data is maintained in independent systems compared with how it is in core system. This means that there are different business rules in these systems. For example, in an independent data system it might be of a habit to use the same record although VAT number changes by just replacing old VAT number with a new one. This is against the business rules of core system since a VAT number change causes the record to be a new business entity. Business rules are describing the What, When, Where, How and Who of the master data. Some of them are automatically validated by SAP when entering information to the system, e.g. same customer number cannot be created more than once, but most of the rules need to be maintained and monitored manually. Business rules can be divided into entity rules, attribute rules, and data dependency rules. Entity rules usually describe

the capability of data, for example issues concerning the data validation, data entry, or data process. With attribute rules the functional requirements of data are viewed as in how the data flows from the core system to business target systems. With data dependency the processes and interface related business rules are reported. I have recorded business rules of case company's core system for MDM project concerning customer domain and they can be found in appendix 4. It is important for data custodians to be able to see the business rules around the master data so that they have the competences to validate and maintain the rules in their work. These rules need to be kept in mind when users of independent systems are requesting changes to core system. This means that managing data quality is not only a onetime task, but an ongoing daily action.

Data quality dimensions can be divided into following categories: Duplicate data, conflicting data, incomplete data, invalid data, and inaccurate data. Data quality rules must be determined in a way that the data would be as free as possible of these factors causing data quality problems. There are some basic principles of data quality that should be considered: Data quality must cover the entire lifecycle of the record, it must be applicable to all systems the data is transferring to, it must be measurable over time, it must be monitored and communicated, and data quality requires training.

5.2 Data quality rules

Data quality can be determined with rules. In addition to determining data quality rules, they also need to be communicated to the users of data. According to the existing literature, quality rules can be divided into following groups: Data completeness, data correctness, data accuracy, data precision, data uniqueness, and data consistency. Data completeness and correctness mean that data needs to be reliable, whereas data accuracy and data uniqueness are referring to validity of data. Data precision is determining the usability of data, and data consistency refers to its timeliness. To define and report data quality rules for MDM project I used the predefined determinants used for the project and noticed that they are very useful and I would suggest that they would be used also in the future when reviewing data quality rules. Below these data determinants are introduced separately.

Determinant	Description
Timeliness	Describes the availability of data throughout its lifecycle
Validity	Describes the degree to which the data reflects against data standards
Reliability	Describes the degree to which the data can be trusted
Usability	Describes the degree to which the record meets the needs of the data users

Figure 10 Determinants of data quality

In order to record data quality rules, the determinants can be divided into dimensions and reviewed according to company's master data. The determinants and their dimensions can be found from the below picture.

Determinant	Dimension	Description	Comment
Timeliness	Age of data	The degree of how current the data is.	If a new version of the record exists elsewhere, the record is not current.
	Availability	The measure of when the data must be available in order to fulfil business needs.	Policies per system level define when the data must be available.
Validity	Completeness	The measure of data content in order to fulfil business needs.	Mandatory information must be available.
	Accuracy	The measure of correctness of data.	The record must follow the data standards and procedures.
	Relevancy	The measure of how much the needs of the data users are met.	If an attribute is passed on to a system that does not require it, the data is not relevant.
	Uniqueness	The measure of redundancy of a data.	The data standards define which attributes must be classified as unique.
Reliability	Consistency	The measure of how the data follows the standards over time.	The data should be used for only the purpose it has been defined for.
	Standardization	The measure of how the data is standardized.	All records should follow the plan determined for the organization.
Usability	Accessibility	The measure of how the data is accessible to the required users.	If the record is available but not accessible then it is not considered acessible.
	Usefulness	The degree of how the data can be used to fulfil business requirements.	Attributes must have certain business impact.

Figure 11 Dimensions of data quality

After data quality rules are divided into the dimensions, they can be reviewed against some additional data standards and used in those data quality rules where possible

and needed. These additional data standards for ensuring data quality are introduced in the following picture and an example is provided from each standard.

Data standards	Description	Example
Value range	Describes the allowed value ranges for the attribute.	Payment term has to be between Advance payment and 160 days net.
Value list	Describes the allowed values for the attribute	Business areas defined for the company.
Text column rules	Describes the rules for textual descriptions.	International language characters should not be used on the global data level.
Character patterns	Describes the character patterns used for the attribute	Customer number for internal customer has to start with the company code.
Existence check	Describes the accepted values for existence checks.	If existence check can be conducted, then use Y, if not use N.
Formats	Describes the accepted formats used in attributes.	Attribute "VAT registration number" is alphanumeric.

Figure 12 Additional data standards for data quality

I have defined the data quality rules for MDM project for customer domain and they can be found in appendix 5. The business rules and data cleaning rules were designed for the MDM program to be implemented in 2014. Therefore, they are reflecting the situation as it is now and are likely to be modified through time as processes change.

5.3 Data controlling

According to my experiences, defining and communicating the data quality rules is usually not enough, and data controlling is necessary. I will now give some suggestions on how data can be controlled to ensure data quality.

Access rights need to be systematically recorded and reviewed in order to have a clear view who are the employees having maintenance rights for the data. Only qualified persons should be able to modify the data.

Systematic checking of data quality is important. This can be done by manually testing the data against data quality rules and reporting the results to the data owner. There

are also systems available for testing the data against format rules, such as SAP Information Steward. For controlling the validity of the data, external validation services can be used.

Training is an important issue in data controlling. This applies both to data custodians and users of data. Knowledge on new country specific data rules and business requirements are examples of important information to data maintenance team. Data quality rules need to be communicated also to the users of data, as they are the ones requesting for new customer records. When a request is already corresponding to data quality rules, it saves time as the request can be handled in one go. In case the request does not correspond to data quality rules, data custodian needs to send it back for corrections. Usually new record creations are urgent and time is wasted when a request has to be sent back and forth. A data custodian might think he or she is saving time by making the corrections needed instead of sending it for corrections to the requestor. There are two reasons why this is wrong. First of all, for internal control purposes every request must be archived in the same form it was approved by an authorized person. Another reason is that if a data custodian is always correcting the errors made by requestors, it is a vicious cycle and the data custodian is carrying out tasks that should belong to the requestor. One might wonder why the requestors are not trained to fill in the requests according to data quality rules. This is because people tend to forget what they were taught. For this reason training should be systematic, not only focusing on new data users. This is also important because rules might change through time as processes or policies change.

Customer lifecycle management should be handled by business users, to reduce the amount of inactive customers still active in the system. Not only do they take up memory space in vain making the system reaching its maximum capacity sooner, but they also make the reporting slower. Additionally, it complicates the search for correct record, and in worst case wrong record is used. For this reason, business and data quality rules should be enforced and not only communicated to users of data.

5.4 Data cleaning project

Although business and data quality rules are enforced and data controlling is taking place in data domains, there will most probably still be errors in the data. One simple reason for this is typos and the fact that personnel is changing. In case it is decided that data quality management as an ongoing process is not enough and a data cleaning project should be conducted, I have prepared some models to be used in planning and implementing data cleaning actions.

5.4.1 Planning the project

Gartner, Inc "Seven building blocks of MDM" was introduced earlier in this study. Before starting to conduct actual data cleaning actions, also here it is necessary to get everyone who needs to be involved aware of the project. Similarly to Gartner's model, I have defined steps to be considered in the early stages of a data cleaning project.

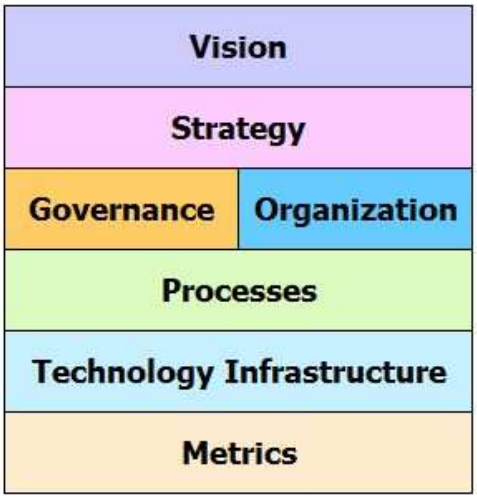

Figure 13 Planning a data cleaning project

Vision: Here the scope of a data cleaning project can be described, for example which systems will be affected and what are the main problems of data to be focused on.

Strategy: How the data cleaning actions will be realized and who is supervising the project.

Governance: Contact persons for the project should be defined and contacted before any actions are initiated. Most likely the contact persons are users of data who understand the scheme of data structure and can act as representatives of the group of users. Data cleaning actions need to be defined together with the governance.

Organization: Personnel executing the cleaning actions are nominated and a communication plan and an action plan will be prepared.

Technology Infrastructure: Determining the tools used in the cleaning project and are they provided in-house or from external sources.

Metrics: Before starting the data cleaning actions, statistics on current situation on the scoped data should be taken out in order to compare the figures before and after the project. Additionally, it might be also useful to take statistics during the project to record the process of cleaning actions.

5.4.2 Implementing the project

When starting to implement the project it should be done in an organized manner and each step should be reported.

Data cleaning phases can be divided into three steps:

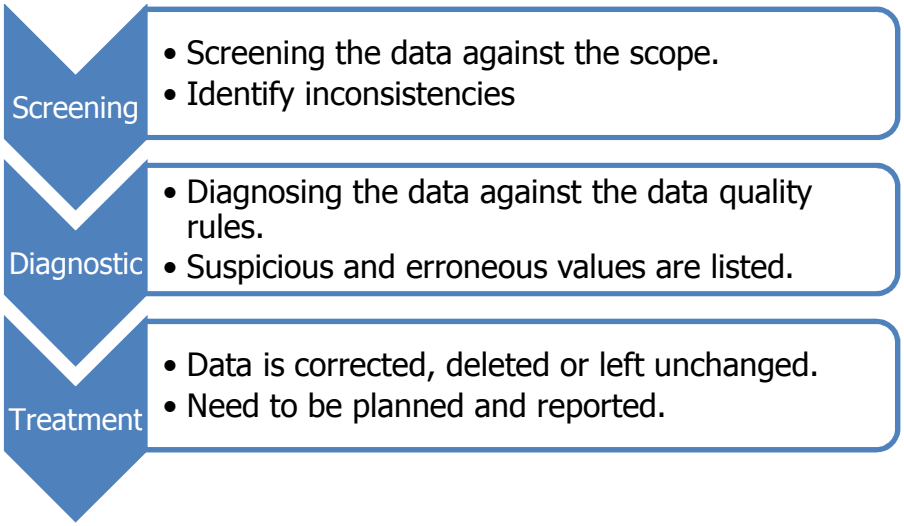

Figure 14 Data cleaning steps

Before starting the Treatment phase, data cleaning method needs to be determined. Possible methods can be divided into three different actions. Reporting integrity violations but not making the correction is called Auditing data. In case excess data or invalid values are found in the data, removal of those values or records is called Filtering data. When new values are needed to replace the inaccurate, invalid or conflicting data, it is called Correcting data. Priority of each task should be determined at this point as well.

Usually there is a need for approval before actual corrective actions can be made to the records. Although it is clear that the violation exists, there might be a business or system reason behind it that we are not aware of. These reasons can be for example business target system requirements such as postal code format for logistics system, or that a tax code is required to be visible on address line for the customs. For approval process, the appropriate contact persons need to be decided and contacted. They can be the same persons as in the governance mentioned in chapter 5.4.1.

To determine the schedule and order in which actions are executed and the dependency of each action, a Gantt chart can be used. A Gantt chart is commonly used in project management, as it displays activities against time. In a data cleaning project there are actions that can and should be executed first, before some other actions are relevant. It might also be the case that there are actions that are difficult or impossible to execute before another task is completed. Furthermore, there might be actions that do not need business approval which means that they can be executed first, or while waiting for answers from business. A Gantt chart designed for the 2014 data cleaning project can be seen in appendix 6.

Actions in data cleaning project can be divided into three main groups; data cleaning, data enrichment, and data harmonization, depending on the nature of tasks inside them. Cleaning of data refers to getting rid of redundant data. These can be for example duplicate customers, dormant customers, or a record with invalid data such as outdated VAT numbers. With data enrichment activities incomplete data is reviewed and corrected. Example can be for example missing business area codes which results in identification challenges or missing tax numbers. Conflicting and inaccurate data is reviewed and corrected when possible with data harmonization. Conflicting data might

be for example region information written in a same field after the city name instead of using the Region field, whereas inaccurate data may refer to a customer's address not corresponding to the country indicating that a wrong country code has been chosen for the record. There can be several country specific rules that need to be taken into consideration. I have prepared for this reason a spreadsheet listing the country specific rules we are aware of. They are introduced in Appendix 6.

The Implementation plan prepared for Data cleaning project in 2014 can be seen in appendix 7.

All data cleaning actions and results have to be carefully reported so they can be presented to the governance group, and also to have them reported for future cleaning projects.

6 Conclusions

Master data management and data quality management is important because data is a valuable asset to a company. It is intellectual property and therefore must be protected, managed and governed for it to stay on a high level. This Master's thesis was conducted for my employer during a Master Data Management project.

The purpose of this study was to provide guidelines and best practices on how to achieve and maintain high level integrity of customer data. In this thesis, first the aspects of Master Data Management and data quality are introduced, followed by a current state analysis, and then practical guidelines are provided based on the existing literature and my own experiences. The research question was how to ensure data integrity in the future. In addition, since I have experience in conducting data cleaning actions without the assistance of a data cleaning tool, another research question was how to clean the data manually. I think this is an important topic and it is very useful to report the ways that data integrity can be kept on a high level in the case company.

The research results can be used for several purposes. Firstly, the Customer Master Data team of the case company may use the results in their future data cleaning projects. In case employees change, or information about the previous projects is not effectively reported, project team can find assistance from the guidelines provided by this research. Secondly, other users of master data can use this guide. There are different teams handling master data in the case company, for example Vendor Master Data team, APAC and NA regional Customer Master Data teams, and business specific master data teams both in the core system and business target systems. In addition, the research results can be used outside the case company when adjusted to their own master data system and business processes.

6.1 Findings

High data quality is important to a company since the costs and consequences resulting from inaccurate information. According to the existing literature, the data quality dimensions fall into the following categories; duplicate, conflicting, incomplete, invalid,

and inaccurate data. In order to protect data from these inconsistencies, clear rules need to be defined. Normally, master data system is programmed in a way that it automatically validates some of the business rules. In addition there are rules that are carefully determined to protect processes and ensuring that the data is relevant for business usage. According to existing literature there are also rules needed to maintain data integrity. Each company must determine their rules on the validation of data, how it is entered into the system and how the record lifecycle is maintained. These rules are called data quality rules.

According to my research, data cleaning actions need to be reported systematically and as accurately as possible. This must be done in a way that the actions done are clearly stated so that a person outside the cleaning project also understands them, or a person involved in the project understands what was done when reviewing the report later. Good communication between data cleaning project group and business representatives is crucial. When there are hundreds of data users, it is not easy to have a single person representing all the users of a certain business or geological area and there is always someone who does not agree with the data cleaning actions taken. There are many reasons for this, for example data might be used against business rules or there is resistance for change. There might also be secondary system requirements which do not follow data quality rules of the core system. These issues need to be handled case by case and both parties need to be able to make compromises. In the 2013 data cleaning project described in this Master's thesis, cleaning tasks were started simultaneously and it was discovered that the data enrichment actions should have been started only after duplicate and inactive customer cleaning, since there are less active customer records left to be enhanced. A Gantt chart can be used to indicate the schedule and dependency of tasks. Additionally, the lists of suspected erroneous values need to be checked as soon as possible from the data extraction date, meaning that the screening and diagnosing of data has to be done promptly followed by sending the lists of suspected erroneous values to business contacts with a deadline. In case the time period is too long, master data might be modified in the core system meanwhile. One additional advice would be to keep it simple, meaning that the main quality dimensions to be focused on based on the data quality analysis should be determined, and focus on them. In case new tasks are identified and executed continuously, data cleaning can be seen as an on-going process rather than a project. When the data

cleaning is done as a project it must have an ending data, when the results are reported. In this case the new tasks can be then schedules to the next data cleaning project.

Due to large amount of data users and the historical problems with data quality in case company, I recommend data controlling to protect the data integrity. This can be done by reviewing core system maintenance rights systematically and not only by training the new users of data but also having a follow up on the existing users. Business rules and data quality rules also need to be enforced to business users of data to avoid misuse at their end. In case data controlling is not effective enough, and both business rules and data quality rules are not systematically followed when maintaining data, or there is a history of misuse, I suggest a data cleaning project is needed to check the level of quality and if needed cleaning actions need to take place. According to my research, data cleaning project should be clearly defined before starting the actual cleansing of the data. I have prepared a model and guidelines on how to plan and implement the data cleaning project. This guide is based on the existing theoretical literature and my own experiences. This study has provided me great help when planning and implementing the 2014 data cleaning actions as it captures the existing theory combining it with the processes and resources available in the case company. Hopefully, this study will be of assistance in the future as well.

According to my findings, same issues need to be taken into consideration when planning a data cleaning project and a MDM project so I adapted the "Seven Building Blocks of MDM" from Gartner, Inc for data cleaning planning model. As the aspect of data quality is of high importance in the MDM program, it could be stated that data cleaning is a smaller project inside a MDM project following the same structure.

Theory around data cleaning was found to be versatile, concentrating mostly on the cleaning actions conducted with a system or a tool designed to clean data. It seems that the common view on the matter is that data cleaning cannot be done manually. The data cleaning project introduced in my research was done solely using core system's own queries and the data was analyzed with using Microsoft Excel. On one hand, the risk of human error is present when doing manual cleaning, but systems and programs are not infallible. It is still a human who enters the rules into the system. I have experienced, that it is not usually the maintainer who configures the rules to the sys-

tem, but instead person form IT department. The rules might not be communicated from the data steward to IT in order for the system to do the wanted validation. For example, for a data maintainer some business rules are so obvious that they are accidentally left out of the reported rules, whereas IT configures the rules solely based on the reported rules and does not take into consideration anything else. Resources are one aspect that needs to be considered when choosing the suitable method. How long does it take to program the rules into the system and run the reports, whereas a data custodian knowing the existing data quality rules filters the data with the system's own query transactions and points out the inconsistencies? Weather using a program or manual work, one thing remains the same; the data needs to be analyzed and corrected manually. Therefore, I think that if a data cleaning system is used, it needs to be able to not only screen and filter the data against data quality and business rules, but also make suggestions or even decisions based on very detailed rules and additionally validating the data against external data bases.

6.2 Future studies

The purpose of this Master's thesis was not to create a universal theory for manual data cleaning as a project that would suit all companies, but to provide the case company a guide on how it can establish and maintain data integrity. I found out that there is a gap in the existing literature on how to maintain the data quality manually and how to execute it as a project, not only as an ongoing process. Nowadays data cleaning systems are very popular and manual cleaning is not considered as an option since it is believed to be impossible or too difficult or time consuming. I believe that greater awareness of the possibilities of manual data cleaning should be raised. Firstly, small businesses with limited amount of records are able to maintain their master data manually if they wish so, and many companies do not have the funds to acquire an external data cleaning system. Secondly, as proven in my thesis, it is possible to clean the data manually, without using any external tool. Furthermore, I would be interested on a research on when an external data cleaning system is needed as opposite to cleaning the data manually taking into consideration the benefits and drawbacks of both methods, and is the best method combination of these two?

Since this Master's thesis answers to the question how to achieve and maintain data integrity in the case company manually, I would like to see in the future more theories on manual data cleaning and a cleaning executed as a project in order to further develop the cleaning processes also in the case company. Additionally, it would be interesting to know which data cleaning system would bring most value to the case company compared to manual data cleaning.

7 Summary

This research concentrates on the customer master data cleaning actions due to a Master Data Management (MDM) project in a case company. Goal of this MDM project is to streamline the data in order to offer high quality master data which will support business processes. Master data can be defined as being the core information to the running of business and typically it refers to people, places and items. Data is a very valuable asset of any company and in today's economic uncertainty companies must be able to trust their data when making decisions. Therefore, businesses need to manage their master data and keep it clean and consistent across multiple databases and systems.

Most of the case company's business processes have evolved independently inside the functions which have led to a situation in which nobody has been managing the overall process. As a result the master data has become redundant, inconsistent and inaccurate. This study aims at investigating the aspect of data quality and how it can be managed in the case company by offering guidelines how to maintain data integrity in the core system manually. Results can be used in the future cleaning actions. Qualitative approach was used when conducting this research and the data was collected with using action research.

MDM is a companywide program combining data governance, business processes, data quality, data enrichment, and a technical solution. In the case company it was discovered that poorly handled master data had led to high efforts in analyzing and verification of actual data, and also longer reaction time due to bad data, which can lead to revenue losses. MDM project creates the appropriate policies and rules to keep data consistent and organized from system to another, as data integrity violations usually cascade from system to another.

Data integrity means the accuracy and consistency of stored data when data follows business rules and is timely, and satisfies the needs of the business. Problems with data quality usually fall into following dimensions: Duplicate data, conflicting data, incomplete data, invalid data, and inaccurate data. To avoid these quality issues clear rules should be defined. According to the existing literature, business rules can be di-

vided into entity rules, attribute rules, and data dependency rules. Furthermore, data quality rules can be divided into six groups of data completeness, data correctness, data accuracy, data precision, data uniqueness, and data consistency.

Data cleaning refers to the correction of erroneous data. Data cleaning project, as any other project, should be clearly defined before starting the actions. According to the existing literature, there are three phases in data cleaning: Screening, diagnostic, and treatment phase. There is also a selection of actions on how to deal with integrity violations once encountered. The options are auditing, filtering, and correcting. Data cleaning requires transparency and proper documentation.

Gap analysis was used to determine case company's current state analysis. As the case company's business processes have evolved independently and each business has adapted their own way of maintaining customer data, the master data has become redundant, inconsistent and inaccurate. The desired state is that all necessary master data can be found in the core system without having any duplicate data, and the data is in good quality. The statistics were gathered for data quality assessment of the current situation. Result was that data integrity violations were found within all the data quality dimensions described earlier. To reach the desired future state a data cleaning and enrichment plan was drawn in addition to short-term and long-term plan for the Customer Master Data (CMD) team which has the main responsibility in maintaining the global data in the core system. First part of the cleaning project was conducted during 2013 and will be continued in 2014. Data cleaning actions were conducted using core systems own queries combined with Microsoft Excel. Actions taken in data cleaning project 2013 are described in this study and suggestions for future cleaning actions are pointed out regarding the issues faced during the project.

There were three dimensions in the project. The data cleaning actions refer to the attempt to decrease the number of active records in the system. This is done by identifying and deleting inactive and duplicate records from the master data. With data enrichment the incomplete data was diagnosed and corrected when possible. Data harmonization concentrated on diagnosing and correcting conflicting and inaccurate data such as errors in address formats. With the help of theories surrounding Master Data Management, data integrity and data cleaning, in addition to the experiences in the

data cleaning project, guidelines were prepared on how to ensure data integrity in the future. These guidelines include data quality rules, data control, and data cleaning project.

When creating data quality rules, the determinants, dimensions, and additional data standards need to be taken into consideration, and they are described in this Thesis. Defining and communicating these rules is usually not enough and data controlling is needed in the form of access rights checking, data quality assessment, and training. In case ongoing data cleaning processes are not enough, a data cleaning project should take place. The project can be planned in a similar way to MDM project by defining the vision, strategy, governance, organization, processes, technology, and metrics of the project. The project should be then implemented in an organized manner and clearly reported. The cleaning phases and actions are determined according to the existing theories and introduced in this Master's thesis.

According to the existing literature and experiences in data cleaning project, business and data quality rules, data controlling, and data cleaning project planning and implementation are described in order to maintain data integrity in the case company. In addition, in this research it is proven that data cleaning actions can be done manually, making it more affordable to the case company than acquiring an external data cleaning service. The results of this study may be used in future cleaning projects inside and outside the company. In the future it would be interesting to study what is the most favorable combination of manual and external system based cleaning for the case company.

References

Adelman, Sid, Moss, Larissa and Abai, Majid (2005). Data strategy.

Bischoff, Joyce and Alexander, Ted (1997). Data warehouse: practical advice from the experts.

Ghauri, Pervez and Grønhaug, Kjell (2002). Research Methods in Business Studies - A practical guide. Second edition.

Hewlett-Packard Development Company, LP. (2007). The Seven Deadly Sins of Master Data Management. How to avoid mistakes that can sink your MDM effort. Public release.

Kelly, Sean (1997). Data Warehousing in Action.

O'Brien, Rory (1998). An Overview of the Methological Approach of Action Research. Faculty of Information Studies, University of Toronto.

Osborne, Jason W. (2013). Best Practices in Data Cleaning – A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.

Radcliffe, John (2007). The Seven Building Blocks of MDM: A Framework for Success. Gartner, Inc.

Scheidl, Half Abude (2011). Master Data Management Maturity and Technology Assessment – From theory to practice, Case Ineo Oy. Master's Thesis for International Management of Information Technology, University of Turku.

Shah, Jignesh, Manathara, Mathew and Hoeppe, Alexander (2012). Process-Driven Master Data Management for Dummies.

Strout, Steven B. and Eisenhauer, John A. (2011). The Elephant in the Room: Data – What you need to know to best govern and manage your enterprise data.

Tuck, Steve (2008). Opinion piece – Is MDM the route to the Holy Grail? Journal of Database Marketing & Customer Strategy management Vol. 15, 4, 218-220.

Case company Annual report 2011

Case company Annual report 2012

Case company internal material

Van den Broeck, Jan, Argeseanu Cunningham, Solveig, Eeckels, Roger and Herbst, Kobus (2005) Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLoS Med 2(10): e267.

Walker, Andy and Ganapathy, Jagadeeshwaran (2009). Effective Master Data Management with SAP NetWeaver MDM.

Wheatley, Malcom (2004). Cooperation the Key to Clean Data, retrieved from www.cio.com.

Winter, Richard (1996). Some Principles and Procedures for the Conduct of Action Research. New Direction in Action Research 16-17.